PlantTraitNet: An Uncertainty-Aware Multimodal Framework for Global-Scale Plant Trait Inference from Citizen Science Data

Ayushi Sharma[1], Johanna Trost[1], Daniel Lusk[1], Johannes Dollinger[2], Julian Schrader[3], Christian Rossi[4], Javier Lopatin[5], Etienne Laliberté[6], Simon Haberstroh[7], Jana Eichel[8], Daniel Mederer[9], Jose Miguel Cerda-Paredes[10, 5], Shyam S. Phartyal[11], Lisa-Maricia Schwarz[12, 13], Anja Linstädter[12], Maria Conceição Caldeira[14], Teja Kattenborn[1]

[1]Chair of Sensor-based Geoinformatics, University of Freiburg, Germany
[2]EcoVision Lab, DM3L, University of Zurich, Switzerland
[3]Department of Biological Sciences, Macquarie University, Australia
[4]Swiss National Park, Switzerland
[5]Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Chile
[6]Department of Biological Sciences, Université de Montréal, Canada
[7]Chair of Ecosystem Physiology, University of Freiburg, Germany
[8]Department of Physical Geography, Utrecht University, The Netherlands
[9]Institute for Earth System Science and Remote Sensing, Leipzig University, Germany
[10]Data Observatory, Universidad Adolfo Ibáñez, Chile
[11]Department of Forestry, Mizoram University, India
[12]Biodiversity Research / Systematic Botany, University of Potsdam, Germany
[13]Department of Plant Nutrition, Institute of Crop Science and Resource Conservation, University of Bonn, Germany
[14]Forest Research Centre, School of Agriculture, University of Lisbon, Portugal
ayushi.sharma@geosense.uni-freiburg.de, teja.kattenborn@geosense.uni-freiburg.de

Abstract

Global plant maps of plant traits, such as leaf nitrogen or plant height, are essential for understanding ecosystem processes, including the carbon and energy cycles of the Earth system. However, existing trait maps remain limited by the high cost and sparse geographic coverage of field-based measurements. Citizen science initiatives offer a largely untapped resource to overcome these limitations, with over 50 million geotagged plant photographs worldwide capturing valuable visual information on plant morphology and physiology. In this study, we introduce PlantTraitNet, a multi-modal, multi-task uncertainty-aware deep learning framework that predicts four key plant traits (plant height, leaf area, specific leaf area, and nitrogen content) from citizen science photos using weak supervision. By aggregating individual trait predictions across space, we generate global maps of trait distributions. We validate these maps against independent vegetation survey data (sPlotOpen) and benchmark them against leading global trait products. Our results show that PlantTraitNet consistently outperforms existing trait maps across all evaluated traits, demonstrating that citizen science imagery, when integrated with computer vision and geospatial AI, enables not only scalable but also more accurate global trait mapping. This approach offers a powerful new pathway for ecological research and Earth system modeling.

Code — github.com/GeoSense-Freiburg/PlantTraitNet
Datasets —
huggingface.co/datasets/ayushi3536/PlantTraitNet

Extended version — https://arxiv.org/abs/2511.06943

Introduction

Terrestrial plants, as the largest primary producers on Earth, contribute about 60% to the global net primary productivity (Field et al. 1998) and play a critical role in the carbon and energy cycles of our Earth system (Pan et al. 2011; Schlesinger and Bernhardt 2020). However, understanding how plants influence these cycles is challenging, as the functioning of plants varies profoundly according to their traits. For instance, traits such as canopy height and leaf area control resource acquisition, while leaf tissue properties, such as nitrogen content or dry matter content, are indicators of plant resilience (Díaz et al. 2016). Although these traits are essential for understanding ecosystem processes, the data on such traits is sparse, as their measurement involves costly field surveys and laboratory analysis. Global plant trait databases such as TRY (Kattge et al. 2011) aggregate thousands of trait measurements from numerous studies and regions, providing an invaluable resource for functional biogeography and ecosystem modeling. However, even with these collective efforts, significant gaps persist in the geographic coverage of trait data across biomes, ecosystems, and species, constraining our ability to fully understand and predict global patterns of vegetation function and change (Díaz et al. 2016; Kattge et al. 2020).

Given the strong link between plant morphology and function, plant photographs in concert with computer vision offer a promising avenue for large-scale estimation of plant traits. Citizen science platforms such as iNaturalist (Su

and Maji 2021) and Pl@ntNet (Garcin et al. 2021) have collected more than 50 million research-grade plant photographs around the world, creating a unique resource for uncovering global plant trait distributions (Wolf et al. 2022). These datasets, primarily curated for the identification of plant species, provide plant images and species labels, but do not provide trait annotations (Goëau et al. 2025; Stevens et al. 2024; Van Horn et al. 2018). However, prior work has demonstrated that trait information for these species can be indirectly obtained by linking species names from citizen science records to trait databases such as TRY (Schiller et al. 2021; Wolf et al. 2022). Through this species-level matching, trait values can be weakly assigned to images, enabling the construction of large-scale, trait-annotated image datasets. These datasets can then be used to train scalable computer vision models for trait prediction from images (Schiller et al. 2021). Such models enable direct trait estimation from photographs, independent of whether the species is known or if a record exists in a trait database. Here, we attempt to advance this approach by predicting multiple traits simultaneously, leveraging shared visual features and underlying trait correlations. Subsequently, we spatially aggregate trait predictions derived from individual geotagged photographs to create global, gridded geospatial maps representing the trait distributions across plant communities and ecosystems (Schiller et al. 2021).

The geolocation of each photograph not only enables the spatial aggregation of the predictions into geospatial maps but also allows for the integration of spatial context into the prediction process itself (Schiller et al. 2021). For instance, climate, including temperature and precipitation, or phenological information from satellite data is known to be the key variable shaping global trait distributions, making it a promising predictor (Bruelheide et al. 2018; Schiller et al. 2021; Joswig et al. 2022). However, integrating large-scale geospatial products can be challenging due to data gaps and feature selection (Lusk et al. 2025).

Recent advances in geospatial foundation models (GeoFMs) now support the seamless integration of such context into downstream tasks. Examples include Climplicit (Dollinger et al. 2025), which encodes climate information, and SatCLIP (Klemmer et al. 2025), which leverages satellite Earth observation data. Such GeoFMs have demonstrated strong generalization in global mapping applications. In this study, we test the integration of such GeoFMs into visual trait prediction to enhance performance through geospatial context.

