OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

# Unsupervised Representation Learning for Leaf Trait Prediction and Spectral Novelty Detection Across 8 EcoSIS Datasets

Vidyut Krishnan
vidyut.krishnan12@gmail.com

## ABSTRACT

Non-destructive monitoring of plant traits through leaf-level reflectance spectroscopy is essential for understanding the functioning of global ecosystems and biodiversity across different biomes. Nevertheless, there are difficulties in combining results of diverse field experiments because of differences in hardware, spectral resolution, and sampling methods. Most workflows do not enable creation of coherent representations without extensive annotation of traits in diverse plant communities. In order to solve this problem, we have combined 8 EcoSIS datasets representing both temperate and tropical communities by bringing 13,549 leaf-level spectra to 201 consistent bands (400-2400 nm) by means of Savitzky-Golay smoothing and SNV normalization and then trained an unsupervised deep autoencoder to obtain 192-dimensional embeddings. When evaluating representations using group-based cross-validation in order to avoid taxa leakage, we found out that raw spectra yielded the best results for structural and water traits (R2 up to 0.92 for leaf mass per area and equivalent water thickness), whereas 192-D embeddings were comparable or better for estimating pigments (R2 = 0.94 - 0.96 on the Angers dataset). In the third stage, the latent space allowed us to perform the discovery task that enables the estimation of species-level spectral novelty based on scan frequency. This discovery task ranked Carex Richardsonii as the most spectrally novel species, with a score of 0.9525.

## INTRODUCTION

Leaf reflectance spectroscopy is a commonly used technique for investigating plant biochemistry and physiology using the reflection and absorption properties of leaves across the electromagnetic spectrum. [1] The absorption of specific wavelengths by certain pigments such as chlorophyll and carotenoids, and the reflection or transmission of other wavelengths in accordance with the leaf's chemical composition, water content, and structure produces these spectra. [4, 5] This spectrum contains detailed information on plant characteristics and environment. [6-10] Because this type of analysis is non-destructive and highly sensitive, it is widely used in plant ecology, agriculture, and remote sensing. [5, 6, 11] It allows researchers to estimate various traits, such as nitrogen content, leaf mass per area, and pigment concentration non-invasively. [8, 9]

Reflectance spectroscopy utilizes the physical interactions between radiation and matter, where the structure and chemistry of matter become responsible for the reflectance and scattering of electromagnetic

energy. [1, 5] In the visible range (400 - 700 nm), pigments are mainly responsible for the differences in the spectra. [5, 10] In the near-infrared range (700-1300 nm), leaf inner structure defines the reflectance. [5, 10] In the shortwave infrared range (1300 - 2500 nm) water and biochemically defined molecules, such as cellulose, lignin, and proteins, define the presence of unique absorption bands. [5, 10] This allows defining the optical fingerprints of the physiological and biochemical composition of the leaves. [4, 6] Studies have established the connection between the results of optical spectroscopy measurements and the processes of photosynthesis and stress. [7] Hyperspectral sensing technology has revolutionized the application of leaf reflectance spectroscopy. [6, 11, 25] Although multispectral systems are only able to record a limited number of broad bands of wavelengths, hyperspectral sensors can detect several hundreds, and sometimes even thousands of narrow contiguous wavelength bands making it possible to detect minor features related to plant biochemistry and physiology. [6, 25] Hyperspectral sensors help discriminate between closely related plant species, determine nutrient deficiencies, measure the level of drought, and estimate biodiversity of a landscape. [11-14]

The combination of machine learning and spectroscopy reflects a broader shift toward data-driven scientific discovery. [1-3] Models based on ML algorithms have been applied in plant science for predicting foliar nitrogen, phosphorus, and metal content, for species classification, for disease detection, as well as for estimating stomatal conductance and photosynthetic potential. [8, 15, 16] Notably, the ML-based models allow generalization for analyzing the spectral data obtained in different environmental settings and in different ecosystems. [2] Studies [1] have illustrated how machine learning is revolutionizing spectroscopy through enhanced computational capabilities, increased number of synthetic training sets, and high-throughput analysis of experimental and theoretical spectra.

The primary problem associated with applying artificial intelligence to biological spectroscopy concerns generalization, the capability of the model to be accurate on different datasets which were obtained through different sampling protocols and instruments in different environments. [2] Overfitting to one dataset makes a model unable to work properly on another one. Artificial intelligence generalization in the area of molecular and biological datasets is focused on the necessity of cross-validation and independent testing, as well as the importance of having a representative training dataset. [2] This is crucial in ecological spectroscopy since spectra depend on species presence, lighting conditions, etc.

In spite of the increasing amount of open leaf spectroscopy datasets which can be obtained from EcoSIS, there is no common cross-dataset latent space. Furthermore, no spectral novelty ranking accounts for confidence: existing outlier detectors flag unusual points but do not weigh them by the number of supporting scans, leaving small-sample species over-represented in discovery rankings. This study addresses three objectives: (1) to train a self-supervised autoencoder on pooled EcoSIS spectra and evaluate whether the learned 192-dimensional embedding outperforms raw spectra and PCA for downstream trait prediction across eight datasets; (2) to characterize cross-domain structure in the latent space by measuring how separable the eight datasets remain in raw and embedded form, and how well the pooled encoder reconstructs each constituent dataset; and (3) to combine embedding-space novelty

detection with scan-count-weighted confidence to identify spectrally unusual candidate species for targeted follow-up.

By combining leaf reflectance spectroscopy with machine learning, this project seeks to better understand how hyperspectral signatures relate to various plant traits across ecological contexts. Spectra from many leaf-level datasets were analyzed to determine if ML models can be trained to accurately predict structural and biochemical traits. The resulting models provide a foundation for scalable, non-destructive trait prediction and estimation, and contribute to efforts in ecological monitoring, biodiversity assessment, and sustainable resource management. [11, 18, 22, 23] More broadly, this work shows how advances in spectroscopy, machine learning, and plant science are enabling us to reveal new insights into the biochemical and physiological diversity of the natural world.

## METHODS

### Pipeline overview

This project developed a three-stage hyperspectral leaf analysis pipeline **(Figure 1)** that combines supervised trait prediction, novelty ranking, and representation learning. Stage 1 trains an unsupervised autoencoder on the pooled spectra and evaluates the resulting embeddings against raw spectra for per-dataset trait prediction; Stage 2 characterizes cross-domain structure through pairwise dataset discriminability and a leave-one-dataset-out evaluation; and Stage 3 ranks spectrally novel candidate species using confidence-weighted novelty scores. The Methods, Results, and Discussion are organized by these three stages, and the input, output, objective, data partition, and evaluation metric of each stage are consolidated in **(Table III)**.

| Dataset | n spectra | n samples | n species | Biome / setting | Instrument | Spectral range | Traits available | Split group | n after outlier filtering |
|---|---|---|---|---|---|---|---|---|---|
| AQGRAD | 190 | 190 | 13 | Water-affinity gradient, Mantua lakes, Italy | Field leaf-clip spectrometer | 400-2400 nm | Species labels and foliar traits | `Species` | 190 |
| Talladega 2018 | 156 | 156 | 10 | Talladega National Forest, Alabama, USA | Field leaf-clip spectrometer | 400-2400 nm | Species labels and foliar traits | `Sample_ID` | 156 |
| Angers 2003 | 276 | 276 | 43 | Laboratory leaf optical properties database | Laboratory spectrometer / integrating sphere | 400-2400 nm | Carotenoid, chlorophyll a, chlorophyll a+b, chlorophyll b, EWT, LMA, SPAD, structure parameter | `Sample_#` | 276 |
| Blackhawk Island WI | 3248 | 3248 | 10 | Seasonal fresh leaves, Blackhawk Island, Wisconsin, USA | Field leaf-clip spectrometer | 400-2400 nm | `chl_a_mmol_m2`, `lma_g_m2` | `Species` | 3248 |
| CABO 2018-2019 v2 | 7821 | 7821 | 103 | Broad multispecies leaf-level trait library | Field leaf-level spectrometer | 400-2400 nm | LMA, EWT, LDMC, C/N ratio, pigments, cellulose, hemicellulose, lignin, solubles, and multiple elemental traits | `Sample_ID` | 7821 |
| NGEE Tropics Panama 2016 | 855 | 855 | 48 | Tropical leaf reflectance, Panama | Field leaf-clip spectrometer | 400-2400 nm | CNratio, Cmass_pc, H20_pc, LMA_gDW_m2, LWP_bar, Nmass_pc, SLA_cm2_gDW | `Sample_ID` | 855 |
| LOPEX93 | 660 | 325 | 46 | Laboratory leaf optical properties database | Laboratory spectrometer / integrating sphere | 400-2400 nm | Chlorophyll a+b, EWT, LMA, protein, cellulose, carotenoid | `Species` | 640 |
| Sierra Nevada 2023 | 363 | 363 | 15 | Field spectra, Sierra Nevada, California, USA | Field leaf-clip spectrometer | 400-2400 nm | `LMA_g_m2`, `LWC_percent`, `Chlorophyll_mg_m2` | `sample_number` | 363 |

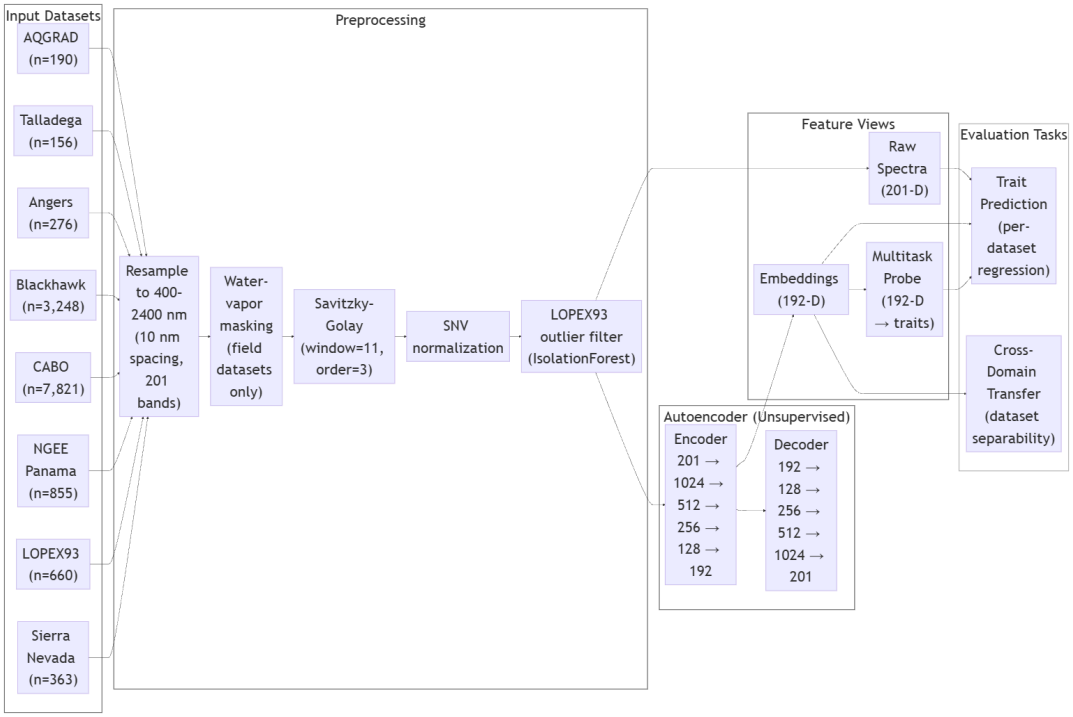



**Figure 1:** Pipeline Overview

metadata were available without the need for importing metadata files locally.

