

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

Ayushi Sharma¹, Johanna Trost¹, Daniel Lusk¹, Johannes Dollinger², Julian Schrader³, Christian Rossi⁴, Javier Lopatin⁵, Etienne Laliberté⁶, Simon Haberstroh⁷, Jana Eichel⁸, Daniel Mederer⁸, Jose Miguel Cerda-Paredes^{10, 5}, Shyam S. Phartyal¹¹, Lisa-Maricia Schwarz^{12, 13}, Anja Linstädter¹², Maria Conceição Caldeira¹⁴, Teja Kattenborn¹

¹Chair of Sensor-based Geoinformatics, University of Freiburg,

²EcoVision Lab, DM3L, University of Zurich

³Department of Biological Sciences, Macquarie University

⁴Swiss National Park

⁵Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez

⁶Université de Montréal

⁷Ecosystem Physiology, University of Freiburg

⁸Department of Physical Geography, Utrecht University

⁹Universität Leipzig

¹⁰Data Observatory, Santiago

¹¹Department of Forestry, Mizoram University

¹²Biodiversity Research / Systematic Botany, University of Potsdam

¹³Institute of Crop Science and Resource Conservation, Department of Plant Nutrition, University of Bonn

¹⁴Forest Research Centre, School of Agriculture, University of Lisbon

{ayushi.sharma, 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/⟨⟩>

Copyright © 2026, Authors. Preprint version. Accepted at AAAI 2026.

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 modelling. 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 pho-

tographs 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 (Bruehlheide 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 joint 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* (Bruehlheide 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.
- 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.

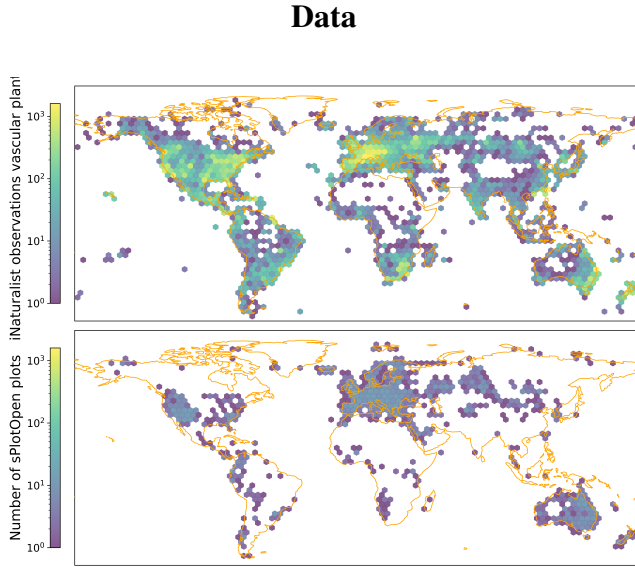


Figure 1: Geographic coverage of the citizen science data (top) and independent benchmark reference data (bottom) from vegetation surveys (sPlotOpen, (Sabatini et al. 2021)).

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 generally exceeds intraspecific variation (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 intraspecific variability across developmental stages (e.g., juvenile trees assigned mature height).

We further reduce noise through model-driven uncertainty estimates (see Methodology). The final data includes 230K training images across 5K species and 80K images in validation set.

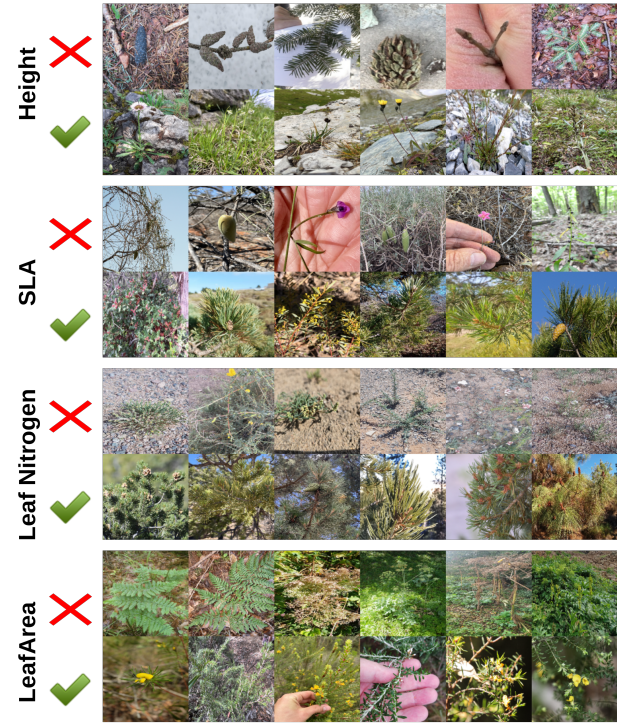


Figure 2: Randomly sampled images showing highest/lowest predictive uncertainty (see Methodology). **Observations:** **Height** uncertainty often from unsuitable contexts (winter scenes, fruits, hands). **SLA** uncertainty from images lacking visible leaves (bare branches, flowers, buds). **Leaf Nitrogen:** 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.

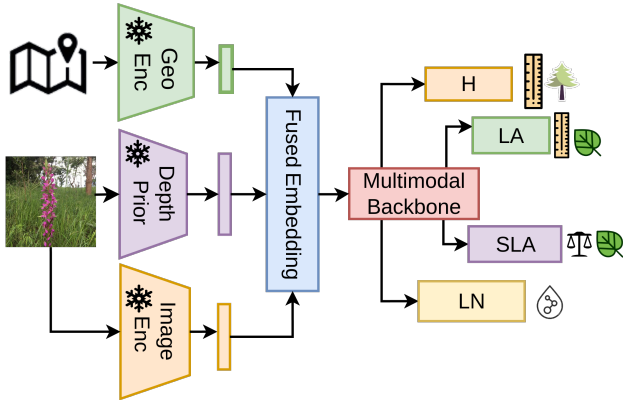


Figure 3: The model integrates image, depth, and geospatial embeddings. These are fused within a multimodal backbone, which then uses individual heads to predict height (H), leaf area (LA), specific leaf area (SLA), and leaf nitrogen (LN).

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 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 Depth Anything 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 of 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 $m \in \{1, \dots, M\}$, 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

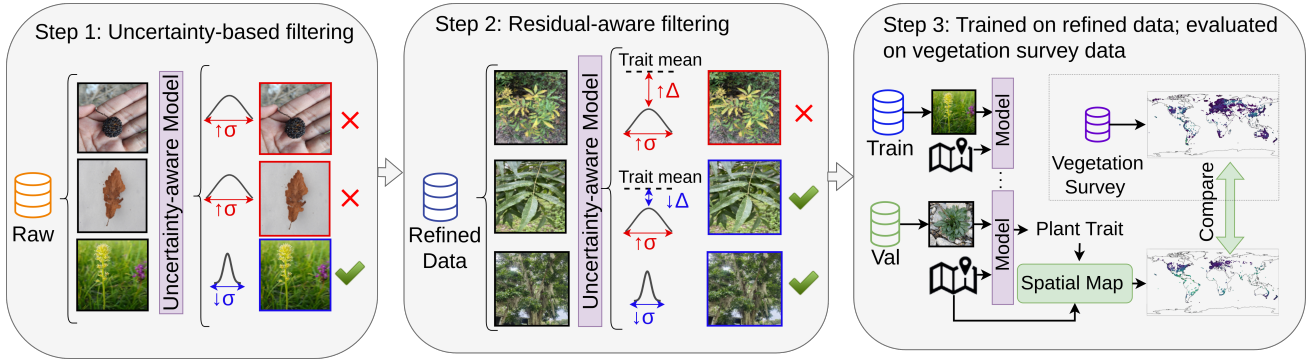


Figure 4: Overview of the pipeline. We filter weakly labeled citizen science data (Raw data) based on high model uncertainty (Step 1) and large residuals from species trait medians (Step 2). We use this refined data for training the models (Step 3), which are evaluated by comparing spatially aggregated 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 Plant-TraitNet 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 ~ 890 grid cells. These predictions were compared against the sPlot-based benchmark dataset using metrics R^2 , mean absolute error (nMAE) normalized by trait range, and Pearson’s correlation (r) on log-transformed values. 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 about ~ 90 M trainable parameters was trained for up to 30 epochs with a batch size of 256 on a single NVIDIA RTX A6000 GPU [using 20GB VRAM]. For selecting 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. Using (Lacoste et al. 2019), we estimate a total of 93.86 kg CO_2 emissions for all experiments across seeds. This does not include testing or failed runs and therefore likely underestimates the overall emissions, but it provides a reasonable guideline for future model training.