An important challenge for computer vision with citizen science data is its inherent heterogeneity and noise (Sierra et al. 2024), ranging from inconsistent image quality due to varied photo acquisition methods (feature noise) to ambiguous trait annotations from weak supervision (label noise). Such data characteristics may result in implausible predictions and leave an imprint on the aggregation into global trait products. Moreover, such training data noise can substantially degrade a model's ability to generalize to unseen data (Lu and He 2022; Arpit et al. 2017; Zhang et al. 2021). To overcome both feature and label noise in the citizen-science data, we propose an uncertainty-aware probabilistic deep learning framework that estimates predictive uncertainty. The predicted uncertainty is used to dynamically down-weight highly noisy samples and to filter out unreliable data points, thereby reducing overfitting to spurious patterns.

Overall, our contributions are summarized as follows:

- We introduce the first machine learning–ready dataset that systematically links crowd-sourced plant photographs from citizen science platforms to species-level trait values derived from global trait databases.
- We present **PlantTraitNet**, the first uncertainty-aware, multimodal, multi-task deep learning model for global-scale prediction of four key plant traits: height (H), leaf area (LA), specific leaf area (SLA), and leaf nitrogen content (LN).
- We apply **PlantTraitNet** on more than 300K independent samples of citizen science photos and spatially aggregate the predictions to global trait maps. A benchmark against globally distributed vegetation survey data (*sPlotOpen*) revealed that these **PlantTraitNet**-derived traits maps consistently outperform previous global trait products.

Related Work

Pioneering work by Schiller et al. (2021) showed that plant traits, such as height, nitrogen content, specific leaf area, or leaf area, can be predicted from citizen science images using weak supervision, where species-level trait labels are derived from the TRY database (Kattge et al. 2020). While Schiller et al. (2021) focused on single-task models, Cherif et al. (2023) showed that predicting multiple plant traits simultaneously can exploit trait-trait correlations and shared features in the predictor data.

However, Schiller et al. (2021) did not assess whether weak supervision enables capturing within-species trait variation (e.g., size differences among individuals of the same species). Moreover, Schiller et al. (2021) did not test how aggregating individual predictions on a global scale resembles large-scale trait variation across the biosphere. Wolf et al. (2022) provided an approach to validate global trait maps using vegetation survey data of plant communities from the collaborative initiative *sPlot* (Bruelheide et al. 2018; Sabatini et al. 2021) linked with trait data from the TRY database (Kattge et al. 2020). This approach provides an effective means to evaluate the potential of computer vision models for generating trait maps at global scale.

A persistent challenge with citizen science data is the noise in both images and labels (Sierra et al. 2024; Schiller et al. 2021), often structured spatially. Such noise can bias both inference and training, as deep networks tend to memorize noisy labels, compromising generalization (Lu and He 2022; Arpit et al. 2017; Zhang et al. 2021).

Here, we build on previous work and advance the global trait mapping from citizen science imagery along the following aspects:

- Using visual and depth-based foundation models (Oquab et al. 2023; Yang et al. 2024) to better represent heterogeneous plant imagery.

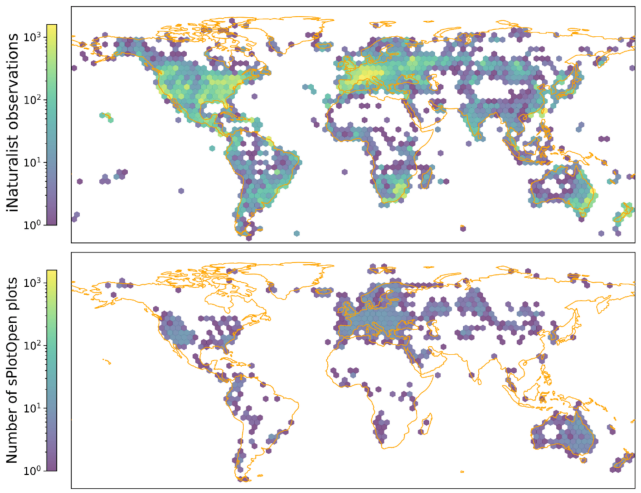

Figure 1: Geographic coverage of the citizen science data (top) and independent benchmark reference data (bottom) from vegetation surveys (sPlotOpen, Sabatini et al. 2021).

- Leveraging multi-task learning to exploit trait correlations (Cherif et al. 2023).
- Incorporating uncertainty-aware training to address label noise (Yeo, Kar, and Zamir 2021; Jiang et al. 2024).
- Benchmarking global trait predictions against sPlot vegetation survey data (Wolf et al. 2022).
- Qualitatively evaluating within-species trait variation.
- Exploring geospatial fusion to enrich trait mapping.

Data

Weakly Labeled Citizen Science Photographs

To predict plant traits at a global scale, we utilize two large-scale citizen science datasets: iNaturalist (GBIF.org 2025; Su and Maji 2021) and Pl@ntNet-300K (Garcin et al. 2021). These datasets consist of plant images annotated with species labels and geolocations but lack direct trait measurements. Following Schiller et al. (2021), we weakly annotate each image using species-level trait distributions from the TRY database (Kattge et al. 2020), based on the premise that interspecific trait variation (variation between species) generally exceeds intraspecific variation (variation within species) (Dong et al. 2020; Wright et al. 2017).

We model each trait as a normal distribution per species, using TRY-derived means and standard deviations, and sample trait values within the interquartile range to reduce outlier influence. To account for intraspecific variability, we resample traits for each image at every training epoch (Schiller et al. 2021).

This weak supervision introduces label noise, especially for traits with strong intraspecies variability across developmental stages (e.g., juvenile trees assigned mature height).

We further reduce noise through model-driven uncertainty estimates (see Methodology). The final dataset includes approximately 220K training images across 5K species and over 80K images in the validation set.

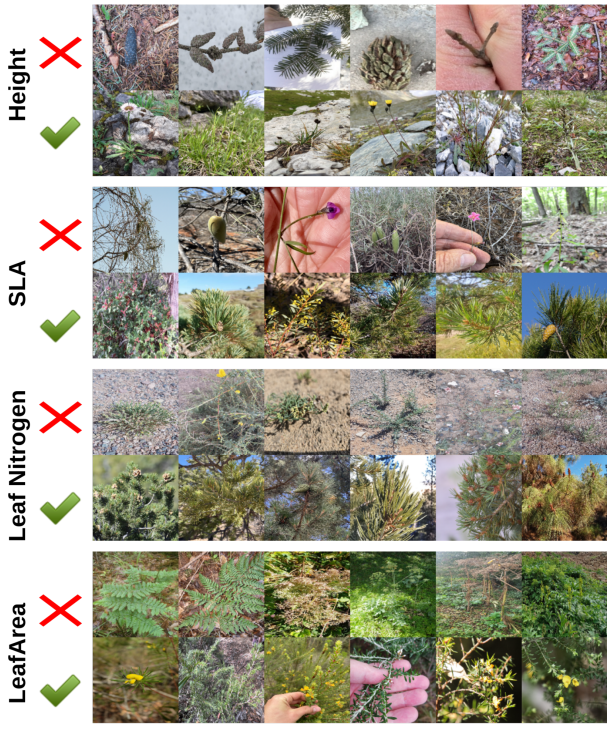

Figure 2: Randomly sampled images showing highest/lowest predictive uncertainty (see Methodology). **Observations**: **Height** uncertainty often from unsuitable contexts (winter scenes, fruits, hands). **SLA** uncertainty from images lacking visible leaves (bare branches, flowers, buds). **Leaf Nitrogen**: low-quality/blurry images. **Leaf Area**: exotic leaf types (e.g., ferns).