**Table I:** Overview of Datasets

| Dataset | n spectra | n samples | n species | Biome / setting | Instrument | Spectral range | Traits available | Split group | n after outlier filtering |
|---|---|---|---|---|---|---|---|---|---|
| AQGRAD | 190 | 190 | 13 | Water-affinity gradient, Mantua lakes, Italy | Field leaf-clip spectrometer | 400-2400 nm | Species labels and foliar traits | `Species` | 190 |
| Talladega 2018 | 156 | 156 | 10 | Talladega National Forest, Alabama, USA | Field leaf-clip spectrometer | 400-2400 nm | Species labels and foliar traits | `Sample_ID` | 156 |
| Angers 2003 | 276 | 276 | 43 | Laboratory leaf optical properties database | Laboratory spectrometer / integrating sphere | 400-2400 nm | Carotenoid, chlorophyll a, chlorophyll a+b, chlorophyll b, EWT, LMA, SPAD, structure parameter | `Sample_#` | 276 |
| Blackhawk Island WI | 3248 | 3248 | 10 | Seasonal fresh leaves, Blackhawk Island, Wisconsin, USA | Field leaf-clip spectrometer | 400-2400 nm | `chl_a_mmol_m2`, `lma_g_m2` | `Species` | 3248 |
| CABO 2018-2019 v2 | 7821 | 7821 | 103 | Broad multispecies leaf-level trait library | Field leaf-level spectrometer | 400-2400 nm | LMA, EWT, LDMC, C/N ratio, pigments, cellulose, hemicellulose, lignin, solubles, and multiple elemental traits | `Sample_ID` | 7821 |
| NGEE Tropics Panama 2016 | 855 | 855 | 48 | Tropical leaf reflectance, Panama | Field leaf-clip spectrometer | 400-2400 nm | CNratio, Cmass_pc, H20_pc, LMA_gDW_m2, LWP_bar, Nmass_pc, SLA_cm2_gDW | `Sample_ID` | 855 |
| LOPEX93 | 660 | 325 | 46 | Laboratory leaf optical properties database | Laboratory spectrometer / integrating sphere | 400-2400 nm | Chlorophyll a+b, EWT, LMA, protein, cellulose, carotenoid | `Species` | 640 |
| Sierra Nevada 2023 | 363 | 363 | 15 | Field spectra, Sierra Nevada, California, USA | Field leaf-clip spectrometer | 400-2400 nm | `LMA_g_m2`, `LWC_percent`, `Chlorophyll_mg_m2` | `sample_number` | 363 |

## Dataset sources

The pipeline trains on eight public EcoSIS leaf spectra datasets **(Table I)** including AQGRAD, Angers 2003, Blackhawk Island WI, CABO v2, LOPEX93, NGEE Panama 2016, Sierra Nevada 2023, and Talladega 2018 26-33. [26-33] Data was imported through the EcoSIS API, and spectra along with metadata were available without the need for importing metadata files locally.

Traits present in two or more datasets were harmonized to common SI units **(Table II)**. Both harmonized and dataset-specific traits were evaluated; unharmonized traits were retained and analyzed within their respective datasets. Harmonized traits included leaf mass per area (g/cm2), chlorophyll a+b (µg/cm2), equivalent water thickness (g/cm2), carotenoid (µg/cm2), leaf water content (%), and C/N ratio (dimensionless). Additionally, any dataset with fewer than 10 valid

| Harmonized trait | Target unit | Datasets |
|---|---|---|
| `LMA_g_cm2` | g/cm² | LOPEX93, Angers, CABO, NGEE Panama, Sierra Nevada, Blackhawk |
| `Chlorophyll_a_b_ug_cm2` | µg/cm² | LOPEX93, Angers |
| `EWT_g_cm2` | g/cm² | LOPEX93, Angers, CABO |
| `Carotenoid_ug_cm2` | µg/cm² | LOPEX93, Angers |
| `LWC_percent` | % | Sierra Nevada, NGEE Panama |
| `CN_ratio` | ratio | CABO, NGEE Panama |

**Table II:** Harmonization of Traits

observations for a given harmonized trait was excluded from the analysis.

**Preprocessing**

Spectra were smoothed using the Savitzky-Golay method (window length 11, polynomial order 3) at native resolution, scaled using standard normal variate normalization (SNV), and for field datasets (Sierra Nevada, Talladega, NGEE Panama), the ~1350-1450 nm and ~1800-1950 nm water-vapor absorption regions were masked and replaced by linear interpolation from adjacent valid bands. Finally, spectra were interpolated to a shared 400-2400 nm wavelength space at 10 nm spacing, yielding 201 bands per scan. **(Figure 2)** Note that raw spectra were also scaled using SNV, not just the embeddings.

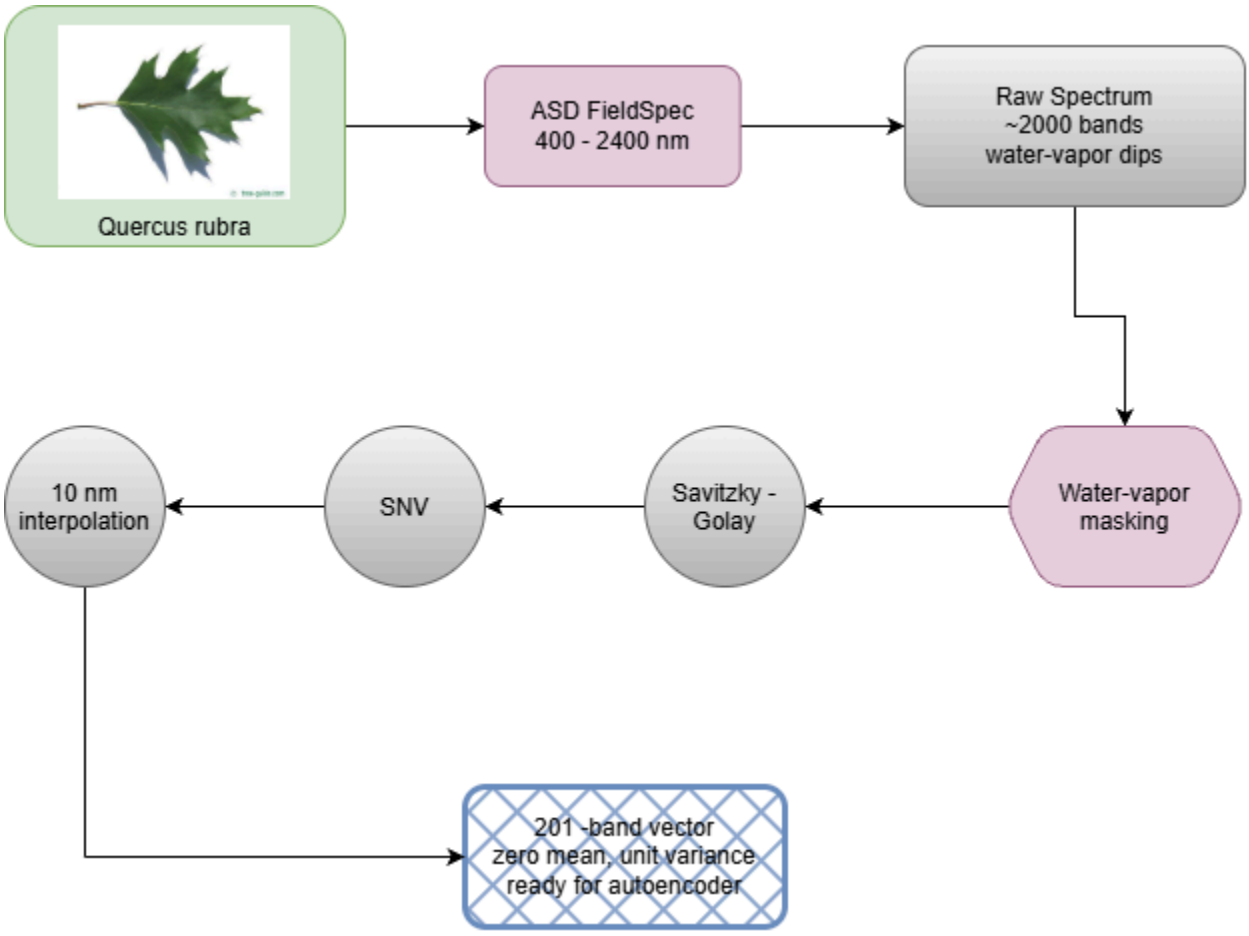



**Figure 2:** Preprocessing of a Quercus rubra (Northern red oak) sample. Note that the image provided is

not an actual picture of a sample, and is an image of a Red Oak leaf. Additionally, LOPEX93 outliers were filtered using Isolation Forest (Contamination = 0.03, n_estimators = 100, random_state = 42), reducing the influence these outliers had on the learned representation. We decided to do this since analysis of the LOPEX93 dataset showed various anomalies such as NaN readings. Datasets with multiple scans per biological sample were grouped during downstream evaluation to reduce leakage between train and test splits. Sample-level aggregation was used for LOPEX93 trait prediction when appropriate.

## Data partitioning, evaluation protocol, and leakage considerations

Because the three stages use the data differently, the partitioning used in each stage is stated explicitly here and summarized in Table V. The Stage-1 autoencoder was trained on all 13,549 pooled, preprocessed spectra; no spectra were withheld from autoencoder training, and early stopping monitored the training loss rather than a validation loss. The only exception is the latent-dimension sweep (Table III), which used a random 90%/10% train/validation split. Consequently, the encoder has seen every spectrum that is later used as a test input in the downstream, cross-domain, and novelty analyses. Because the autoencoder never receives trait labels, this does not leak label information into any supervised evaluation; however, it does mean that all embedding-based results are transductive with respect to the spectra themselves and should not be read as estimates of performance on spectra unseen by the encoder.