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), Van Bodegom (Van Bodegom, Douma, and Verheijen 2014).

Method	Metric	H	LA	SLA	LN
PlantTraitNet (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
PlantTraitNet (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	–
Van Bodegom	$R^2 \uparrow$	–	–	-1.00	–
	nMAE $\downarrow$	–	–	0.24	–
	$r \uparrow$	–	–	0.33	–

Results

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. 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

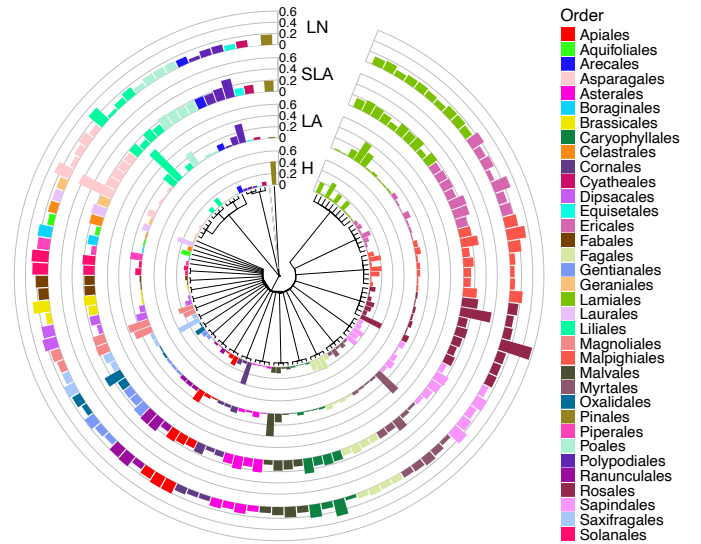


Figure 5: Mean relative prediction error (MRPE) computed on validation data at the family level, visualized along the taxonomic tree, for height (H), leaf area (LA), specific leaf area (SLA) and leaf nitrogen (LN).

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 the inferences and species information of the validation data, we show that residuals from PlantTraitNet are largely unsystematically distributed in 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 is, which is more sensitive to fine-scale signal among closely related species (Blomberg, Garland Jr, and Ives 2003). For SLA and leaf nitrogen, the phylogenetic signal is weak ($\lambda = 0.04$ and 0.15 ; $K = 0.0053$ and 0.0076), suggesting that prediction errors are largely independent of species relatedness. Although errors for height ($\lambda = 0.80$, $K = 0.018$) and leaf area ($\lambda = 0.56$, $K = 0.0067$) 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 on a 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 within-species variability for the traits). This suggests that the model is not simply regressing to a species-level mean but is sensitive to morphological cues in

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.

	Image	Geo	Depth	H	LA	SLA	LN	# Top ranks
Multi-Task	DinoV2	X	X	0.15 ± 0.00	0.31 ± 0.00	0.32 ± 0.00	0.14 ± 0.01	1
	BioCLIP	X	X	0.15 ± 0.01	0.3 ± 0.00	0.32 ± 0.01	0.15 ± 0.04	1
	DinoV2	SatCLIP	X	0.16 ± 0.02	0.27 ± 0.04	0.25 ± 0.02	0.11 ± 0.05	0
	DinoV2	GeoCLIP	X	0.17 ± 0.01	0.33 ± 0.01	0.32 ± 0.00	0.15 ± 0.02	3
	DinoV2	Climplicit	X	0.19 ± 0.01	0.32 ± 0.01	0.31 ± 0.01	0.16 ± 0.06	3
	BioCLIP	Climplicit	X	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	–

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) (trained on satellite imagery) for its ability to capture 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. 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 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 like 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 enhance 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 estab-

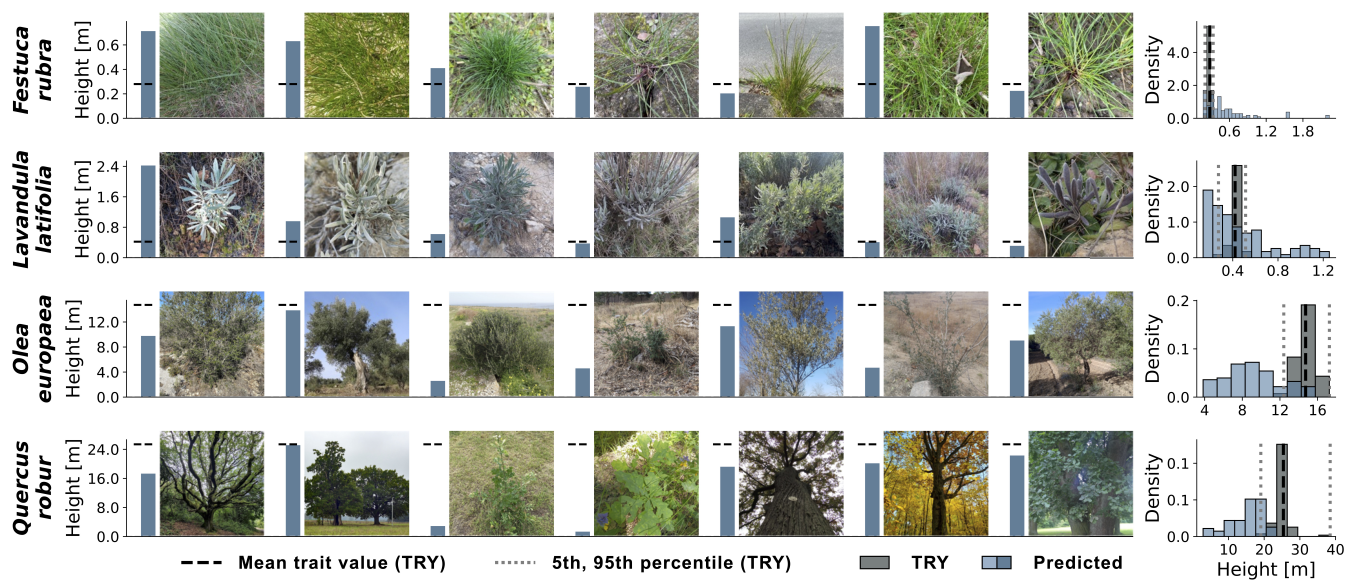


Figure 6: Intraspecific variation in predicted height for four species. Bar plots (left) show model predictions; histograms (right) show height distributions predicted from up to 100 images compared with up to 100 measurements from the TRY database.

lishes 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.; Jastrzebski, 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.; Hattings, 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.; Chytrý, 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.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

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.; Bönsch, 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.; Bönsch, 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.; Bruehlheide, 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.
- Paradis, E.; and Schliep, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3): 526–528.
- Sabatini, F. M.; Lenoir, J.; Hattab, T.; Arnst, E. A.; Chytrý, 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.; Bruehlheide, 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.

Zanaga, D.; Van De Kerchove, R.; De Keersmaecker, W.; Souverijns, N.; Brockmann, C.; Quast, R.; Wevers, J.; Grosu, A.; Paccini, A.; Vergnaud, S.; et al. 2021. ESA WorldCover 10 m 2020 v100. 2021.

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.

Supplemental Material

Data

Citizen Science Data Preparation The preparation of our citizen science training data involved several preprocessing steps to ensure quality and ecological representativeness. Using the geolocation metadata of the images, we removed samples from aquatic and urban areas based on the ESA WorldCover 10m v100 land cover map (Zanaga et al. 2021).