Vegetation Survey Data - sPlotOpen

For evaluation, we use the sPlotOpen database (Sabatini et al. 2021). The georeferenced sPlotOpen records represent plant community compositions, which were linked with trait data from the TRY database (Kattge et al. 2020). This data provides global trait maps of community-weighted mean (CWM) trait values.

Reference Data

To aid uncertainty-based filtering, we curated a small dataset of 780 species with images and trait measurements taken from the same individual at the same time including observations from diverse regions such as Germany, La Palma, India, Australia etc. (See Appendix for details)

Methodology

The PlantTraitNet architecture (Fig. 3), uses a general-purpose vision encoder. In addition to image features, we incorporate depth and geospatial priors. These modality-specific embeddings are fused using simple concatenation. The fused representation is passed through a shared multi-modal backbone, followed by trait-specific linear heads.

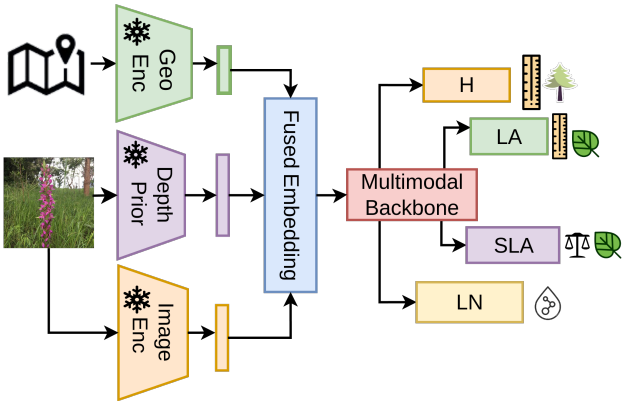

Figure 3: The model integrates image, depth, and geospatial embeddings. These are fused within a multimodal backbone, which then uses individual heads to predict height (H), leaf area (LA), specific leaf area (SLA), and leaf nitrogen (LN).

Image Encoder

Given an input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, we use the pretrained DINOv2 ViT-B/14 encoder to extract a sequence of patch-level feature embeddings $\in \mathbb{R}^{N\times C}$ (Oquab et al. 2023).

We apply adaptive average pooling along the patch tokens N, reducing it to a 32-dimensional representation. This pooling operation is parameter-free and preserves condensed spatial structure before projection. This pooled output is flattened and passed through a multi-layer perceptron (MLP) to generate the embedding of dimension 768.

Depth Priors from Foundation Models

A novel addition to our architecture is the use of depth priors from foundation models for monocular depth estimation. While standard 2D RGB images lack explicit three-dimensional spatial cues, depth information encodes the distance between the sensor and surface points on the plant, enabling a more accurate reconstruction of plant morphology and structure. To incorporate depth, we use the pretrained and frozen encoder from the Depth-Anything-V2 (DA-V2) model (Yang et al. 2024), denoted as h. Although various models could be used, we adopt DA-V2 for its strong generalization capabilities, attributed to its training on large-scale labeled and unlabeled datasets and its student-teacher distillation framework. We use the ViT-B variant, which outputs a set of embeddings $h(I) \in \mathbb{R}^{N\times C}$. Similar to the image encoder, we apply adaptive average pooling along the patch token dimension N, reducing it to a 64-dimensional representation. This pooled output is flattened and passed through a MLP to generate the depth prior embedding of dimension 768.

Geospatial Priors from Foundation Models

Plants are tailored to local climatic conditions, such as precipitation and temperature, through their traits (Joswig et al. 2022).

To incorporate this climatic context as a cue in the prediction process, we integrate Climplicit (Dollinger et al. 2025) into our architecture, a spatio-temporal geo-location encoder trained on the CHELSA climate dataset (Karger et al. 2017). Climplicit maps latitude, longitude, and month of the year to a continuous embedding that implicitly captures climatic factors such as temperature and precipitation. To incorporate seasonal trends, we concatenate the embeddings for the months of March, June, September, and December.

Multimodal and Multi-Task Backbone

Let $\mathbf{X}_{\text{img}} \in \mathbb{R}^{768}$ denote the image embedding obtained from the pretrained DINOv2 encoder, and $\mathbf{X}_{\text{depth}} \in \mathbb{R}^{768}$ denote the depth embedding obtained from the DA-V2 encoder. To incorporate geospatial context, we project the 1024-dimensional embedding produced by Climplicit denoted as $\mathbf{X}_{\text{geo}} \in \mathbb{R}^{1024}$ to a 256-dimensional vector using a trainable linear projection.

The multimodal representation is formed by concatenating all embeddings and is then projected to a 1024-dimensional representation via a linear layer: $\mathbf{Z} = \text{Proj}(\text{concat}(\mathbf{X}_{\text{img}}, \mathbf{X}_{\text{depth}}, \text{Proj}(\mathbf{X}_{\text{geo}}))) \in \mathbb{R}^{1024}$. The resulting embedding is passed through a residual network of 8 residual blocks with hidden dimension twice the embedding size. This architecture and embedding dimensions were chosen based on an ablation across multiple configurations (see Appendix). Finally, the output feature representation is passed to four independent heads for trait prediction in our multi-task architecture.

Uncertainty Estimation

To capture uncertainty in plant trait prediction, each trait-specific prediction head outputs both the predicted value and its associated uncertainty, following the method by (Jiang et al. 2024). For each trait $m \in \{1, \ldots, M\}$, the model predicts two values for each sample n: the mean $\hat{\mu}_n^m$ and the log-scale parameter $\hat{s}_n^m$, where the scale or standard deviation is given by $b_n^m = \exp(\hat{s}_n^m)$.

We model the predictive distribution differently for each trait based on its statistical characteristics. For Leaf Area (LA), which exhibits a long-tailed distribution, we use a Laplace distribution parameterized by mean $\hat{\mu}_n^m$ and scale $b_n^m = \exp(\hat{s}_n^m)$. The Laplace distribution is more suitable for modeling long-tailed distributions compared to Gaussian distributions (Jiang et al. 2024).