For trait prediction (the Stage-1 downstream evaluation), models are fit separately within each dataset. Within a dataset, rows with at least one trait value were split once into 75% training and 25% test rows using a group-aware shuffle split (random seed 42), where the grouping variable is the "Split group" column of Table I (biological sample identifier where available, otherwise species), so that replicate scans of the same sample, or spectra of the same species where sample identifiers are unavailable, never appear on both sides of the split. Model selection used an inner group k-fold cross-validation (up to five folds) on the training rows only, and the reported test $R^2$ values are computed on the untouched 25% test rows; the "CV $R^2$" values quoted for CABO refer to this inner cross-validation on the training rows. If the common split left fewer than 12 training or 5 test rows for a particular trait, the split was re-drawn for that trait with the same procedure. Because trait prediction is performed within each dataset, differences in dataset size (Table I) affect the number of training examples available to each dataset but do not create cross-dataset class imbalance.

For Stage 2, all spectra of both datasets in a pair were used without subsampling or class weighting; each dataset was split 80%/20% into training and test spectra by a seeded random permutation within each dataset, and domain-classification accuracy was computed on the pooled 20% test spectra. In the leave-one-dataset-out evaluation, up to 500 randomly drawn spectra of the held-out dataset were classified against all spectra of the remaining seven datasets under the same 80%/20% protocol, and reconstruction error was computed on the same held-out subset using the pooled Stage-1 encoder without retraining. The class sizes in these experiments are therefore highly unequal (for example, 7,821 CABO spectra versus 190 AQGRAD spectra in the pairwise task, and 13,359 "other" spectra versus 190 held-out AQGRAD spectra in the leave-one-dataset-out task), and the reported metric is standard (unbalanced)

accuracy. The consequences of this imbalance for interpreting Stage 2 are addressed in the Results and Discussion. For Stage 3, Isolation Forests are fit and min-max normalized within each dataset, so that the largest dataset (CABO, 7,821 of the 13,549 pooled spectra) cannot dominate the cross-dataset ranking, and species-level scores are down-weighted by scan count to limit the influence of species represented by very few spectra.

| **Stage** | **Input** | **Output** | **Objective** | **Data partition** | **Metric(s)** |
|---|---|---|---|---|---|
| Stage 1 – Autoencoder | 13,549 pooled SNV-normalized spectra (201 bands, 400–2400 nm) | 192-D embedding per spectrum; reconstructed spectrum | Unsupervised: minimize reconstruction MSE | All pooled spectra used for training (no held-out spectra); 90%/10% split only in the latent-dimension sweep | Reconstruction MSE; kNN batch-mixing entropy; Silhouette coefficient |
| Stage 1 – Trait prediction (three feature views) | Per dataset: (i) raw 201-band SNV spectrum; (ii) 192-D embedding; (iii) 192-D embedding fed to the multitask probe. Examples – CABO v2: 7,821 spectra; Angers 2003: 276 spectra; LOPEX93: 325 sample-aggregated spectra | Predicted trait value(s) per spectrum/sample. Examples – CABO v2: LMA, EWT, carbon mass per area, solubles per area, and other harmonized traits; Angers 2003: chlorophyll a, chlorophyll b, chlorophyll a+b, carotenoid, EWT, LMA; LOPEX93: chlorophyll a+b, carotenoid, EWT, LMA | Supervised regression: ridge / PLS / gradient-boosted trees selected by inner CV for (i) and (ii); masked multitask MSE for (iii) | Within each dataset: 75% train / 25% test, group-aware split on the Table I "Split group"; inner group k-fold CV (≤5 folds) on training rows | Test $R^2$ (and inner CV $R^2$) |
| Stage 2 – Cross-domain transfer | Raw spectra or embeddings of two datasets (56 directed pairs), or of one held-out dataset versus the remaining seven | Domain label (source vs. target) from logistic regression; reconstructions of held-out spectra | Quantify dataset discriminability (a proxy for domain divergence) and per-dataset reconstruction fidelity of the pooled encoder | 80%/20% train/test within each dataset; held-out set ≤500 random spectra; all spectra of the other datasets; encoder not retrained; no class weighting | Balanced accuracy; reconstruction MSE |
| Stage 3 – Novelty ranking | 192-D embeddings with dataset, species, and scan-count metadata | Ranked list of candidate species with confidence-adjusted novelty scores | Within-dataset Isolation Forest anomaly score, 95th-percentile species aggregation, confidence weight $1 - e^{(-n/k)}$ (Eq. 1) | Unsupervised: Isolation Forest fit within each dataset ($n \geq 10$) on all its spectra; global fallback for $n < 10$ | Adjusted novelty score (Eq. 1) |

**Table III:** Stage-wise summary of inputs, outputs, objectives, data partitions, and evaluation metrics.

## Stage 1: Unsupervised representation learning

In Stage 1, a deep autoencoder was trained on the pooled preprocessed spectra to learn a latent representation without the need for trait labels. The encoder comprised 4 fully connected layers with widths [1024, 512, 256, 128] and GELU activations, producing a 192-dimensional latent embedding; the decoder mirrored this architecture.

The autoencoder's latent dimension was chosen using a sweep over 8, 16, 32, 64, 128, and 192 **(Table IV)**. For each dimension, the model was trained on a random 90% training split, evaluated on reconstruction MSE on the remaining 10% validation split, and evaluated on downstream LMA prediction $R^2$ across three datasets (LOPEX93, CABO, NGEE Panama). Embedding $R^2$ showed an increase with the increase in latent dimension (0.640 for dim 8, 0.806 for dim 32, 0.856 for dim 192), where the major contribution was provided by the dim 32 and then onwards the returns started to diminish.

| Latent dim | Val MSE | $R^2$ (LMA) | Δ $R^2$ from dim 8 |
|---|---|---|---|
| 8 | 0.00316 | 0.640 | — |
| 16 | 0.00150 | 0.691 | +0.051 |
| 32 | 0.00153 | 0.806 | +0.166 |
| 64 | 0.00244 | 0.833 | +0.193 |
| 128 | 0.00187 | 0.848 | +0.208 |
| 192 | 0.00202 | 0.856 | +0.216 |

**Table IV:** Latent Dimension Sweep Results

The network was trained to minimize reconstruction error using the AdamW optimizer with a learning rate of 7.5e-4, batch size of 64, for up to 100 epochs with early stopping based on training loss improvement (patience = 8, minimum epochs = 35, minimum delta = 3e-4). The final autoencoder was trained on all pooled spectra; no validation or test spectra were withheld at this stage (see Data partitioning above). These embeddings were later used as features in downstream regression and visual PCA analysis.

## K nearest neighbor batch mixing entropy

Stage 1 calculates both k nearest neighbor batch mixing entropy as well as Silhouette coefficient. These

are calculated to quantify domain shift in the latent space, telling us if the encoder preserves dataset identity or learns a dataset-invariant representation. For each of the 13,549 embeddings, the 15 nearest neighbors were located in Euclidean space. The class proportions among those neighbors form a probability vector p, and Shannon entropy $H = -\Sigma\, p_i \log(p_i)$ was computed with base equal to the number of datasets (8), yielding $H \in [0, 1]$ where 0 indicates all 15 neighbors belong to the same dataset and 1 indicates uniform mixing across all eight datasets.

## Stage 1 downstream evaluation: Trait prediction under three feature views

Trait prediction was evaluated under three feature views: (i) the raw 201-band SNV-normalized spectrum, (ii) the frozen 192-dimensional embedding, and (iii) the frozen embedding passed through a supervised multitask probe. For views (i) and (ii), three candidate regressors were considered for each trait: ridge regression on standardized inputs with the regularization strength selected from a logarithmic grid ($10^{-3}$ to $10^{3}$), partial least squares regression, and histogram-based gradient-boosted trees. The candidate with the highest inner cross-validated $R^2$ on the training rows was refit on the full training split and scored once on the held-out test split; the regressor selected for each panel is indicated in the figure titles (e.g., "ridge_cv"). For LOPEX93, replicate scans were aggregated to the sample level before regression.

The multitask supervised probe is a small neural network that evaluates whether the embedding space captures trait-relevant information when jointly predicting multiple traits. The probe architecture consists of three fully connected layers: a hidden layer (128 units) with LayerNorm and GELU activation, a bottleneck layer (64 units) with GELU, and an output layer with linear activation producing predictions for all traits simultaneously. The probe is trained on the 192-dimensional frozen embeddings using AdamW optimization (learning rate 1e-3, weight decay 1e-4) with a masked mean squared error loss to handle missing trait values. Training uses early stopping (patience 5, minimum 12 epochs) on the training loss. The probe is fit within each dataset's 75% training split (the same group-aware split used for views (i) and (ii)) and evaluated on the held-out 25% test split, providing a supervised baseline that assesses the linear and nonlinear predictability of traits learned from representations.

## Stage 2: Cross-domain transfer

We evaluated the extent to which dataset (domain) identity is encoded in raw spectra and in the Stage-1 embeddings across eight leaf spectroscopy datasets spanning lab and field measurement conditions. These 8 datasets provide us with 56 dataset pairs. This is because each of the 8 datasets have 7 possible partners, which provides 28 unique pairs. Each pair is evaluated in both directions (i.e. CABO → LOPEX and LOPEX → CABO are different ) since domain discriminability is asymmetric. In the pairwise transfer task, a logistic regression classifier was trained to distinguish source from target domain spectra using either raw reflectance or 192-dimensional encoder embeddings **(Table V)**. The classifier was an L2-regularized logistic regression (C = 1, maximum 500 iterations) trained without class weighting on the SNV-normalized spectra or the embeddings. Balanced accuracy was used throughout to correct for class imbalance between source and target datasets, which varies substantially across pairs (e.g., 190 AQGRAD

spectra versus 7,821 CABO spectra). Following the domain-adaptation literature, the balanced accuracy of such a domain classifier is interpreted as a proxy for the divergence between the two domains [51]: the easier it is to tell two datasets apart in a given representation, the larger the domain shift between them in that representation, and the less domain-invariant the representation is. The results of this experiment are reported in Results, Stage 2. To examine per-dataset reconstruction fidelity and discriminability for each dataset in turn, a leave-one-dataset-out evaluation (Figure 3) was performed using the pooled Stage-1 autoencoder. We emphasize that the encoder was not retrained with the held-out dataset excluded: the held-out spectra were part of the pooled autoencoder training set, so this evaluation measures how well the single pooled encoder represents each constituent dataset rather than generalization to a dataset unseen during training. Up to 500 spectra of the held-out dataset were drawn at random and (a) passed through the encoder–decoder to compute reconstruction MSE and (b) classified against all spectra of the remaining seven datasets using the same logistic-regression protocol, with balanced accuracy reported to account for the large class-size asymmetry (Table IV). To identify the spectral regions driving dataset separability, the domain-classification logistic regression was refit on standardized raw spectra for all 56 directed pairs, and the absolute coefficient magnitudes were averaged across pairs and plotted against wavelength.