We weakly annotate these images with species-level trait values from the TRY database (Kattge et al. 2020). To reduce label noise due to sparsely represented species, we excluded species with fewer than three trait observations in the TRY database. Additionally, to further mitigate the impact of outliers, we computed the species-level median trait values after removing observations below the 5th and above the 99th percentiles.

Finally, we constructed the training and validation splits using an 80-20 partition, stratified by plant growth form (trees, shrubs, grasses) to preserve diversity across the subsets. Information on plant growth form was derived from the TRY database (Kattge et al. 2020, Trait-ID 4). For each species, we standardized the provided growth forms into 3 classes (Tree, Shrub, Grassland) and subsequently applied a majority vote per species (Lusk et al. 2025).

For benchmarking PlantTraitNet against other products, we jointly evaluate the model on a comprehensive evaluation set. This benchmark consists of a large-scale, uncurated set of 300K randomly sampled citizen science images, which were downloaded without any filtering for species name or quality, and about 80K images from the validation split of our curated training data.

Reference Data to Aid Uncertainty-Guided Data Cleaning

Our uncertainty-guided data cleaning strategy is grounded in the concept of a turning point—a stage during training when the model transitions from learning generalizable patterns to overfitting noisy labels (Lu and He 2022). Detecting this turning point is critical for effective uncertainty-aware data cleaning. However, in weakly supervised settings, where ground truth labels are absent or imprecise, identifying this transition reliably is challenging. To overcome this limitation, we curated a high-quality reference dataset consisting of plant images paired with accurate trait measurements. This dataset consists of smaller datasets distributed globally from various previous studies and a range of collaborators. Each collaborator contributed plant trait measurements and associated images from ecologically diverse locations, including Germany, La Palma, India, Australia, Panama, Canada, Indonesia, Switzerland, Portugal, and Namibia (see Figure 8). Detailed statistics on sample sizes and taxonomic diversity, quantified as the number of unique species, are reported in Table 3. Because different collaborators focused on specific traits, we maintain separate datasets per trait.

Table 3: Metadata of the curated plant reference dataset to aid uncertainty-guided data cleaning.

Trait	Units	#Data points	#Species
Plant height (H)	m	4478	452
Leaf Area (LA)	cm ²	2386	136
Specific Leaf Area (SLA)	mm ² /mg	1796	199
Leaf Nitrogen Content (LN)	mg/g	773	73

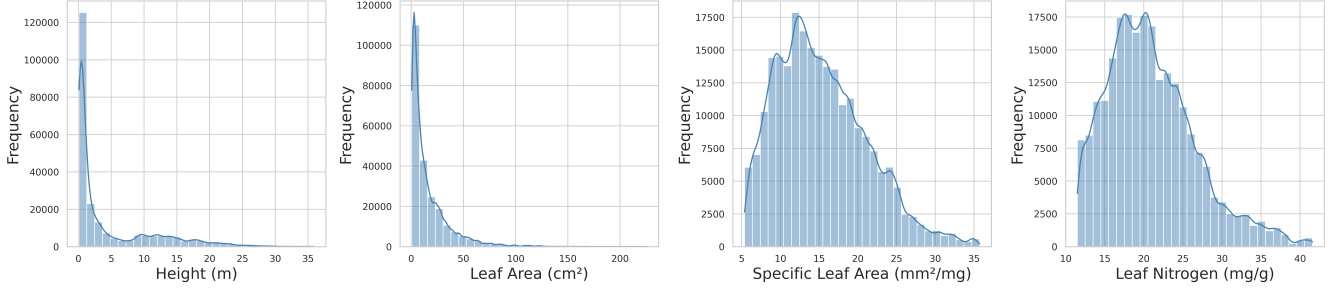


Figure 7: Distribution of median trait values at the species level in the weakly labeled citizen science training data.

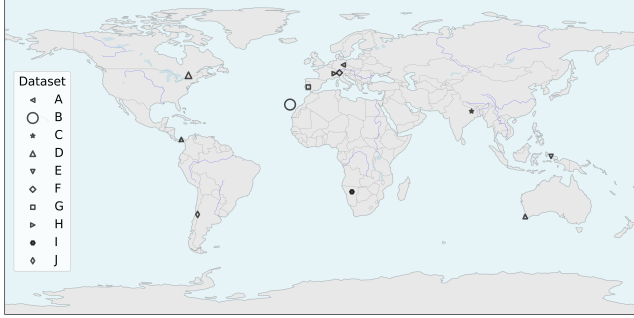


Figure 8: Locations of scientifically curated reference datasets. Marker symbols indicate distinct test data sources from Germany (A), La Palma (B), India (C), Australia (D), Panama (D), Canada (D), Indonesia (E), Switzerland (F), Portugal (G), Switzerland (H), Namibia (I), and Chile (J). Marker sizes are proportional to the corresponding dataset size.

Methodology

Uncertainty Estimation To capture predictive uncertainty across traits, we model the output distribution for each trait type, denoted as $m \in \{1, \dots, M\}$. Due to its long-tailed nature, Leaf Area (LA) is modeled using a Laplace distribution, parameterized by a predicted mean $\hat{\mu}_n^m$ and scale $b_n^m = \exp(\hat{s}_n^m)$, where $\hat{s}_n^m$ is the predicted log-scale.

For Height (H), Specific Leaf Area (SLA), and Leaf Nitrogen (LN), we assume a Gaussian distribution with predicted mean $\hat{\mu}_n^m$ and variance $\sigma_n^{2,m} = \exp(\hat{s}_n^m)$, where $\hat{s}_n^m$ is the predicted log-variance. The negative log-likelihood (NLL) is used as the training objective: for LA, the Laplace NLL is

$$\mathcal{L}_{\text{NLL}}^m = \frac{1}{N} \sum_{n=1}^N \left[\frac{|y_n^m - \hat{\mu}_n^m|}{\exp(\hat{s}_n^m)} + \hat{s}_n^m \right],$$

and for the remaining traits, the Gaussian NLL is

$$\mathcal{L}_{\text{NLL}}^m = \frac{1}{N} \sum_{n=1}^N \left[\frac{(y_n^m - \hat{\mu}_n^m)^2}{2 \exp(\hat{s}_n^m)} + \frac{1}{2} \hat{s}_n^m \right].$$

This formulation enables the model to learn both the central tendency and the predictive uncertainty for each trait. We visualize predicted log-variance across traits in Fig. 11

Uncertainty-Guided Data Cleaning Loop

To enable scalable data curation in the presence of label noise, we propose a two-step data cleaning strategy guided by model-predicted uncertainty and residual error.

Stage 1: Uncertainty-Aware Filtering We employ an iterative, uncertainty-aware filtering strategy to clean the training data while a model is being trained.

The model is first trained for one epoch on the full dataset. After this initial pass, we perform inference on the training set and compute predictive uncertainty across all four trait heads for each image. We then remove samples where the joint uncertainty falls within the top 5%, as we hypothesize these are likely to be visually ambiguous or noisy. Examples of such samples include poorly lit or occluded images, or those containing non-foliage content like twigs or fruit.

The model’s training then continues from its current state on this filtered dataset. This filtering–retraining cycle is repeated for up to two iterations or until the number of high-uncertainty samples drops below a predefined threshold.

Step 2: Residual-Aware Filtering As visualized in Figure 9, we observe that predictive uncertainty for traits like height increases for taller individuals, a phenomenon we hypothesize is due to heteroscedasticity. This is consistent with our observation that residuals for taller plants are generally larger than for smaller plants, leading the model to output a correspondingly higher uncertainty for taller plants. Consequently, a naive filtering approach that relies solely on high predictive uncertainty could disproportionately remove samples from larger growth forms (e.g. trees) and fail to distinguish genuinely mislabeled data from correctly labeled but uncertain predictions.