For the remaining traits, Height (H), Specific Leaf Area (SLA), and Leaf Nitrogen (LN), we assume a Gaussian distribution with mean $\hat{\mu}_n^m$ and standard deviation $\sigma_n^m = \exp(\hat{s}_n^m)$. Although plant height has strong skewness (dominance of small plants), we employ stratified sampling based on plant functional types during training. This ensures that each mini-batch contains approximately equal representation of grasses, shrubs, and trees, which may make the Gaussian assumption more suitable for modeling this trait (respective ablations are described in the Appendix).

Uncertainty-Guided Data Cleaning Loop

Citizen science image datasets offer large-scale and diverse data for plant trait modeling but suffer from substantial noise and inconsistencies. Common issues include the presence

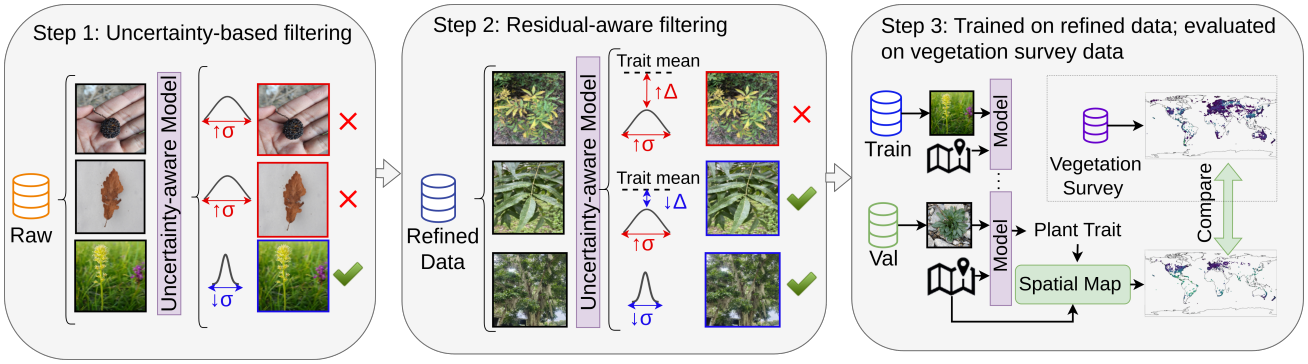

Figure 4: Overview of the pipeline. We filter weakly labeled citizen science data (Raw data) based on high model uncertainty (Step 1) and large residuals from species trait medians (Step 2). We use this refined data for training the models (Step 3), which are evaluated by comparing spatially aggregated predictions (1° resolution) against overlapping vegetation surveys (sPlotOpen).

of non-plant objects, non-representative plant parts, uninformative specimens, and scenes that are too dense, distant, or blurred (Fig. 2). Manual curation at this scale is impractical, and species-level trait annotations often ignore individual variation. Neural networks often memorize noisy labels which harms generalization and makes noise handling essential (Lu and He 2022; Arpit et al. 2017; Zhang et al. 2021). Lu and He (2022) shows models initially learn from clean samples but, past a 'turning point', begin memorizing noise, leading to poorer generalization.

Building on this insight, we implement a two-step data cleaning loop guided by model-predicted uncertainty (Fig. 4). The first step applies uncertainty-based filtering: after early training for a single epoch on raw data, all training images are inferred and ranked by trait-wise uncertainty, and those exceeding a joint threshold across all traits are filtered out. This process continues iteratively until the number of samples jointly flagged as uncertain across traits falls below a predefined threshold.

However, uncertainty alone can be unreliable for heteroscedastic traits such as plant height, where variance naturally increases with trait magnitude. In such cases, high uncertainty may reflect genuine biological variability rather than label noise. Consequently, filtering solely by uncertainty risks biasing the cleaned dataset toward lower-variance samples. To mitigate this, the second stage performs residual-aware filtering, combining uncertainty and prediction residuals. For this, we identify the 'turning point' while training for each trait. We do so by tracking performance on the reference dataset and selecting the epoch after which trait-wise performance begins to deteriorate. Using predictions from this epoch, we calculate the mean absolute error between predicted trait values and species-level means for samples with high uncertainty. Images with high uncertainty and large residuals are filtered from the dataset. The cleaning loop terminates when the number of samples satisfying the filtering criteria becomes negligible. Further details are provided in Appendix.

Model Evaluation and Selection

Following the approach of previous studies (Wolf et al. 2022; Dechant et al. 2024), we aggregated the plot-level trait values from sPlotOpen to a 1-degree spatial resolution to generate a global benchmark dataset. We then applied PlantTraitNet to predict trait values using more than 300K globally distributed citizen science observations. The predicted trait values were aggregated to the same 1-degree resolution and filtered to include only grid cells with at least 20 observations, resulting in over 890 grid cells. To ensure a robust and interpretable evaluation, we used complementary metrics that capture different aspects of model performance: the coefficient of determination (R^2) quantifies the proportion of variance in observed trait values explained by the predictions, the normalized mean absolute error (nMAE) measures the average deviation normalized by the observed trait range, and Pearson's correlation coefficient (r) on log-transformed values quantifies the linear relationship between predicted and observed data. To account for spatial sampling bias, all metrics are weighted by the area of each 1-degree grid cell. We also compare previously published trait maps against the same sPlotOpen CWM values on overlapping grid cells (Boonman et al. 2020; Butler et al. 2017; Madani et al. 2018; Moreno-Martínez et al. 2018; Schiller et al. 2021; Van Bodegom, Douma, and Verheijen 2014; Wolf et al. 2022). For the ablation study, all models were assessed solely on our validation dataset.

The final model, with approximately 90M trainable parameters, was trained for up to 30 epochs with a batch size of 256 on a single NVIDIA RTX A6000 GPU (using approximately 20 GB VRAM). To select the optimal model checkpoint across all traits, we compute the Pareto front using the Non-Dominated Sorting (NDS) algorithm (Deb et al. 2002). We then calculate the hypervolume for all candidate checkpoints on this front (Zitzler, Laumanns, and Thiele 2001) and select the checkpoint that maximizes the hypervolume. Following Lacoste et al. (2019), we estimate a total of 93.86 kg CO_2 emissions for all experiments across seeds. This estimate excludes testing and failed runs and therefore likely underestimates the total emissions, but it provides a reasonable guideline for future model training.