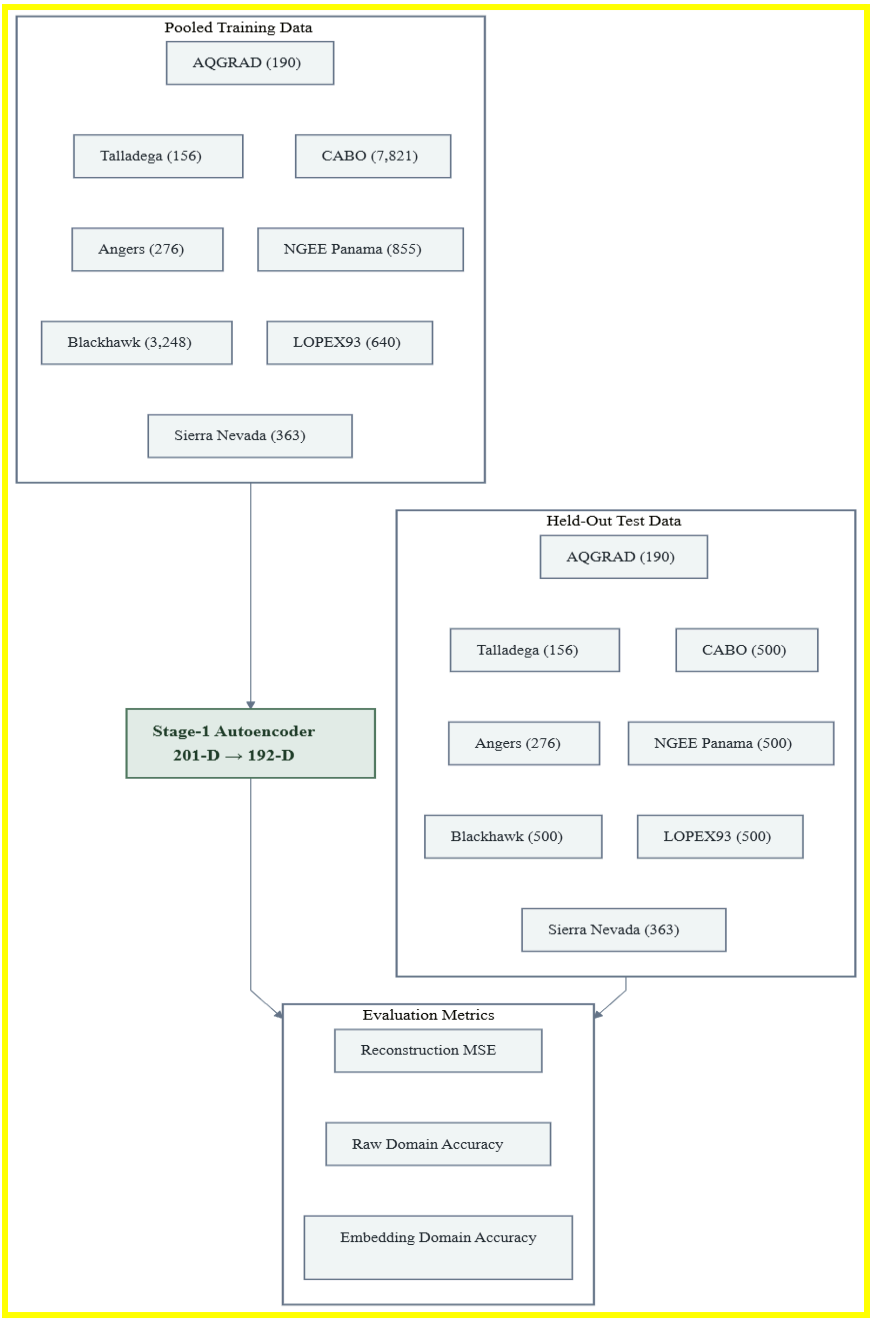



**Figure 3:** Leave-one-dataset-out experiment pipeline (the Stage-1 encoder was trained on the pooled data, which include the held-out subset; see Data partitioning)

### Stage 3: Novelty scoring

Novel candidate discovery operates on 192-dimensional spectral embeddings Z produced by the Stage-1 autoencoder encoder. Following the formulation of Isolation Forest 49, an anomaly score $s(x, \psi)$ is computed for each embedding $zi \in Z$ via the Isolation Forest ensemble, where higher scores indicate greater structural deviation from the training distribution. However, rather than applying a single global model across all concatenated datasets—which would conflate inter-dataset spectral baselines with genuine novelty—the scoring is stratified by source dataset. A global Isolation Forest ($n_{estimators}$ = 200, contamination = 0.1) is fit once on to provide fallback scores for datasets with n < 10 spectra. For each dataset $d$ with $n_d \geq 10$, a separate Isolation Forest is fit within $Zd$, and the resulting anomaly scores $a_d$ are min-max normalized to $0, 1$ to control dataset-specific spectral baselines. This within-dataset normalization prevents the combined novelty ranking from being dominated by whichever dataset occupies the densest region of the joint embedding space. At the species level, the composite novelty score for species $j$ is defined as the 95th-percentile normalized novelty $a\tilde{}j, 0.95$, aggregating across all spectra belonging to that species. A confidence weight $\omega(n_j) = 1 - e^{-n_j/k}$ down-weights species observed in fewer spectra, yielding the final adjusted score used for the ranked candidate output (1). Because each Isolation Forest is fit and normalized within its own dataset, the number of spectra contributed by each dataset (Table I) does not set the novelty scale across datasets, and the scan-count weight additionally limits the influence of species represented by very few spectra.

$$a_j^* = \tilde{a}j,\ 0.95 \cdot \omega(n_j) \qquad (1)$$

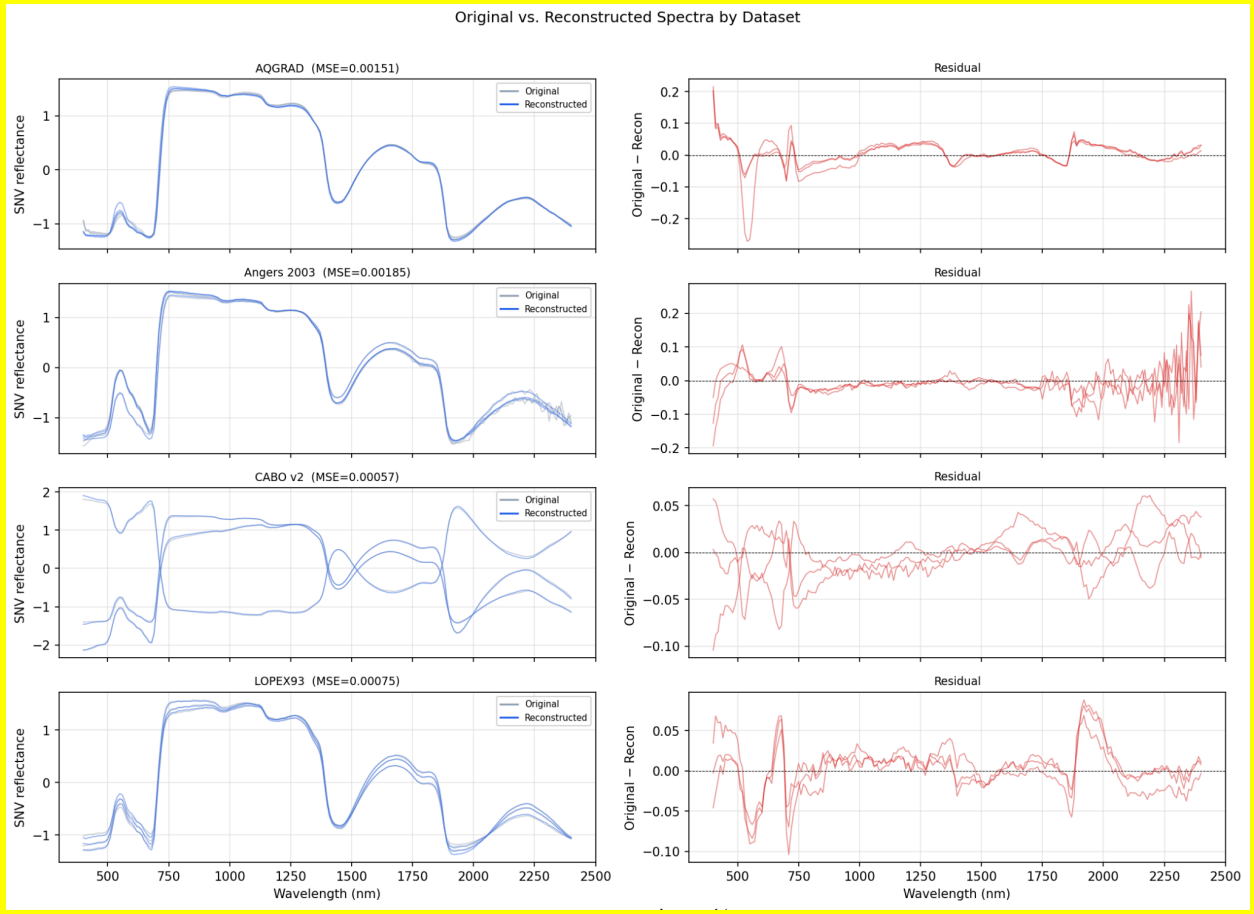
## RESULTS

In this section, the performance of the shared unsupervised autoencoder is described, including the comparison of the trait prediction results on the eight harmonized EcoSIS datasets, as well as the resulting novelty rankings adjusted for confidence. Please note that the results covered in this section may not represent the latest results of the latest pipeline. We trained a shared autoencoder on pooled leaf-level hyperspectral reflectance from eight EcoSIS datasets after harmonizing all spectra to a common 400-2400 nm grid at 10 nm spacing (201 bands). After preprocessing with Savitzky-Golay smoothing and standard normal variate normalization, the final pooled training set contained 13,549 spectra after LOPEX93 outlier filtering. The encoder compressed each spectrum into a 192-dimensional latent representation, which was then evaluated in three ways, mirroring the three stages of the Methods: first (Stage 1), by comparing raw spectra against learned embeddings for dataset-specific trait prediction; second (Stage 2), by measuring dataset discriminability and per-dataset reconstruction error; and third (Stage 3), by ranking

spectrally unusual candidate species through within-dataset normalized novelty scoring. To avoid label leakage into the representation used for transfer and discovery, the primary encoder remained unsupervised, while supervised comparisons were performed only downstream. This encoder trained on the pooled spectra including test spectra, not just their labels. This implies that all embedding-based results are transductive by nature and are not fully held out.