To address this, we introduce residual-aware filtering. For each trait, we identify a ‘turning point’, the training epoch after which performance on an external scientific reference set begins to decline, suggesting the onset of memorization. At this checkpoint, we compute normalized mean absolute error (nMAE) between predicted trait values and species-level medians. We hypothesize that samples exhibiting both high uncertainty (above the 95th percentile) and large residual error (above 50%) are likely mislabeled or unreliable. These samples are removed over four iterations, each followed by training from scratch.

We visualize examples of such filtered samples for height



Figure 9: Visualizations of predictive uncertainty for plant height during residual-aware filtering. Example images with high uncertainty (top) and low uncertainty (bottom) are shown.



Figure 10: Examples of images with high uncertainty and high residual error identified during residual-aware filtering for Height (m) (top) and Leaf Area (cm^2) (bottom). Predicted trait values are shown as bars, with species median values indicated by black dashed lines. For Height, juvenile or undeveloped individuals exhibit low predicted height but high residual error due to elevated species-level medians, leading to their removal. For Leaf Area, high residuals are associated with exotic ferns, images not focused on leaves, and small leaves from species with large median leaf areas.

and leaf area in Figure 10. For height, we often observe juvenile or undeveloped individuals with low predicted values but large residuals due to high species-level medians. For leaf area, examples for uncertainty samples are exotic ferns or unsharp photographs.

We continue the process until the number of samples with high-residual and high-uncertainty becomes negligible. In total, we remove approximately 500 species from an initial pool of 5,500. The resulting trait distributions are shown in Figure 7. While some borderline cases may be excluded, we argue that the volume of citizen science data compensates for a reduction of the total sample size and number of species. Furthermore, given the difficulty of verifying trait labels, particularly for biochemical traits such as specific leaf area (SLA) or leaf nitrogen, we prefer cautious removal over attempting label correction.

Experiment

Experimental Setup Data Preprocessing. In both the global plant kingdom and citizen science datasets, grassland species are the most frequently observed growth form, followed by shrubs, with tree species being the least represented. Blindly training on such imbalanced data would likely bias the model toward frequent grassland species,

reducing its ability to generalize to less common growth forms. To mitigate this, we stratified the data by growth forms and applied weighted sampling during training. Each batch was constructed to contain a balanced mix of grasses, shrubs, and trees, ensuring more equitable learning across growth forms. Lastly, all continuous target traits were normalized to the range $[0, 1]$ using MinMax scaling.

Training Procedure. The model is trained for 30 epochs using the AdamW optimizer with $\beta = (0.9, 0.999)$, and a weight decay of 5×10^{-5} . A cosine annealing learning rate schedule decays the learning rate from an initial value of 1×10^{-5} to a minimum of 5×10^{-6} . Training stability is further ensured by gradient clipping with a maximum norm of 1.0.

Results

Global Maps Trait Maps derived from Aggregated PlantTraitNet Predictions Figure 12 show global maps at 1 degree resolution obtained from aggregating predicted traits from PlantTraitNet on 300k citizen science images. In addition to the quantitative evaluation against vegetation surveys (sPlotOpen, next section), we found that the global patterns reproduce expected global trait patterns (Lusk et al. 2025; Wolf et al. 2022). For instance, for height and leaf

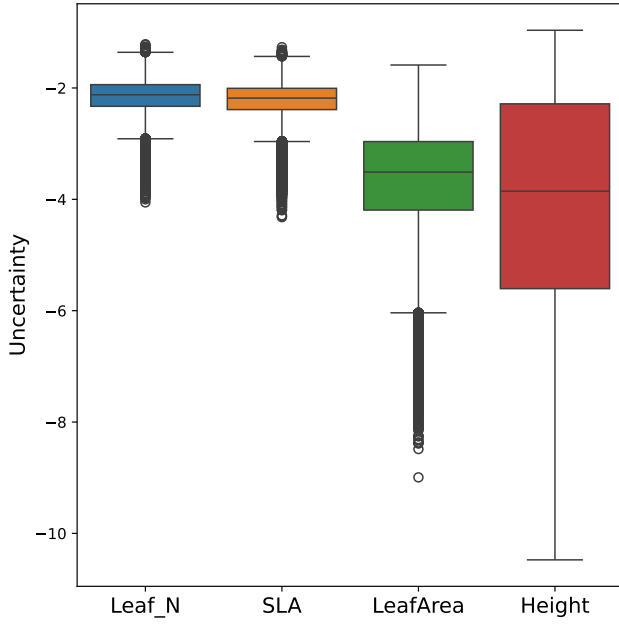


Figure 11: Visualization of the predicted log-variance values across all traits during the uncertainty-aware filtering process. Here, Leaf_N denotes leaf nitrogen, and SLA refers to specific leaf area.

area, we find expected longitudinal variation, such as large plants with big leaves towards the equator. Most robust plants (high specific leaf area and nitrogen content) are found in temperate and boreal zones.

Evaluating Global Trait Maps from PlantTraitNet with Vegetation Survey data (sPlotOpen) To empirically demonstrate the effectiveness of our uncertainty-guided data cleaning process, we compare model performance on raw (unfiltered) and refined dataset (Table 4). Performance is evaluated using R^2 , nMAE, and Pearson's r , with results averaged over three independent runs.

We observe that our data refinement process leads to consistent improvements in predictive performance. Notably, R^2 for SLA and LA increases from 0.23 to 0.27 and 0.30 to 0.34, respectively, demonstrating enhanced predictive reliability. For Leaf Nitrogen (LN), although the model's overall performance is limited (negative R^2), the refined dataset still yields a marginal improvement in correlation ($r = 0.50$ vs. 0.49).

The results, visualized in Figure 13, indicate that filtering noisy samples is a key step toward more robust and reliable trait predictions from our citizen science data.

Qualitative Assessment of Intraspecific Trait Variability Trait values within a species can vary across environmental gradients and hence an accurate representation of intraspecific variability can be important to create global trait maps. However, the weak annotations used for model training were based on species-level matching of citizen science photographs and trait values from the TRY database.

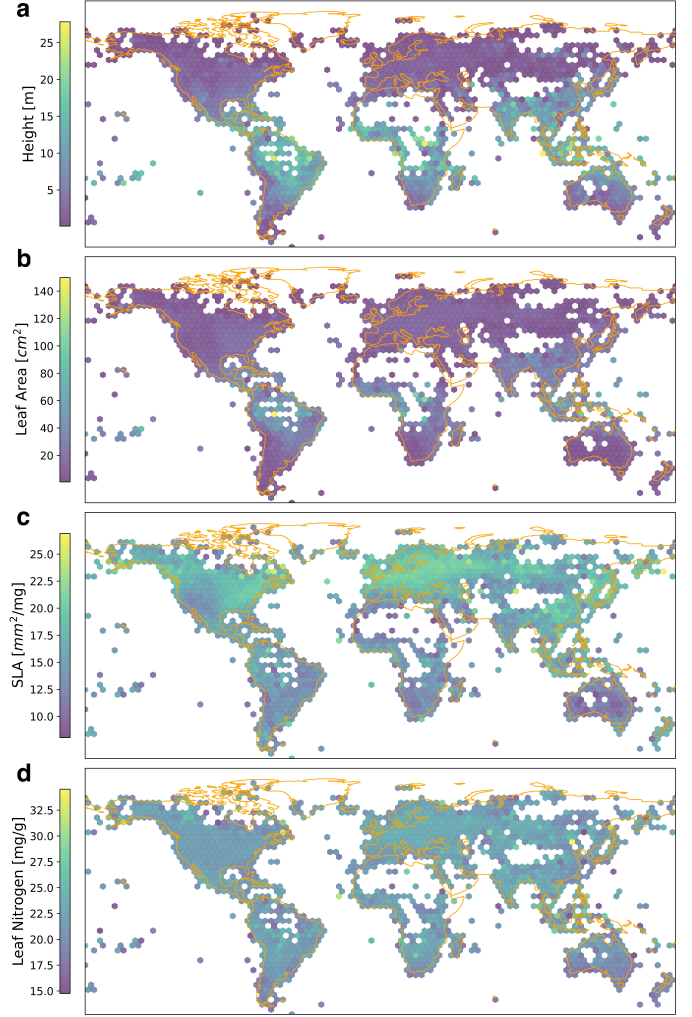


Figure 12: Global trait maps derived from spatially aggregating PlantTraitNet predictions at 1 degree resolution. The global trait maps comprise height (a), SLA (b), leaf area (c), and leaf Nitrogen (d) based on the spatial aggregation of citizen science photos from the validation dataset.