Results

Method	Metric	H	LA	SLA	LN
Ours (Raw)	$R^2 \uparrow$	**0.19**	*0.30*	*0.23*	-0.16
	nMAE $\downarrow$	**0.22**	**0.14**	*0.14*	*0.17*
	$r \uparrow$	**0.45**	*0.56*	**0.59**	*0.49*
Ours (Refined)	$R^2 \uparrow$	*0.18*	**0.34**	**0.27**	*-0.12*
	nMAE $\downarrow$	**0.22**	**0.14**	**0.13**	*0.17*
	$r \uparrow$	**0.45**	**0.57**	**0.59**	**0.50**
Schiller	$R^2 \uparrow$	-0.32	0.11	0.16	**0.06**
	nMAE $\downarrow$	*0.28*	*0.17*	*0.14*	**0.14**
	$r \uparrow$	0.42	0.52	*0.53*	0.40
Wolf	$R^2 \uparrow$	-0.61	-0.02	0.02	-0.20
	nMAE $\downarrow$	0.31	0.18	0.16	0.18
	$r \uparrow$	*0.43*	0.53	0.50	0.41
Moreno	$R^2 \uparrow$	–	–	-0.72	-0.85
	nMAE $\downarrow$	–	–	0.23	0.22
	$r \uparrow$	–	–	0.23	0.17
Butler	$R^2 \uparrow$	–	–	-0.17	-0.50
	nMAE $\downarrow$	–	–	0.18	0.20
	$r \uparrow$	–	–	0.29	0.32
Boonman	$R^2 \uparrow$	–	–	0.03	-0.37
	nMAE $\downarrow$	–	–	0.16	0.18
	$r \uparrow$	–	–	0.49	0.20
Madani	$R^2 \uparrow$	–	–	-0.76	–
	nMAE $\downarrow$	–	–	0.23	–
	$r \uparrow$	–	–	-0.07	–
Bodegom	$R^2 \uparrow$	–	–	-1.00	–
	nMAE $\downarrow$	–	–	0.24	–
	$r \uparrow$	–	–	0.33	–

Table 1: Global trait map benchmarking against sPlotOpen CWMs (1° resolution). **Best**. *Second-best*. External products: Schiller (Schiller et al. 2021), Wolf (Wolf et al. 2022), Moreno (Moreno-Martínez et al. 2018), Butler (Butler et al. 2017), Boonman (Boonman et al. 2020), Madani (Madani et al. 2018), Bodegom (Van Bodegom, Douma, and Verheijen 2014). PlantTraitNet (Ours) is evaluated after training on both raw and refined datasets.

Benchmarking Against Vegetation Survey Data

We benchmark our global trait maps derived from models trained on both raw and filtered data and those of previous studies using R^2, Pearson's r, and nMAE (Table 1). Depending on the metric, our model for LN delivers comparable results with those by Schiller et al. (Schiller et al. 2021). For the other three traits, our models consistently achieve higher performance than previously published plant trait maps. The overall improvement reflects the model's ability to capture complex and variable patterns in large-scale trait prediction. While the r scores suggest that the maps capture relative differences, substantially lower R^2 scores indicate that PlantTraitNet maps and all other products are systematically biased (also see Appendix), underscoring the inherent challenge of revealing morphological and physiological ecosystem patterns at global scale.

Model Performance Across and Within Species

To reveal trait variability within the biosphere on a global scale, a computer vision model must be robust across species, and thus phylogenetic lineages. Using inferences and species information from the validation data, we show that the residuals of PlantTraitNet are largely unsystematically distributed in the phylogenetic space (Fig. 5). To quantify this, we use two standard metrics of phylogenetic signal:

Pagel's λ, which captures broad-scale phylogenetic autocorrelation in residual covariance (Pagel 1999) and Blomberg's K, which is more sensitive to fine-scale signals among closely related species (Blomberg, Garland Jr, and Ives 2003). For SLA and leaf nitrogen, the phylogenetic signal is weak (λ = 0.04 and 0.15; K = 0.0053 and 0.0076), suggesting that prediction errors are largely independent of species relatedness. Although errors for height (λ = 0.80, K = 0.018) and leaf area (λ = 0.56, K = 0.0067) show some phylogenetic autocorrelation, the consistently low K values indicate that even closely related species do not share systematic prediction biases (see Appendix for details). Although PlantTraitNet was trained using weak annotations at the species level, these findings underscore the model's strong generalizability and robustness across the plant tree of life.

Despite weak supervision at the species level, PlantTraitNet captures within-species variability in trait expression. This is particularly evident in the case of height prediction (Fig. 6), where the model reflects differences across growth forms and developmental stages within individual species (see Appendix for additional examples of within-species variability across traits). This suggests that the model is not simply regressing to a species-level mean but is sensitive to morphological cues in the images that reflect ecological and ontogenetic variation.

Effect of Input Modalities

To evaluate the contribution of each input modality to trait prediction, we conduct an ablation study using different combinations of image, geospatial, and depth information (Table 2). Our goal is to understand how each modality influences model performance across key plant functional traits: H, LA, SLA, and LN. For this ablation study, we also experimented with a pretrained BioCLIP (Stevens et al. 2024) encoder as an alternative to DinoV2. For BioCLIP, we extracted the embedding from its classification token, as it empirically showed superior performance (detailed in the Appendix). For geospatial priors, we assess SatCLIP (Klemmer et al. 2025), which is trained on satellite imagery and captures vegetation density and phenology; GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023), a geo-localization model trained on natural images; and Climplicit (Dollinger et al. 2025)

We find that image features alone provide a strong baseline, with DINOv2 and BioCLIP performing comparably.