## Stage 1: Representation learning, trait prediction under three feature views, and reconstruction fidelity

To verify that the 192-dimensional bottleneck preserves the information content of the original spectra, we examined reconstruction fidelity per dataset. Mean squared error between original and reconstructed spectra was uniformly low across the pooled training set (AQGRAD: 0.00151; Angers 2003: 0.00185; CABO v2: 0.00057; LOPEX93: 0.00075; **(Figure 4)**). In all four datasets, reconstructed spectra closely track the originals across the visible and near-infrared regions, where the dominant pigment and structural absorption features lie. Residuals are largest at the blue end of the spectrum (400–500 nm), where leaf reflectance is lowest and sensor signal-to-noise is poorest, and in the shortwave infrared tail above 2000 nm, particularly for Angers 2003, whose instrument records noisier measurements in that region. CABO v2 attains the lowest MSE despite displaying the widest spread of input spectra, reflecting both its larger training contribution (7,821 of 13,549 pooled spectra) and its greater species and functional-type diversity, which is amplified by SNV scaling of spectrally dissimilar samples. Together, these results indicate that the encoder retains nearly all of the spectrally relevant signal under a 192-dimensional



bottleneck, and that reconstruction quality does not depend on any single dataset's measurement regime.

**Figure 4**: Original and reconstructed spectra for three random samples from each of four datasets (left), with per-wavelength residuals (original − reconstructed, right)

Trait prediction performance was driven mainly by three factors: dataset size, trait richness, and label structure. Trait-dense datasets like CABO produced the most stable cross-validation results **(Figure 5)**. Smaller field datasets such as Sierra Nevada showed more variability. Trait type was also an important factor as water-related and pigment-related traits tended to be more predictable than some structural or compositional traits. Embeddings performed either on par or better than raw spectra for several pigment-heavy tasks, particularly in Angers 2003 **(Figure 6)** and parts of LOPEX93 and CABO. Finally, candidate discovery ranking was strongly influenced by replicate count. This was addressed by applying a scan-count confidence penalty, which downweighed single-scan species and produced a more defensible top-candidate list.

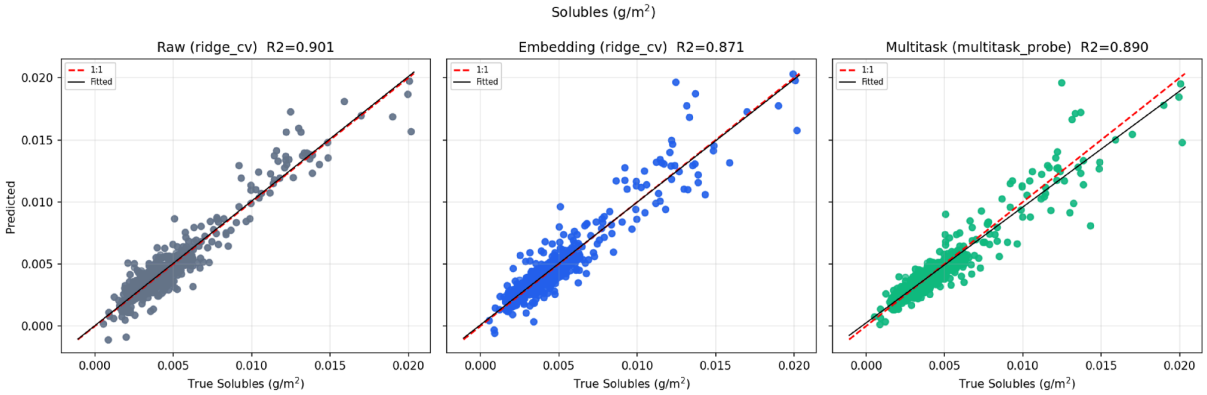



**Figure 5:** CABO v2 Solubles area trait regression.

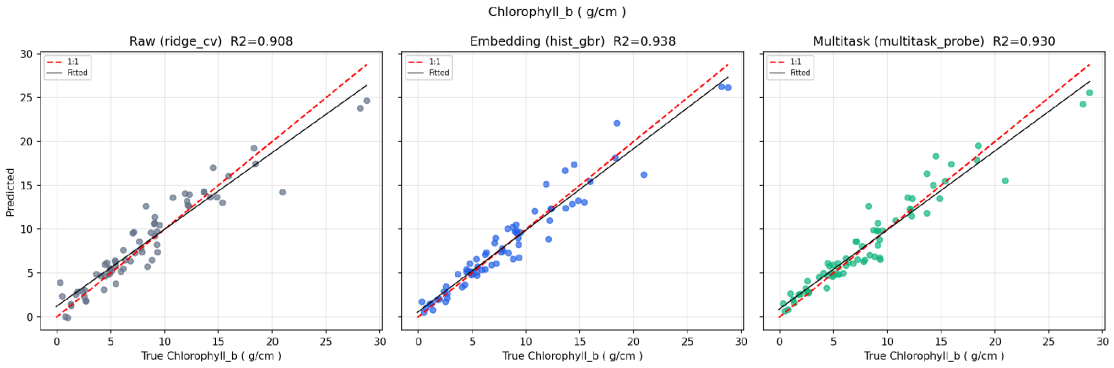


**Figure 6:** Angers 2003 Chlorophyll B trait regression.

The strongest benchmark results came from multiple datasets rather than CABO alone. In Angers 2003, substantial variance was explained for several optical and pigment traits, including Leaf mass per area (test $R^2$ = 0.894, raw), Chlorophyll a (test $R^2$ = 0.956, embeddings), Chlorophyll a+b ($R^2$ = 0.958, embeddings), Chlorophyll b (test $R^2$ = 0.940, embeddings), and Equivalent Water Thickness (test $R^2$ = 0.876, multitask). These values align with published PROSPECT and PLSR inversion benchmarks on the Angers dataset 39, 45.

In LOPEX93, Equivalent Water Thickness **(Figure 7)** remained a strong cross-species result (test $R^2$ = 0.918, raw). Raw models yielded negative or near-zero test performance on difficult pigment traits such as Chlorophyll a+b (test $R^2$ = -0.074 raw) and Carotenoid ($R^2$ = 0.074 raw). Supervised multitask learning showed a meaningful increase in predictability and showed an increase in both Chlorophyll a+b (test $R^2$ = 0.203 multitask) and Carotenoids (test $R^2$ = 0.148 multitask).

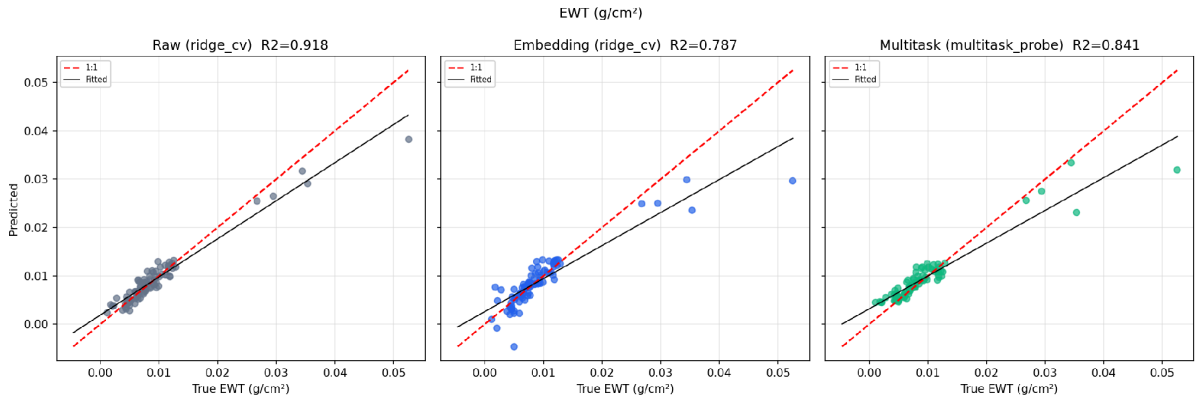



**Figure 7:** LOPEX93 Equivalent Water Thickness trait regression

In NGEE Panama, structurally meaningful traits such as leaf mass per unit area (LMA) (test $R^2$ = 0.841, raw) and specific leaf per unit area (test $R^2$ = 0.813, embeddings) confirmed that the pipeline generalized beyond temperate lab datasets into tropical leaf chemistry and economics. In Blackhawk Island, LMA was also strongly recoverable (test $R^2$ = 0.920, raw), while chlorophyll-a was predicted with moderately high accuracy (test $R^2$ = 0.872, raw).

CABO remained the broadest benchmark overall, with strong variance retrieval across many traits, including LMA (CV $R^2$ = 0.880, test $R^2$ = 0.863, raw), equivalent water thickness (EWT) (test $R^2$ = 0.897, multitask), carbon mass per unit area (test $R^2$ = 0.890, raw), and soluble mass per unit area (test $R^2$ = 0.901, raw). Despite these results, CABO no longer dominated the interpretation of the full benchmark set.