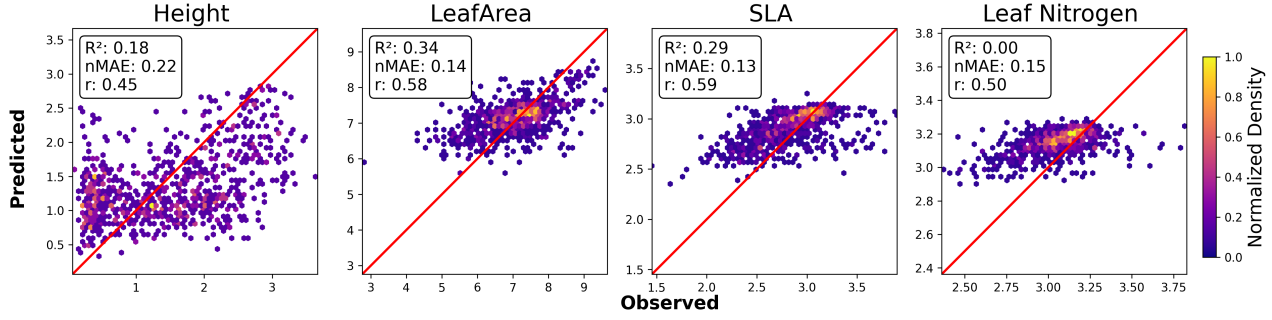


Figure 13: Global trait predictions obtained from PlantTraitNet against globally distributed vegetation survey data (sPlotOpen). Trait predictions of geotagged images are aggregated at a 1° spatial resolution using image geolocations (Predicted). These aggregated predictions are then compared with averaged community-weighted trait values derived from sPlotOpen (Observed).

Table 4: Comparison of PlantTraitNet with sPlotOpen community weighted trait means at 1° resolution. We report mean $\pm$ standard deviation across 3 runs with different random initializations. Raw refers to models trained with unfiltered citizen science data, while refined refers models based on the uncertainty-aware training.

Method	Metric	H	LA	SLA	LN
Ours (Raw)	$R^2 \uparrow$	0.19 ± 0.01	0.30 ± 0.04	0.23 ± 0.01	-0.16 ± 0.03
	nMAE $\downarrow$	0.22 ± 0.00	0.14 ± 0.00	0.14 ± 0.00	0.17 ± 0.00
	$r \uparrow$	0.45 ± 0.01	0.56 ± 0.01	0.59 ± 0.01	0.49 ± 0.01
Ours (Refined)	$R^2 \uparrow$	0.18 ± 0.00	0.34 ± 0.01	0.27 ± 0.02	-0.12 ± 0.14
	nMAE $\downarrow$	0.22 ± 0.0	0.14 ± 0.0	0.13 ± 0.00	0.17 ± 0.01
	$r \uparrow$	0.45 ± 0.00	0.57 ± 0.00	0.59 ± 0.01	0.50 ± 0.00

Accordingly, the model was not explicitly trained to capture within-species variability. To qualitatively evaluate if PlantTraitNet still captures intraspecific trait variation, we analyzed model predictions for ecologically diverse species across major growth forms, including grasses, shrubs, and trees. For each selected species, we conducted two complementary analyses:

1. We visualized predicted trait values for a small, held-out subset of seven individuals that exhibited notable visual variation in developmental stage, size or structure.
2. We compared the distribution of predicted trait values from up to 100 training images to the corresponding distribution of observed trait values from up to 100 samples in the TRY database.

This analysis provides insight into how well the model reflects trait variability within species, relative to aggregated trait observations. Our results, visualized in Figures 15 to 18, consistently show that the model learns to represent a wide range of intraspecific variation for traits like Height (H) and Leaf Area (LA). In these cases, the predicted trait ranges often exceeded the variation captured in the original TRY database observations. This seems plausible, given that trait observations are typically performed on healthy and adult plants. Compared to Leaf Area (LA) and Plant Height (H), for physiological traits such as Specific Leaf Area (SLA) and Leaf Nitrogen (LN), we observe little predicted variation. We hypothesize this is a positive outcome, as these leaf traits

are ecologically expected to show less intraspecific variability on the premise that interspecific trait variation generally exceeds intraspecific variation (Dong et al. 2020; Wright et al. 2017). For improved visibility, only predicted values between the 5th and 95th percentiles are shown.

Table 5: Phylogenetic signal metrics with corresponding p-values

Trait Error	K	K*	Lambda
H	0.018	0.015	0.801
	(p = 0.001)	(p = 0.001)	(p = 0.001)
LA	0.007	0.007	0.556
	(p = 0.064)	(p = 0.079)	(p = 0.001)
SLA	0.005	0.006	0.042
	(p = 0.441)	(p = 0.399)	(p = 0.003)
LN	0.008	0.008	0.150
	(p = 0.001)	(p = 0.001)	(p = 0.001)

Trait Prediction Error Along the Taxonomic Tree A phylogenetic tree was constructed from hierarchical taxonomic information (Kingdom to Species) using `as.phylo()` from the `ape` package (Paradis and Schliep 2019). Mean relative errors (MRE) were computed per species, normalized by the trait-specific error range, and matched with taxonomic metadata. In Figure 14, MREs are shown at the fam-

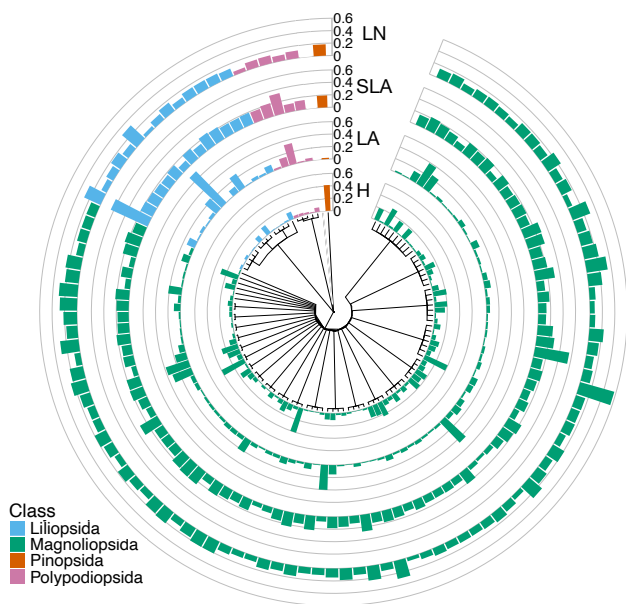


Figure 14: Mean relative prediction error (MRPE) computed on validation data at the family level, visualized along the taxonomic tree, for height (H), leaf area (LA), specific leaf area (SLA) and leaf nitrogen (LN).

ily and class levels, with no clear visual pattern of systematic bias. Moreover, we report standard metrics to quantify the phylogenetic signal in species-level prediction errors, calculated via the `phyloSignal()` function from the `phylosignal` R package. Pagel’s λ (Pagel 1999) measures the overall fit of the data to the phylogeny by scaling internal branches; values close to 1 indicate a strong phylogenetic structure, while values near 0 suggest independence from phylogeny. Statistical significance is assessed by comparing the fitted λ to a model with $\lambda = 0$. Blomberg’s K (Blomberg, Garland Jr, and Ives 2003) compares the observed trait variance across the tree to expectations under a Brownian motion model; values close to 1 indicate strong signal, and values $\ll 1$ suggest weak signal, especially among closely related species. K^* is a variation of K that is less sensitive to tree imbalance and branch length distortions. Together, these metrics offer complementary perspectives on phylogenetic structure at different evolutionary scales (see Table 5). Pagel’s λ captures broad-scale phylogenetic autocorrelation in residual covariance (Pagel 1999), while Blomberg’s K is more sensitive to fine-scale signal among closely related species (Blomberg, Garland Jr, and Ives 2003). For SLA and leaf nitrogen, the phylogenetic signal is weak ($\lambda = 0.04$ and 0.15 ; $K = 0.0053$ and 0.0076), suggesting that prediction errors are largely independent of species relatedness. Although errors for height ($\lambda = 0.80$, $K = 0.018$) and leaf area ($\lambda = 0.56$, $K = 0.0067$) show some phylogenetic autocorrelation, the consistently low K values indicate that even closely related species do not share systematic prediction biases. These results reveal that while slight patterns may emerge at the level of broader plant classes, there is no con-

sistent phylogenetic structure among closely related species. Altogether, this underscores that despite being trained on species-level annotations, the model exhibits strong generalizability and methodological robustness for predicting plant traits across the plant kingdom.