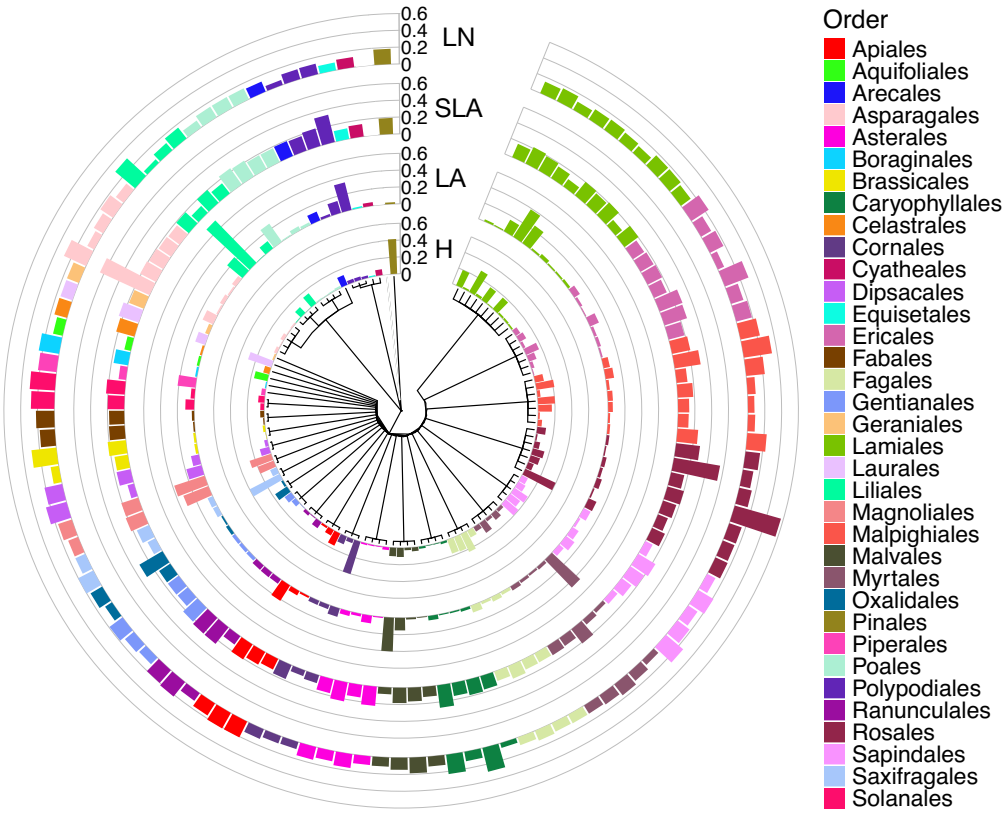

Figure 5: Mean relative prediction error (MRPE) computed on validation data at the family level, visualized along the taxonomic tree, for height (H), leaf area (LA), specific leaf area (SLA) and leaf nitrogen (LN).

	Image	Geo	Depth	H	LA	SLA	LN	# Top ranks
Multi-Task	DinoV2	✗	✗	0.15 ± 0.00	0.31 ± 0.00	**0.32 ± 0.00**	0.14 ± 0.01	1
	BioCLIP	✗	✗	0.15 ± 0.01	0.3 ± 0.00	**0.32 ± 0.01**	0.15 ± 0.04	1
	DinoV2	SatCLIP	✗	0.16 ± 0.02	0.27 ± 0.04	0.25 ± 0.02	0.11 ± 0.05	0
	DinoV2	GeoCLIP	✗	*0.17 ± 0.01*	**0.33 ± 0.01**	**0.32 ± 0.00**	0.15 ± 0.02	3
	DinoV2	Climplicit	✗	**0.19 ± 0.01**	*0.32 ± 0.01*	*0.31 ± 0.01*	0.16 ± 0.06	3
	BioCLIP	Climplicit	✗	**0.19 ± 0.00**	*0.32 ± 0.02*	*0.31 ± 0.01*	0.15 ± 0.06	3
	BioCLIP	Climplicit	DA-V2	0.16 ± 0.01	0.28 ± 0.03	0.30 ± 0.00	**0.19 ± 0.02**	1
	DinoV2	Climplicit	DA-V2	**0.19 ± 0.02**	*0.32 ± 0.01*	*0.31 ± 0.02*	*0.18 ± 0.05*	4
ST	DinoV2	Climplicit	DA-V2	0.12 ± 0.01	**0.34 ± 0.01**	**0.33 ± 0.01**	**0.21 ± 0.02**	–

Table 2: Multi-modal ablation study for plant trait prediction. Results are reported as mean $R^2 \pm$ 1 standard deviation over 3 runs. **Bold** indicates the best result, and *italic* indicates the second-best. '# Top ranks' counts the number of top-two rankings. The last row reports the performance of the best multi-task model when evaluated in a single-task (ST) setting.

Adding geospatial priors from Climplicit consistently improves performance across traits, reflected in higher R^2. Adding depth information on top of the image and climate input leads to marginal changes overall. In general, image features provide a strong foundation for trait inference, while the integration of climate information significantly enhances prediction. Although depth contributes selectively, its inclusion offers a modest gain in average performance, supporting the use of all three modalities in the final model.

Multi-Task versus Single-Task

In Table 2, we also compare the effect of jointly predicting all traits (multi-task) versus independently predicting each trait using the same architecture with single trait heads (single-task). While the single-task architecture yields marginally better performance for LA, SLA, and LN (e.g., higher R^2 and lower nMAE), the multi-task model shows a substantial performance gain for H, improving R^2 from 0.12 to 0.19. Importantly, the multi-task model achieves these results with significantly lower computational cost—training

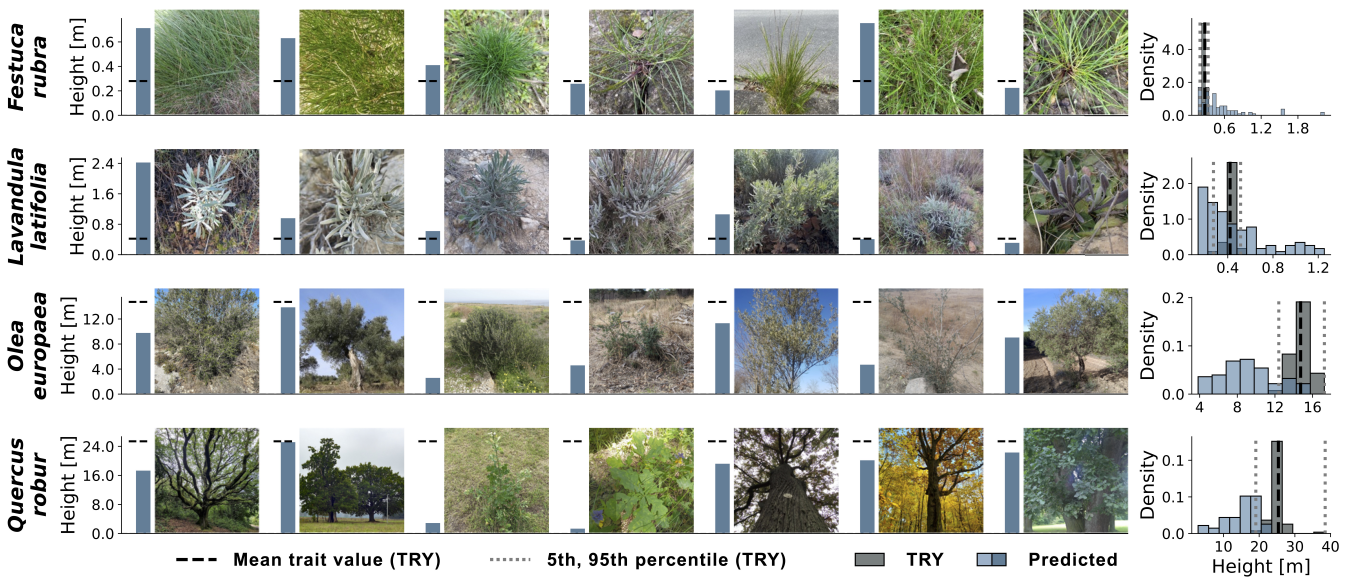

Figure 6: Intraspecific variation in predicted height for four species. Bar plots (left) show model predictions; histograms (right) show height distributions predicted from up to 100 images compared with up to 100 measurements from the TRY database.

a single joint model instead of four separate ones reduces training time and GPU memory usage by approximately 75%. Thus, the multi-task model provides a better overall balance of performance and efficiency.