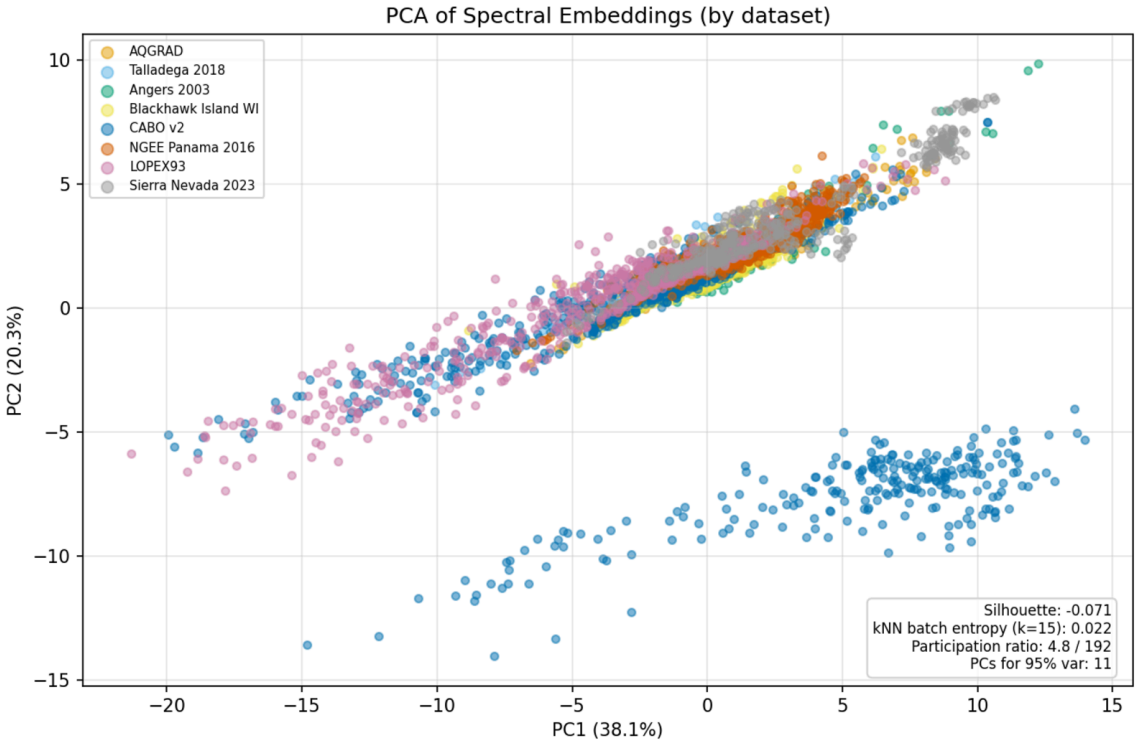



**Figure 8:** PCA projection of embeddings across 8 leaf spectral datasets.

Embedding space analysis supported a complementary conclusion. PCA projection of the embedding **(Figure 8)** revealed that spectra of various datasets existed on the same but non-mixed manifold, which shows that the model learned the biologically relevant common structure while keeping the unique effects of the particular domains of the datasets intact. To quantify this structure, the mean Silhouette coefficient with respect to dataset labels ($S = -0.108$) and the k-nearest-neighbor batch-mixing entropy ($E = 0.033$) were computed. Under the definition given in the Methods, an entropy near zero indicates that almost all of a spectrum's 15 nearest neighbors come from its own dataset, i.e., the datasets are locally unmixed in the latent space, while the near-zero (slightly negative) Silhouette indicates that the datasets do not form compact, well-separated clusters at the global scale. Together, these metrics describe a latent space in which the datasets occupy adjacent and partly overlapping regions of a shared manifold but remain locally distinguishable, which is consistent with the near-perfect dataset discriminability reported under Stage 2 below.

Although the embedding is 192-dimensional, PCA reveals that only 11 principal components are needed to explain 95% of the variance, and the participation ratio is 4.9 (out of 192), indicating that the encoder learns a representation with an effective dimensionality of approximately 5. The first two PCs account for 57.5% of variance (PC1: 37.9%, PC2: 19.6%) and exhibit a mainly one-dimensional structure. A clearly

separated cluster comprising about 1,900 points (i.e., 24% of CABO spectra) is visible along PC2, hinting at the existence of a separate CABO group – maybe due to different sampling protocols (adaxial vs. abaxial leaf surfaces) or due to separate batches of sensor calibrations in this particular dataset. This is what we expect from a partially collapsed latent space, explaining the success of the multitask supervised probe compared to the frozen embeddings.

## Stage 2: Cross-domain transfer

Across all 56 directed dataset pairs, logistic regression distinguished the two datasets with high balanced accuracy using either raw spectra (0.966–1.000) or embeddings (0.922–1.000). Discriminability was not confined to size-imbalanced pairs: pairs of datasets of comparable size, such as AQGRAD (190 spectra) versus Talladega 2018 (156 spectra) or Angers 2003 (276 spectra) versus Sierra Nevada 2023 (363 spectra), were separated with a balanced accuracy of 1.000 in both feature views and in both directions. The lowest accuracies occurred between NGEE Panama and Sierra Nevada (raw 0.966 – 0.979; embedding 0.922 – 0.939), the two field datasets that share the same instrument class and the same water-vapor masking. These results indicate that dataset identity, arising from instrument, protocol, and ecological differences, remains strongly encoded in both the raw spectra and the learned embedding; the embedding provides a shared coordinate system but is not a domain-invariant representation [51].

In the leave-one-dataset-out evaluation (Table IV), the reconstruction MSE of the pooled encoder on each held-out subset was consistently low (0.0010–0.0036 in standardized units), and the balanced accuracy of held-out versus remaining spectra ranged from 0.868 to 1.000 for raw spectra and from 0.746 to 1.000 for embeddings. Two caveats govern the interpretation of these values. First, because the encoder was trained on the pooled data that include the held-out spectra, the low reconstruction error shows that the pooled encoder represents every constituent dataset with similar fidelity; it does not show that the encoder generalizes to a dataset unseen during training. Second, the held-out and "other" sets are strongly unbalanced (e.g., 190 AQGRAD spectra versus 13,359 other spectra, Table IV), so standard accuracy would be inflated by majority-class defaults. Balanced accuracy, which was used here, corrects for this by averaging recall across both classes. This correction reveals a meaningful spread that plain accuracy concealed: the embedding's balanced accuracy on Sierra Nevada falls to 0.746, and on AQGRAD to 0.842, indicating that the encoder's representation of these small, distinctive datasets is less separable from the pooled remainder than the raw spectra suggest. The pairwise comparisons between similarly sized datasets reported above remain the more informative evidence that the datasets are separable in both feature views.

In order to identify which wavelengths account for the high dataset discriminability observed above, we investigated the importance of features for domain-classification logistic regressions **(Figure 9)**. As the coefficients of the raw spectra classifiers correspond to the 201 bands, the latter can be mapped back to the standardized spectral region. Upon averaging the absolute values of the coefficients across all 56 directed pairs, two regions became apparent: 1400–1440 nm and 1820–1930 nm. They overlap nearly perfectly with the water vapor absorption bands that are masked and linearly interpolated for the three

field datasets (Sierra Nevada 2023, Talladega 2018, NGEE Panama 2016) but retained as measured for the five laboratory datasets. In other words, the classifiers use the preprocessed signatures, not the biological differences between the datasets. A smaller elevation at 400–450 nm may correspond to the different signal-to-noise ratios at the pigment absorption edge, where the leaf reflectance is minimal. This explains the Stage 2 results above: directed pairs that share the same preprocessing step (e.g., NGEE Panama and Sierra Nevada) are the most difficult to distinguish, while any field vs laboratory pair is distinguished by the latter two regions only.

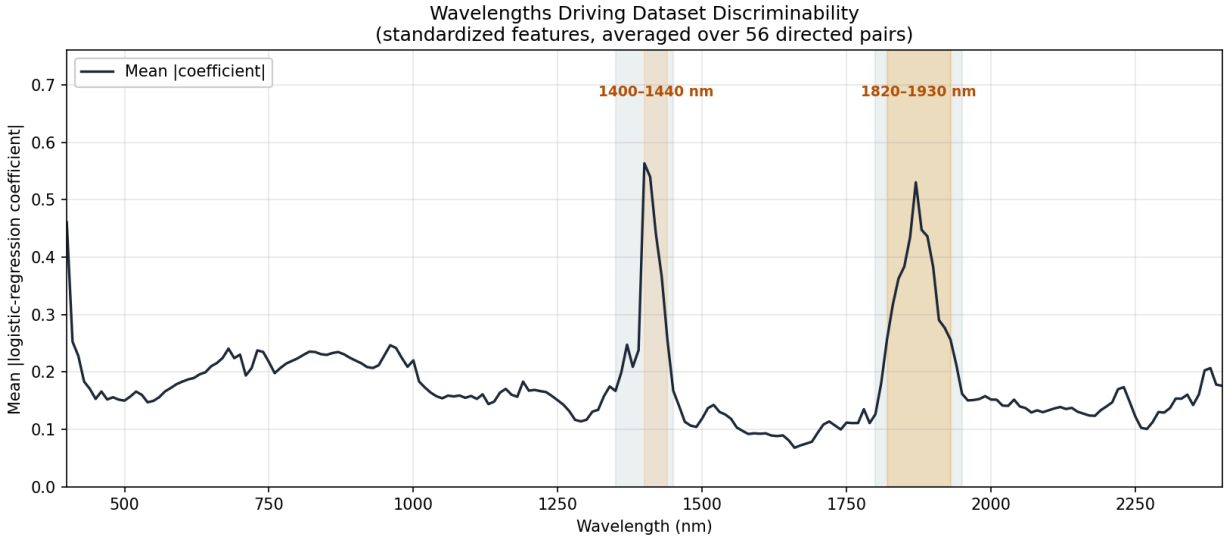



**Figure 9:** Mean absolute logistic-regression coefficient magnitudes across the 201 standardized wavelength bands, averaged over all 56 directed dataset pairs.

| Held out | n_other | n_test | Recon MSE | Raw Bal. Acc. | Emb Bal. Acc. |
|---|---|---|---|---|---|
| AQGRAD | 13,359 | 190 | 0.0026 | 0.868 | 0.842 |
| Talladega 2018 | 13,393 | 156 | 0.0019 | 0.922 | 1.000 |
| Angers 2003 | 13,273 | 276 | 0.0036 | 0.955 | 0.928 |
| Blackhawk WI | 10,301 | 500 | 0.0010 | 1.000 | 1.000 |
| CABO v2 | 5,728 | 500 | 0.0019 | 0.999 | 1.000 |
| NGEE Panama | 12,694 | 500 | 0.0014 | 1.000 | 0.973 |
| LOPEX93 | 12,909 | 500 | 0.0018 | 0.935 | 0.954 |
| Sierra Nevada 2023 | 13,186 | 363 | 0.0028 | 0.945 | 0.746 |

**Table V:** Overview of Cross-Domain – leave-one-dataset-out results: number of "other" and held-out test spectra, reconstruction MSE of the pooled encoder on the held-out subset, and domain-classification accuracy (held-out vs. other) for raw spectra and embeddings

### Stage 3: Confidence-weighted novelty ranking

Novelty scoring was computed within each dataset in stage 3. After computation, results were normalized to [0,1], aggregated to the species level using the 95th percentile scan score, and then penalized by replicate count. This adjustment substantially improved ranking stability: after the penalty was added, the top candidates were no longer dominated by single-scan species and instead shifted toward species with both high novelty and reasonable support, such as Carex richardsonii, Castanea sativa, Ceanothus parvifolius, and several high-novelty CABO taxa. Overall, the results suggest that raw spectra remain the strongest choice for direct supervised trait prediction, whereas the learned embeddings are most valuable as a shared latent space for cross-dataset comparison, pigment-sensitive tasks, and confidence-adjusted novelty discovery.