Ablation Study

Effect of Trait-Specific Loss Functions

Plant traits can differ strongly in their distributions, which in turn can affect the model training. To investigate the influence of loss function choice on trait prediction, we compare Gaussian and trait-specific Laplace loss formulations, as shown in Table 6. Given the long-tailed distributions observed for certain traits, notably leaf area (LA) and height (H) (see Figure 7), we hypothesize that the heavier-tailed Laplace loss may improve model robustness and calibration. Applying the Laplace loss to LA results in performance improvements compared to the Gaussian loss for both LA and height, while performance for SLA and LN remains similar. However, applying the Laplace loss to both LA and H decreases performance for height. This outcome may be related to the height distribution being heavily influenced by grass species, causing imbalance. Since we mitigate this imbalance through stratified sampling across plant functional types such as grasses, shrubs, and trees, the Gaussian loss appears more appropriate for modeling height under these conditions. In summary, our results suggest that trait-specific loss functions can benefit LA prediction. Based on this analysis, we adopt a hybrid approach in the final model, using the Laplace loss for LA and Gaussian losses for the other traits. This approach is associated with improved predictive performance across traits and underlines that model training requires careful consideration of multivariate ecological complexity.

Effect of Token Pooling Strategy on Trait Prediction Performance While the classification token (‘[CLS]’) is a common choice for a global image representation, the individual patch tokens of vision transformers contain rich, spatially-aware information that can be more beneficial for downstream tasks (Caron et al. 2021). We therefore perform an ablation study to determine the optimal token representation for our trait prediction task (Table 7).

We observe that for DinoV2 patch token pooling improves R^2 scores for SLA (from 0.30 to 0.32) and LN (from 0.11 to 0.14), while Height performs better with the classification token (R^2 of 0.18 vs. 0.15), and Leaf Area shows no change. Based on the net gains for SLA and LN, we use patch token pooling in our final configuration.

For BioCLIP, we conduct a similar ablation and observe that the classification token consistently performs better on average. We hypothesize that this may be due to the contrastive training objective of BioCLIP, which aligns the classification token with the projected textual representation in a joint embedding space, making it a more robust representation for this model.

Effect of Embedding Dimensionality on Multimodal Trait Prediction The choice of embedding dimensional-

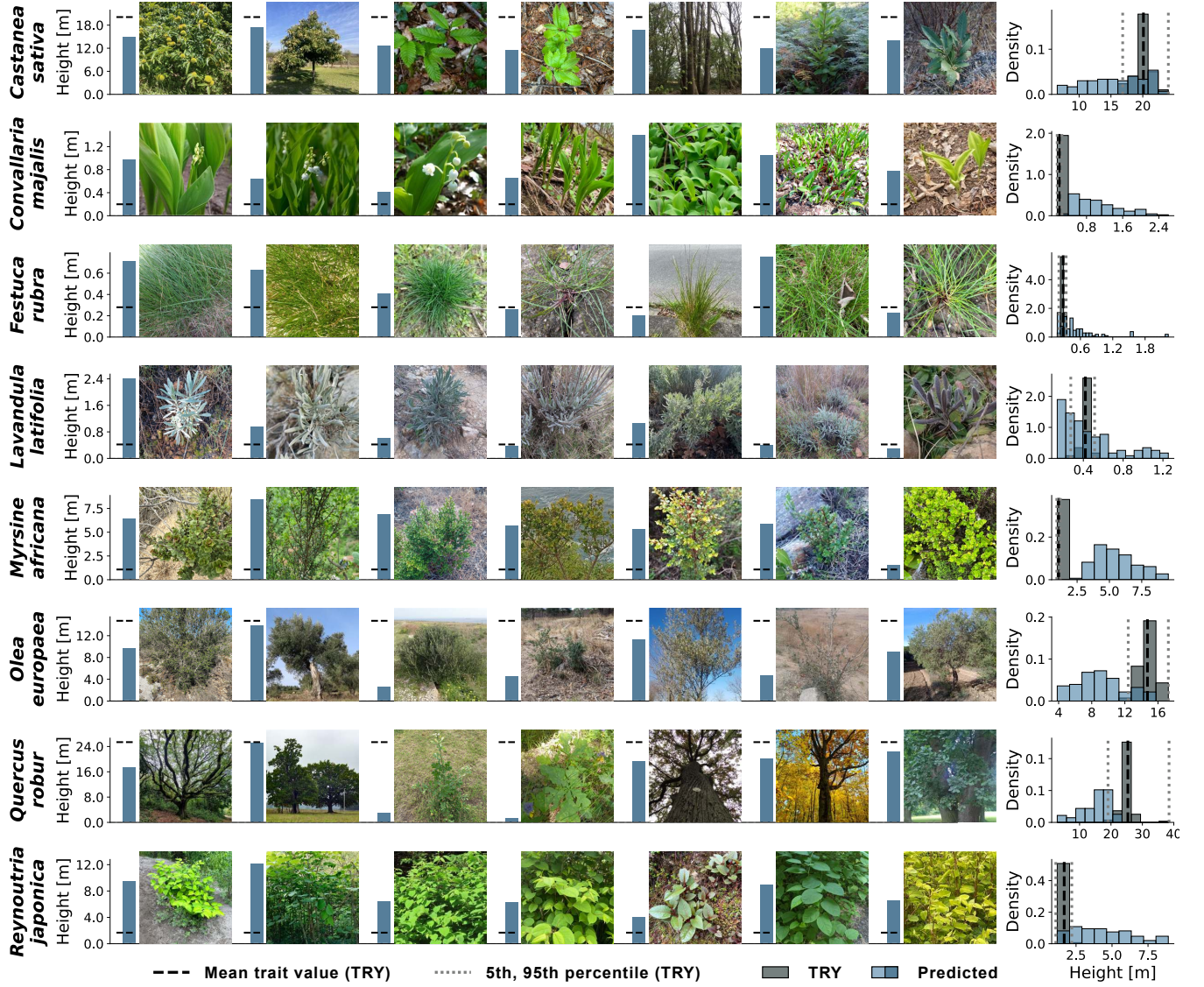


Figure 15: Intraspecific variation in predicted plant height compared to TRY-derived trait means.

Table 6: Impact of trait-specific loss functions on predictive performance. Reported values are mean $R^2 \pm$ one standard deviation across three independently initialized training runs for each loss configuration.

Method	H	LA	SLA	LN
Gaussian	0.18 ± 0.02	0.31 ± 0.01	0.31 ± 0.01	0.18 ± 0.01
Laplace (LA + H)	0.10 ± 0.01	0.32 ± 0.01	0.32 ± 0.02	0.18 ± 0.03
Laplace (LA)	0.19 ± 0.02	0.32 ± 0.01	0.31 ± 0.02	0.18 ± 0.05

Table 7: Integrating embeddings from patch token vs. classification token (CLS) in image encoders. We report $R^2 \pm$ 1 standard deviation across 3 runs.