Discussion

Predicting plant traits from citizen science photos is challenging due to data variability and biases, including spatial and taxonomic bias, and overrepresentation of smaller growth forms such as grasses and herbs (Di Cecco et al. 2021; Sierra et al. 2024). Ecological complexity adds difficulty, as traits vary across biomes, with generalists showing common traits and specialists distinct ones, resulting in skewed, long-tailed distributions (see Appendix). Unlike animals with fixed body plans, e.g. with symmetric and fixed numbers of legs or arms, plants have a comparably flexible morphology, resulting in varying numbers of plant organs, such as leaves or branches. This variability complicates trait prediction via computer vision. Despite these challenges, our results show promising potential. Future work should focus on reducing biases through targeted data acquisition. Increased acquisition of reference data to enable better 'turning point' selection and incorporate label correction to improve model robustness and generalization in ecological contexts.

Conclusion

Our understanding of plant–environment interactions is limited by the sparse geographic and taxonomic coverage of morphological and physiological trait data. We demonstrate that citizen science plant images, combined with machine learning can be used to predict and map global distributions of key ecological plant traits using only geolocated images making the approach highly scalable across biomes. Despite relying on weak supervision via species-level trait annotations, our models capture consistent intraspecific variation. Integrating geospatial context through Earth observation foundation models (GeoFMs) and structural cues via depth priors improves predictive performance and model robustness. Our multi-task framework enables simultaneous prediction of multiple traits, capturing inter-trait dependencies while improving computational efficiency. Benchmarking against existing global trait maps shows that our approach achieves state-of-the-art performance. This establishes a new baseline for large-scale trait inference from image data, offering a powerful alternative to traditional mapping based on field sampling and extrapolation. By leveraging abundant publicly available plant images, our method enables automated, global retrieval of core traits, offering new opportunities to explore functional diversity and improve ecosystem modeling under global change.

Acknowledgements

This study was funded by the German Research Foundation (DFG) within the project PANOPS (Revealing Earth's plant functional diversity with citizen science; project no. 504978936)

References

Arpit, D.; Jastrz_kebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, 233–242. PMLR.

Blomberg, S. P.; Garland Jr, T.; and Ives, A. R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4): 717–745.

Boonman, C. C.; Benítez-López, A.; Schipper, A. M.; Thuiller, W.; Anand, M.; Cerabolini, B. E.; Cornelissen, J. H.; Gonzalez-Melo, A.; Hattingh, W. N.; Higuchi, P.; et al. 2020. Assessing the reliability of predicted plant trait distributions at the global scale. *Global Ecology and Biogeography*, 29(6): 1034–1051.

Bruelheide, H.; Dengler, J.; Purschke, O.; Lenoir, J.; Jiménez-Alfaro, B.; Hennekens, S. M.; Botta-Dukát, Z.; Chytrỳ, M.; Field, R.; Jansen, F.; et al. 2018. Global trait–environment relationships of plant communities. *Nature ecology & evolution*, 2(12): 1906–1917.

Butler, E. E.; Datta, A.; Flores-Moreno, H.; Chen, M.; Wythers, K. R.; Fazayeli, F.; Banerjee, A.; Atkin, O. K.; Kattge, J.; Amiaud, B.; et al. 2017. Mapping local and global variability in plant trait distributions. *Proceedings of the National Academy of Sciences*, 114(51): E10937–E10946.

Cherif, E.; Feilhauer, H.; Berger, K.; Dao, P. D.; Ewald, M.; Hank, T. B.; He, Y.; Kovach, K. R.; Lu, B.; Townsend, P. A.; et al. 2023. From spectra to plant functional traits: Transferable multi-trait models from heterogeneous and sparse data. *Remote Sensing of Environment*, 292: 113580.

Deb, K.; Pratap, A.; Agarwal, S.; and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197.

Dechant, B.; Kattge, J.; Pavlick, R.; Schneider, F. D.; Sabatini, F. M.; Moreno-Martínez, Á.; Butler, E. E.; van Bodegom, P. M.; Vallicrosa, H.; Kattenborn, T.; et al. 2024. Intercomparison of global foliar trait maps reveals fundamental differences and limitations of upscaling approaches. *Remote Sensing of Environment*, 311: 114276.

Di Cecco, G. J.; Barve, V.; Belitz, M. W.; Stucky, B. J.; Guralnick, R. P.; and Hurlbert, A. H. 2021. Observing the observers: How participants contribute data to iNaturalist and implications for biodiversity science. *BioScience*, 71(11): 1179–1188.

Díaz, S.; Kattge, J.; Cornelissen, J. H.; Wright, I. J.; Lavorel, S.; Dray, S.; Reu, B.; Kleyer, M.; Wirth, C.; Colin Prentice, I.; et al. 2016. The global spectrum of plant form and function. *Nature*, 529(7585): 167–171.

Dollinger, J.; Robert, D.; Plekhanova, E.; Drees, L.; and Wegner, J. D. 2025. Climplicit: Climatic Implicit Embeddings for Global Ecological Tasks. *arXiv preprint arXiv:2504.05089*.

Dong, N.; Prentice, I. C.; Wright, I. J.; Evans, B. J.; Togashi, H. F.; Caddy-Retalic, S.; McInerney, F. A.; Sparrow, B.; Leitch, E.; and Lowe, A. J. 2020. Components of leaf-trait variation along environmental gradients. *New Phytologist*, 228(1): 82–94.

Field, C. B.; Behrenfeld, M. J.; Randerson, J. T.; and Falkowski, P. 1998. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*, 281(5374): 237–240.

Garcin, C.; Joly, A.; Bonnet, P.; Affouard, A.; Lombardo, J.-C.; Chouet, M.; Servajean, M.; Lorieul, T.; and Salmon, J. 2021. Pl@ntnet-300k: A plant image dataset with high label ambiguity and a long-tailed distribution. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.

GBIF.org. 2025. GBIF Occurrence Download. Accessed: 2025-07-03.