## DISCUSSION

In the current study, we demonstrate that a shared hyperspectral embedding can bring multiple leaf-level EcoSIS datasets together in one latent space; however, this representation is not always better than the original raw spectra when it comes to the trait prediction. The encoder successfully learned the structured manifold which captured the major variance in the data but maintained the domain-specific distinction between the measurement protocol and ecology (Figure 6). The latter is important since the embedding keeps enough expressiveness to be used as a feature space to compare different datasets and identify

novelties but does not compress so much as to remove the dataset identity. In other words, we show that the learned embedding is a valuable representation for the analysis; however, it should not be considered a universal replacement of the raw spectrum.

## Stage 1: Representation quality and trait prediction

One of the general conclusions from the benchmark set is that the direct regression on raw spectra remains the best baseline in cases of many structurally driven traits. This approach worked best especially in large datasets with rich trait diversity such as CABO and NGEE Panama. On the contrary, the embedding performed well in predicting pigment-related and some chemistry-related traits in Angers 2003 and LOPEX93 subsets. The above pattern indicates that the autoencoder is encoding optically significant variance, but the loss of information caused by compression cannot be ignored in the majority of quantifying applications. In other words, the resulting embedding seems to retain broader biochemical information rather than specific information of the traits. This is a good thing for transfer and discovery, but it imposes a limit on the supervised improvement of predicting traits.

Dataset size and label structure strongly influenced performance. CABO, the largest and most diverse trait-bearing dataset, gave the most stable benchmark behavior and supported strong prediction across a wide range of mass-, area-, pigment-, and mineral-related traits. NGEE Panama and Blackhawk Island similarly provided coherent leaf-level trait signals, though raw spectra usually remained better than embeddings. Angers 2003 was especially informative because pigment traits were predicted with very low residual variance, and several of those traits benefited from the embeddings. On the other hand, LOPEX93 was hard to predict even for species held out, which is not surprising considering the smaller effective number of samples and cross-species generalization difficulty. Sierra Nevada behaved as a small auxiliary stress-test dataset. It was useful for robustness checks but lacked the size to make broad claims about representation quality.

An important perspective was added via the multitask probe. As the autoencoder was still unsupervised itself, the inclusion of the multitask model checked whether supervising traits in the embedding space could reveal extra information without interfering with the shared encoder which was responsible for transfer and novelty detection. In several datasets such as CABO v2 and Angers 2003, the multitask probe helped to close the gap between frozen embeddings and raw-spectrum baselines for some traits, suggesting that the learned latent space contains biologically organized signals that can be exploited more effectively when related traits are modeled jointly. Despite these results, the probe did not overturn the broader conclusion that raw spectra remain the most reliable input for direct trait prediction and regression. The key takeaway is therefore not that supervised learning “solves” the embedding gap, but that multitask supervision can extract more value from an already biologically rich representation.

## Stage 2: What dataset discriminability does and does not show

The Stage 2 results should be interpreted as a characterization of domain shift rather than as evidence of

robustness to it. In the domain-adaptation literature, the error of a classifier trained to tell two domains apart is a standard proxy for the divergence between them [51]; near-perfect accuracy therefore means that the domains are far apart in the chosen representation, which is the opposite of domain invariance. In this light, the pairwise results show that the autoencoder, trained purely to reconstruct spectra, preserves rather than removes the dataset-specific signal introduced by instruments, sampling protocols, and ecology. This is consistent with the low batch-mixing entropy (E = 0.033) and the visibly non-mixed manifold in Figure 6, and it is the expected behavior of a reconstruction objective, which has no incentive to discard systematic between-dataset variation. What the embedding offers is a common coordinate system in which spectra from heterogeneous datasets can be compared and scored (Stage 3), not a representation in which dataset identity has been removed. Two methodological caveats further limit the Stage 2 conclusions. First, the leave-one-dataset-out evaluation used the pooled Stage-1 encoder without retraining, so the held-out spectra were seen during unsupervised training; a true test of generalization to an unseen dataset would require retraining the encoder with that dataset excluded. Second, the domain classifiers were trained without class weighting and evaluated with balanced accuracy on highly unequal class sizes, so the absolute accuracy values in the leave-one-dataset-out setting are inflated by the majority class; the pairwise comparisons between similarly sized datasets (e.g., AQGRAD versus Talladega 2018) are the more reliable indication that the separability is real. Standardized logistic-regression coefficients mapped onto the 201 wavelength bands would indicate, for example, whether the masked and interpolated water-vapor regions (1350–1450 nm and 1800–1950 nm) in the field datasets, the pigment-dominated visible region, or instrument-specific baseline differences carry most of the discriminative weight; this analysis is proposed as future work.

From the wavelength-specific coefficient plot **(Figure 9)**, it becomes clear that the major domain marker used is the water-vapor masking for the field data sets alone, and there is a minor effect due to blue region sensor difference. This is significant because the current high discriminability scores underestimate the real biological difference between the data sets, and in any future studies of the difference between the domains, consistent masking needs to be used for all the data sets or the water-vapor region should be removed before calculating domain shift.

## Stage 3: Novelty ranking as a screening tool

The discovery results strengthen the case for the embeddings as a screening tool rather than a pure predictor. The novelty-ranking pipeline was deliberately made more conservative by normalizing novelty within each dataset and then applying a confidence penalty based on scan count. That adjustment was necessary because uncorrected max-based ranking overemphasized species with only one or two spectra. After the penalty, the shortlist became more plausible and more balanced across datasets, with high-novelty species emerging from AQGRAD, LOPEX93, Sierra Nevada, CABO, and NGEE Panama rather than from a single dominant source. This is a practical advantage of the latent space: it provides a consistent way to compare unusual spectra across heterogeneous datasets, even when those datasets are not fully commensurate in their trait coverage.

## Limitations and future work

Several limitations still remain. Firstly, the pipeline relies on datasets with uneven trait availability, causing some benchmark comparisons to be far richer than others. Second, domain shift across sampling protocols is real, and is likely a cause for the residual gap between raw and embedding performance. Third, the novelty ranking is still sensitive to replicate count and should be interpreted as a candidate-generation tool rather than a ranking of validated biological importance. Fourth, the cross-domain analyses have two weaknesses that should be addressed before the latent space is used for transfer in practice: the leave-one-dataset-out evaluation did not retrain the encoder with the held-out dataset excluded, and the domain classifiers were evaluated with standard accuracy on strongly imbalanced classes rather than with balanced accuracy [52] or class-weighted training. Consequently, the present results establish that dataset identity is preserved in the embedding, but they do not establish how the encoder would behave on a dataset it has never seen. Finally, the current representation was trained unsupervised; a supervised or multitask-pretrained encoder would likely improve trait prediction, but it would also require more careful controls to avoid leaking label information into transfer and discovery analyses. For that reason, the current split between an unsupervised shared encoder and supervised downstream probes is a reasonable design.

It is important to note that the key novelty of this research is neither in the finding of the best-performing predictor nor in any other particular result; rather, it is in providing a well-rounded workflow that allows us to integrate diverse leaf spectrometry data into one framework. The findings show that raw spectra perform best in regression tasks of importance and embedding – in alignment and exploration. That division of labor is scientifically useful: it allows the pipeline to support both prediction and hypothesis generation without conflating the two. The next step is likely not simply adding more datasets but introducing a representation objective that is more trait-aware while still preserving the cross-dataset structure that makes discovery possible.

## CONCLUSION

This study demonstrates that hyperspectral leaf reflectance can be organized into a shared latent space that supports both trait prediction and candidate discovery across heterogeneous EcoSIS datasets. Raw spectra were found to be the best choice in most regression tasks, but the learned latent representation proved to be a good alternative in case of certain pigment- and chemistry-related traits and was helpful in cross-dataset analysis and novelty ranking. In addition, it was shown how the model's performance depends on dataset size, trait structure and measurement regime; larger datasets yielded the clearest trait signals, whereas smaller field datasets served as robustness stress-tests. The cross-domain experiments showed that dataset identity remains strongly encoded in both raw spectra and embeddings; the embedding therefore provides a shared coordinate system for comparing heterogeneous datasets rather than a domain-invariant representation, and its behavior on datasets unseen during encoder training remains to be established.

## DATA AND CODE AVAILABILITY STATEMENT

All spectral data used in this study are publicly available through the Ecological Spectral Information System (EcoSIS) at https://ecosis.org. The eight datasets are: AQGRAD (10.21232/xYgn5GKv), Talladega 2018 (10.21232/kRB5Gjgi), Angers 2003, Blackhawk Island WI, CABO v2 (10.21232/58kApoRy), NGEE Tropics Panama (10.15486/ngt/1478523), LOPEX93, and Sierra Nevada 2023. The complete processing pipeline, including all preprocessing, autoencoder training, downstream regression, cross-domain transfer, and novelty discovery scripts, is available at https://github.com/vamazing/Leaf2Traits. The code is licensed under the MIT license. All trained model weights, intermediate embeddings, and generated figures are included in the repository outputs directory. No proprietary or restricted data was used.

## ACKNOWLEDGEMENTS

I would like to thank Dr. Sayane Shome for her guidance and help on this project. This research would have been much harder without her endless support.