Image Encoder	H	LA	SLA	LN
DINOv2 (CLS token)	0.18 ± 0.01	0.31 ± 0.00	0.30 ± 0.01	0.11 ± 0.01
DINOv2 (patch mean)	0.15 ± 0.00	0.31 ± 0.00	0.32 ± 0.00	0.14 ± 0.01
BioCLIP (CLS token)	0.15 ± 0.01	0.30 ± 0.00	0.32 ± 0.01	0.15 ± 0.04
BioCLIP (patch mean)	0.17 ± 0.01	0.30 ± 0.01	0.31 ± 0.01	0.12 ± 0.05

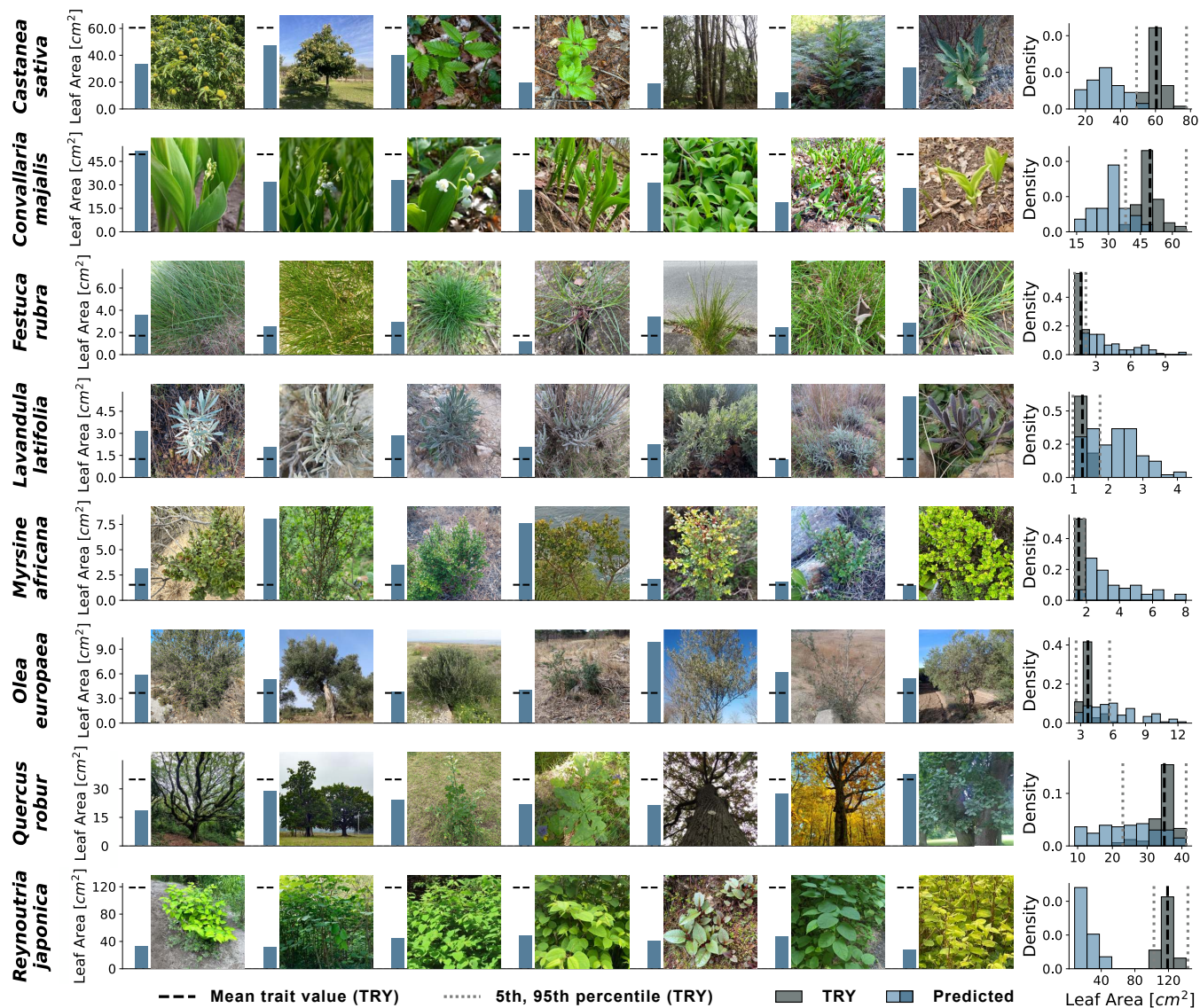


Figure 16: Intraspecific variation in predicted leaf area compared to TRY-derived trait means.

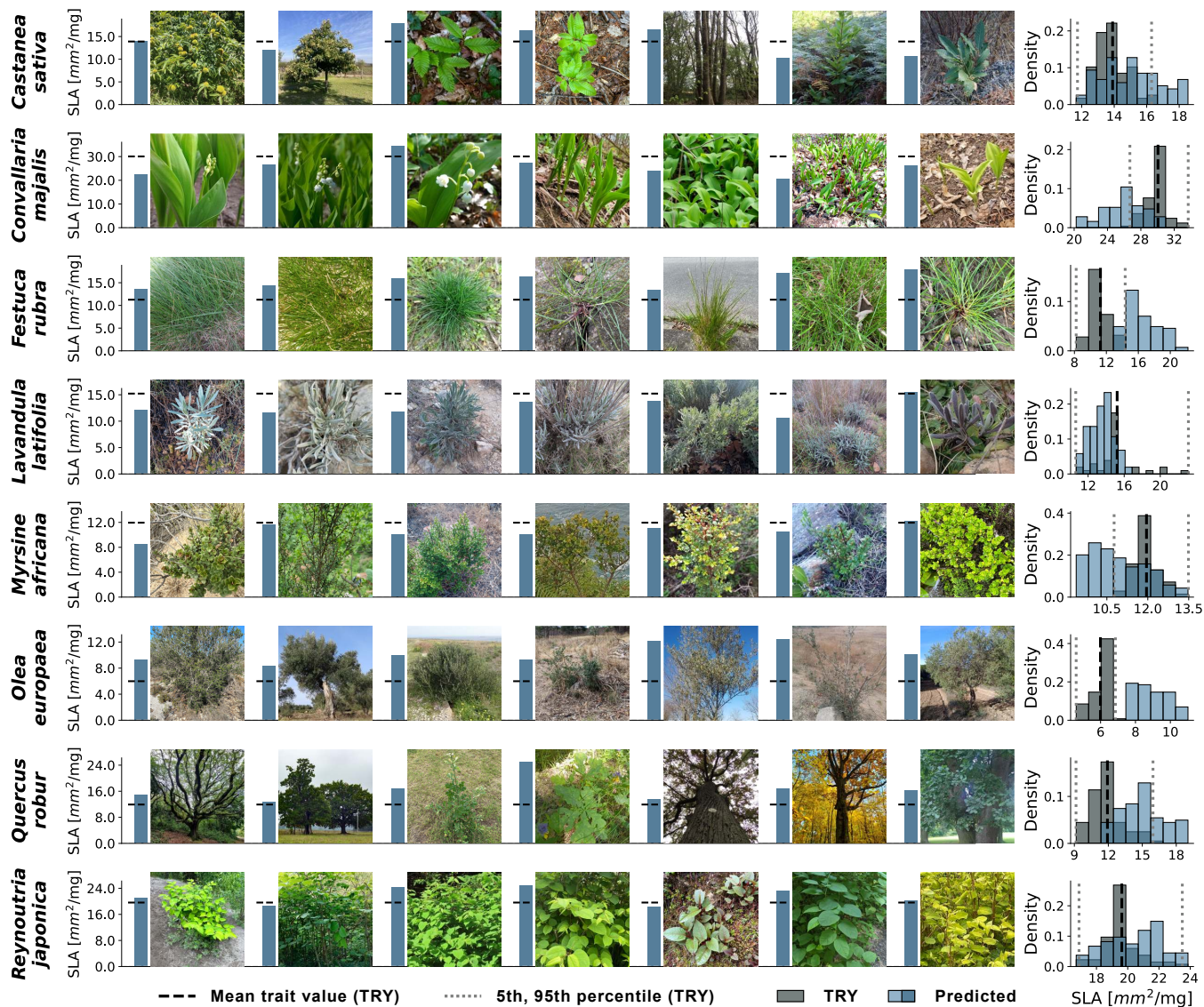


Figure 17: Intraspecific variation in predicted specific leaf area compared to TRY-derived trait means.