Goëau, H.; Martellucci, G.; Bonnet, P.; Vinatier, F.; and Joly, A. 2025. PlantCLEF2025 @ LifeCLEF & CVPR-FGVC. https://kaggle.com/competitions/plantclef-2025. Kaggle.

Jiang, Y.; Garnot, V. S. F.; Schindler, K.; and Wegner, J. D. 2024. Uncertainty Voting Ensemble for Imbalanced Deep Regression. In *DAGM German Conference on Pattern Recognition*, 329–343. Springer.

Joswig, J. S.; Wirth, C.; Schuman, M. C.; Kattge, J.; Reu, B.; Wright, I. J.; Sippel, S. D.; Rüger, N.; Richter, R.; Schaepman, M. E.; et al. 2022. Climatic and soil factors explain the two-dimensional spectrum of global plant trait variation. *Nature ecology & evolution*, 6(1): 36–50.

Karger, D. N.; Conrad, O.; Böhner, J.; Kawohl, T.; Kreft, H.; Soria-Auza, R. W.; Zimmermann, N. E.; Linder, H. P.; and Kessler, M. 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific data*, 4(1): 1–20.

Kattge, J.; Bönisch, G.; Díaz, S.; Lavorel, S.; Prentice, I. C.; Leadley, P.; Tautenhahn, S.; Werner, G. D.; Aakala, T.; Abedi, M.; et al. 2020. TRY plant trait database–enhanced coverage and open access. *Global change biology*, 26(1): 119–188.

Kattge, J.; Diaz, S.; Lavorel, S.; Prentice, I. C.; Leadley, P.; Bönisch, G.; Garnier, E.; Westoby, M.; Reich, P. B.; Wright, I. J.; et al. 2011. TRY–a global database of plant traits. *Global change biology*, 17(9): 2905–2935.

Klemmer, K.; Rolf, E.; Robinson, C.; Mackey, L.; and Rußwurm, M. 2025. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4347–4355.

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lu, Y.; and He, W. 2022. SELC: self-ensemble label correction improves learning with noisy labels. *arXiv preprint arXiv:2205.01156*.

Lusk, D.; Wolf, S.; Svidzinska, D.; Dormann, C. F.; Kattge, J.; Bruelheide, H.; Sabatini, F. M.; Damasceno, G.; Martínez, Á. M.; Violle, C.; et al. 2025. From smartphones to satellites: Uniting crowdsourced biodiversity monitoring and Earth observation to fill the gaps in global plant trait mapping. *bioRxiv*, 2025–03.

Madani, N.; Kimball, J. S.; Ballantyne, A. P.; Affleck, D. L.; Van Bodegom, P. M.; Reich, P. B.; Kattge, J.; Sala, A.; Nazeri, M.; Jones, M. O.; et al. 2018. Future global productivity will be affected by plant trait response to climate. *Scientific reports*, 8(1): 2870.

Moreno-Martínez, Á.; Camps-Valls, G.; Kattge, J.; Robinson, N.; Reichstein, M.; van Bodegom, P.; Kramer, K.; Cornelissen, J. H. C.; Reich, P.; Bahn, M.; et al. 2018. A methodology to derive global maps of leaf traits using remote sensing and climate data. *Remote sensing of environment*, 218: 69–88.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature*, 401(6756): 877–884.

Pan, Y.; Birdsey, R. A.; Fang, J.; Houghton, R.; Kauppi, P. E.; Kurz, W. A.; Phillips, O. L.; Shvidenko, A.; Lewis, S. L.; Canadell, J. G.; Ciais, P.; Jackson, R. B.; Pacala, S. W.; McGuire, A. D.; Piao, S.; Rautiainen, A.; Sitch, S.; and Hayes, D. 2011. A large and persistent carbon sink in the world's forests. *Science*, 333(6045): 988–993.

Sabatini, F. M.; Lenoir, J.; Hattab, T.; Arnst, E. A.; Chytrỳ, M.; Dengler, J.; De Ruffray, P.; Hennekens, S. M.; Jandt, U.; Jansen, F.; et al. 2021. sPlotOpen–An environmentally balanced, open-access, global dataset of vegetation plots. *Global Ecology and Biogeography*, 30(9): 1740–1764.

Schiller, C.; Schmidtlein, S.; Boonman, C.; Moreno-Martínez, A.; and Kattenborn, T. 2021. Deep learning and citizen science enable automated plant trait predictions from photographs. *Scientific Reports*, 11(1): 16395.

Schlesinger, W. H.; and Bernhardt, E. S. 2020. *Biogeochemistry: An analysis of global change*. Academic Press, 4th edition. ISBN 978-0-12-814608-8.

Sierra, E.; Gillespie, L. E.; Soltani, S.; Exposito-Alonso, M.; and Kattenborn, T. 2024. DivShift: Exploring Domain-Specific Distribution Shift in Volunteer-Collected Biodiversity Datasets. *arXiv preprint arXiv:2410.19816*.

Stevens, S.; Wu, J.; Thompson, M. J.; Campolongo, E. G.; Song, C. H.; Carlyn, D. E.; Dong, L.; Dahdul, W. M.; Stewart, C.; Berger-Wolf, T.; et al. 2024. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19412–19424.

Su, J.-C.; and Maji, S. 2021. The semi-supervised inaturalist challenge at the fgvc8 workshop. *arXiv preprint arXiv:2106.01364*.

Van Bodegom, P. M.; Douma, J. C.; and Verheijen, L. M. 2014. A fully traits-based approach to modeling global vegetation distribution. *Proceedings of the National Academy of Sciences*, 111(38): 13733–13738.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Vivanco Cepeda, V.; Nayak, G. K.; and Shah, M. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36: 8690–8701.

Wolf, S.; Mahecha, M. D.; Sabatini, F. M.; Wirth, C.; Bruelheide, H.; Kattge, J.; Moreno Martínez, Á.; Mora, K.; and Kattenborn, T. 2022. Citizen science plant observations encode global trait patterns. *Nature Ecology & Evolution*, 6(12): 1850–1859.

Wright, I. J.; Dong, N.; Maire, V.; Prentice, I. C.; Westoby, M.; Díaz, S.; Gallagher, R. V.; Jacobs, B. F.; Kooyman, R.; Law, E. A.; et al. 2017. Global climatic drivers of leaf size. *Science*, 357(6354): 917–921.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything V2. *arXiv preprint arXiv:2406.09414*.

Yeo, T.; Kar, O. F.; and Zamir, A. 2021. Robustness via cross-domain ensembles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12189–12199.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zitzler, E.; Laumanns, M.; and Thiele, L. 2001. SPEA2: Improving the performance of the strength Pareto evolutionary algorithm. In *Proceedings of the 2001 Congress on Evolutionary Computation (CEC 2001)*, volume 1, 959–966.