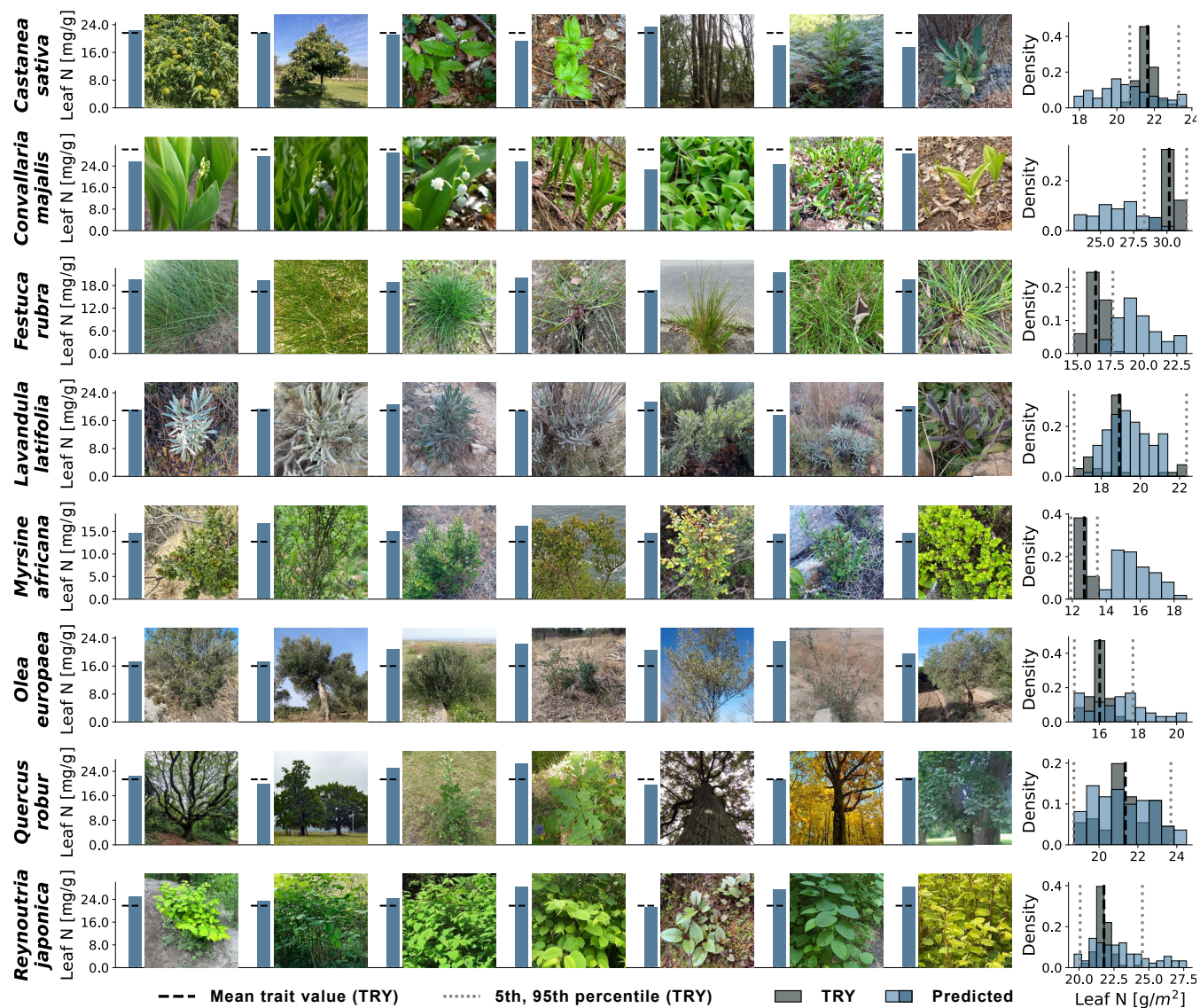


Figure 18: Intraspecific variation in predicted leaf nitrogen compared to TRY-derived trait means.

Table 8: Effect of embedding size for each modality and the multimodal backbone. We perform ablations over different embedding dimensions: [512, 768] for image encoders, [256, 512, 1024] for geospatial encoders, and [768, 1024] for the multimodal fusion layer. We report $R^2 \pm 1$ standard deviation across 3 runs.

Image	Geo	Multimodal Backbone Dim	H	LA	SLA	LN
DinoV2(patch) + Climplicit						
512	256	768	0.18 ± 0.01	0.30 ± 0.01	0.30 ± 0.01	0.13 ± 0.04
768	256	768	0.19 ± 0.01	0.32 ± 0.01	0.31 ± 0.01	0.16 ± 0.06
768	256	1024	0.18 ± 0.02	0.30 ± 0.03	0.29 ± 0.01	0.12 ± 0.02
768	1024	768	0.18 ± 0.00	0.31 ± 0.01	0.29 ± 0.02	0.12 ± 0.03
768	1024	1024	0.17 ± 0.01	0.30 ± 0.03	0.28 ± 0.01	0.13 ± 0.03
768	512	768	0.18 ± 0.01	0.32 ± 0.02	0.29 ± 0.01	0.14 ± 0.05
768	512	1024	0.17 ± 0.01	0.31 ± 0.01	0.3 ± 0.01	0.14 ± 0.02
DinoV2(patch) + SatCLIP						
768	256	768	0.16 ± 0.02	0.27 ± 0.04	0.25 ± 0.02	0.11 ± 0.05
768	256	1024	0.13 ± 0.03	0.26 ± 0.03	0.25 ± 0.01	0.13 ± 0.01
DinoV2(patch) + GeoCLIP						
768	256	768	0.17 ± 0.01	0.33 ± 0.01	0.32 ± 0.00	0.15 ± 0.01
768	256	1024	0.16 ± 0.01	0.33 ± 0.00	0.31 ± 0.01	0.15 ± 0.02
768	512	768	0.17 ± 0.0	0.32 ± 0.01	0.32 ± 0.00	0.13 ± 0.02
768	512	1024	0.17 ± 0.0	0.3 ± 0.0	0.31 ± 0.01	0.13 ± 0.02
DinoV2(patch) + Climplicit(256-D) + Depth(768-D)						
768	256	768	0.17 ± 0.02	0.32 ± 0.02	0.32 ± 0.01	0.16 ± 0.03
768	256	1024	0.19 ± 0.02	0.32 ± 0.01	0.31 ± 0.02	0.18 ± 0.05

ity and model capacity is a critical design decision that balances computational efficiency with representational power. To find the optimal configuration for our model, we conducted a detailed ablation study to systematically evaluate the impact of embedding dimensions across our primary modalities and the multimodal fusion backbone (Table 8).

We vary (i) the image embedding dimension (512 vs. 768), (ii) the geolocation embedding dimension (256, 512, and 1024), and (iii) the hidden size of the multimodal fusion backbone (768 vs. 1024).

We observe that using a 768-dimensional image embedding from DINOv2, a 256-dimensional geolocation embedding, and a 768-dimensional multimodal fusion backbone tends to yield strong performance across traits. This trend is consistent across the tested geolocation encoders: Climplicit (Dollinger et al. 2025), SatCLIP (Klemmer et al. 2025), and GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023). Among these, Climplicit performs best in our experimental setup.

When including depth as an additional modality, performance improves when the multimodal backbone is increased to 1024 dimensions. We hypothesize that this benefit may arise from the increased representational capacity required to integrate the added modality. Based on these findings, we use a 768-dimensional multimodal backbone when the model includes only image and geolocation inputs, and a 1024-dimensional backbone when depth is incorporated.