## REFERENCES

[1] Westermayr J, Marquetand P. Machine learning spectroscopy to advance computation and analysis. Chemical Science. 2025;16(46):21660-21676. Accessed June 9, 2026. https://pubs.rsc.org/sc/article/16/46/21660/911031/Machine-learning-spectroscopy-to-advance

[2] Guo K, Shen Y, Gonzalez-Montiel GA, et al. Artificial Intelligence in Spectroscopy: Advancing Chemistry from Prediction To Generation and Beyond. Ijcaiorg. 2025;12:10445-10454. doi:10.24963/ijcai.2025/1160

[3] Asner, GP; Martin, RE. Spectral and chemical analysis of tropical forests: Scaling from leaf to canopy levels. Epa.gov. 2026. Accessed July 12, 2026. https://hero.epa.gov/reference/2491935/

[4] Asner GP, Martin RE, Anderson CB, Knapp DE. Quantifying forest canopy traits: Imaging spectroscopy versus field survey. Remote Sensing of Environment. 2015;158:15-27. doi:10.1016/j.rse.2014.11.011

[5] Curran PJ. Remote sensing of foliar chemistry. Remote Sensing of Environment. 1989;30(3):271-278. doi:10.1016/0034-4257(89)90069-2

[6] Ustin SL, Gamon JA. Remote sensing of plant functional types. New Phytologist. 2010;186(4):795-816. doi:10.1111/j.1469-8137.2010.03284.x

[7] Gamon JA, Peñuelas J, Field CB. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sensing of Environment. 1992;41(1):35-44. doi:10.1016/0034-4257(92)90059-s

[8] Serbin SP, Singh A, McNeil BE, Kingdon CC, Townsend PA. Spectroscopic determination of leaf morphological and biochemical traits for northern temperate and boreal tree species. Ecological Applications. 2014;24(7):1651-1669. doi:10.1890/13-2110.1

[9] J.-B. Féret, G. le Maire, S. Jay, et al. Estimating leaf mass per area and equivalent water thickness based on leaf optical properties: Potential and limitations of physical modeling and machine learning. Remote Sensing of Environment. 2019;231:110959. doi:10.1016/j.rse.2018.11.002

[10] Kokaly RF, Asner GP, Ollinger SV, Martin ME, Wessman CA. Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. Remote Sensing of Environment. 2009;113:S78-S91. doi:10.1016/j.rse.2008.10.018

[11] Madritch M, Kingdon C, Singh A, Mock K, Lindroth R, Townsend P. Imaging spectroscopy links aspen genotype with below-ground processes at landscape scales. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2014;369:20130194. doi:10.1098/rstb.2013.0194

[12] Singh A, Serbin SP, McNeil BE, Kingdon CC, Townsend PA. Imaging spectroscopy algorithms for mapping canopy foliar chemical and morphological traits and their uncertainties. Ecological Applications. 2015;25(8):2180-2197. doi:10.1890/14-2098.1

[13] Meireles JE, Cavender‐Bares J, Townsend PA, et al. Leaf reflectance spectra capture the evolutionary history of seed plants. New Phytologist. 2020;228(2):485-493. doi:10.1111/nph.16771

[14] Schweiger AK, Cavender-Bares J, Townsend PA, et al. Plant Spectral Diversity Integrates Functional and Phylogenetic Components of Biodiversity and Predicts Ecosystem Function. Nature Ecology & Evolution. 2018;2(6):976-982. doi:10.1038/s41559-018-0551-1

[15] Guerri MF, Cosimo Distante, Spagnolo P, Fares Bougourzi, Taleb-Ahmed A. Deep Learning Techniques for Hyperspectral Image Analysis in Agriculture: A Review. Isprs Open Journal of Photogrammetry and Remote Sensing. 2024;12(7):100062. doi:10.1016/j.ophoto.2024.100062

[16] Verrelst J, Rivera-Caicedo JP, Reyes-Muñoz P, et al. Mapping Landscape Canopy Nitrogen Content from Space Using Prisma Data. Isprs Journal of Photogrammetry and Remote Sensing. 2021;178:382-395. doi:10.1016/j.isprsjprs.2021.06.017

[17] Chan JCW, Paelinckx D. Evaluation of Random Forest and Adaboost Tree-Based Ensemble Classification and Spectral Band Selection for Ecotope Mapping Using Airborne Hyperspectral Imagery. Remote Sensing of Environment. 2008;112(6):2999-3011. doi:10.1016/j.rse.2008.02.011

[18] A. Zhidebayeva, S. Syrlybekkyzy, L. Taizhanova, et al. Emerging Frontiers in Ecosystem and Biodiversity Monitoring Using Remote Sensing and Geographic Information Systems. Global Journal of Environmental Science and Management. 2025;11(4):1791-1818. doi:10.22034/gjesm.2025.04.23

[19] Seeley M, Asner GP. Imaging Spectroscopy for Conservation Applications. Remote Sensing. 2021;13(2):292. doi:10.3390/rs13020292

[20] Shepherd K D, Walsh MG. Infrared Spectroscopy—Enabling an Evidence-Based Diagnostic Surveillance Approach to Agricultural and Environmental Management in Developing Countries. Journal of Near Infrared Spectroscopy. 2007;15(1):1-19. Accessed July 26, 2026. https://opg.optica.org/jnirs/abstract.cfm?uri=jnirs-15-1-1

[21] Kacic P, Kuenzer C. Forest Biodiversity Monitoring Based on Remotely Sensed Spectral Diversity—A Review. Remote Sensing. 2022;14(21):5363. doi:10.3390/rs14215363

[22] USTIN SL, ROBERTS DA, GAMON JA, ASNER GP, GREEN RO. Using Imaging Spectroscopy to Study Ecosystem Processes and Properties. BioScience. 2004;54(6):523. doi:10.1641/0006-3568(2004)054[0523:uistse]2.0.co;2

[23] Ajaoud M, Ciccarelli C, De Mizio M, et al. Bridging Sustainability and Environmental Impact Assessment: Multi-Scale Bioindication and Remote Sensing for Pollution Monitoring in Agroecosystems. Sustainability. 2025;17(9):4115. doi:10.3390/su17094115

[24] Tucker CJ. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. Remote Sensing of Environment. 1979;8(2):127-150. doi:10.1016/0034-4257(79)90013-0

[25] Prasad Thenkabail. Hyperspectral Remote Sensing of Vegetation. ResearchGate. October 1, 2011. doi:10.13140/2.1.2985.3127

[26] Villa P, Dalla VA, Piaser E, Bolpagni R. Leaf Reflectance and Traits of Plants Sampled Along a Water Affinity Gradient (Aqgrad). Ecosis.org. September 30, 2024. Accessed July 19, 2026. https://ecosis.org/package/leaf-reflectance-and-traits-of-plants-sampled-along-a-water-affinity-gradient--aqgrad-

[27] S. Jacquemound, L. Bidel, C. Francois, G. Pavan. Angers Leaf Optical Properties Database (2003). Ecosis.org. November 11, 2016. Accessed July 19, 2026. https://ecosis.org/package/angers-leaf-optical-properties-database--2003-

[28] Chlus A, A. Townsend P. Seasonal Fresh Leaf Spectra and Traits, Blackhawk Island, WI. Ecosis.org. March 2, 2022. Accessed July 19, 2026. https://ecosis.org/package/seasonal-fresh-leaf-spectra-and-traits--blackhawk-island--wi

[29] Kothari S, Beauchamp-Rioux R, Blanchard F, et al. CABO 2018-2019 Leaf-Level Spectra V2. Ecological Spectral Information System (EcoSIS). Published online 2022. doi:10.21232/58kApoRy

[30] Serbin S, Ely K, Rogers A, et al. NGEE Tropics Leaf Spectral Reflectance Measured in Panama Collected February to April 2016. Published online 2019. doi:10.15486/ngt/1478523

[31] Hosgood B, Jacquemound S, Andreeoli G, Verdebout J, Pedrini A, Schmuck G. Leaf Optical Properties Experiment Database (Lopex93). Ecological Spectral Information System (EcoSIS). Published online 1993. Accessed May 10, 2026. http://ecosis.org

[32] Vanden Heuvel C, Queally N, Zheng T, et al. Fresh Leaf Spectra and Measured Traits from the Sierra Nevada (CA) in July 2023. Published online 2023. Accessed July 19, 2026. http://ecosis.org

[33] Kamoske AG, Dahlin KM, Serbin SP, Stark SC. 2018 Talladega National Forest: Leaf Level Reflectance Spectra and Foliar Traits. Ecological Spectral Information System (EcoSIS). Published online 2018. doi:10.21232/kRB5Gjgi

[34] Villa P, Dalla Vecchia A, Piaser E, Bolpagni R. Assessing Prospect Performance on Aquatic Plant Leaves. Remote Sensing of Environment. 2024;301:113926. doi:10.1016/j.rse.2023.113926

[35] Kothari S, Schweiger AK. Plant Spectra as Integrative Measures of Plant Phenotypes. Journal of Ecology. 2022;110(11):2536-2554. doi:10.1111/1365-2745.13972

[36] Kothari S, Beauchamp‐Rioux R, Blanchard F, et al. Predicting Leaf Traits across Functional Groups Using Reflectance Spectroscopy. New Phytologist. 2023;238(2):549-566. doi:10.1111/nph.18713

[37] Wu J, Rogers A, Albert LP, et al. Leaf Reflectance Spectroscopy Captures Variation in Carboxylation Capacity Across Species, Canopy Environment and Leaf Age in Lowland Moist Tropical Forests. New Phytologist. 2019;224(2):663-674. doi:10.1111/nph.16029

[38] Hosgood B, Jacquemoud S, Andreoli G, Verdebout J, Pedrini A, Schmuck G. The JRC Leaf Optical Properties Experiment (LOPEX'93). Jussieu.fr. 2016. Accessed July 21, 2026. http://teledetection.ipgp.jussieu.fr/opticleaf/lopex.htm

[39] Jiang J, Comar A, Burger P, Bancal P, Weiss M, Baret F. Estimation of leaf traits from reflectance measurements: comparison between methods based on vegetation indices and several versions of the PROSPECT model. Plant Methods. 2018;14:23. doi:10.1186/s13007-018-0291-x

[40] Ali AM, Darvishzadeh R, Skidmore AK, van Duren I, Heiden U, Heurich M. Estimating Leaf Functional Traits by Inversion of Prospect: Assessing Leaf Dry Matter Content and Specific Leaf Area in Mixed Mountainous Forest. International Journal of Applied Earth Observation and Geoinformation. 2016;45:66-76. doi:https://doi.org/10.1016/j.jag.2015.11.004

[41] Dechant B, Cuntz M, Vohland M, Schulz E, Doktor D. Estimation of Photosynthesis Traits from Leaf Reflectance Spectra: Correlation to Nitrogen Content as the Dominant Mechanism. Remote Sensing of Environment. 2017;196:279-292. doi:https://doi.org/10.1016/j.rse.2017.05.019

[42] Jay S, Gorretta N, Morel J, et al. Estimating Leaf Chlorophyll Content in Sugar Beet Canopies Using Millimeter- to Centimeter-Scale Reflectance Imagery. Remote Sensing of Environment. 2017;198:173-186. doi:https://doi.org/10.1016/j.rse.2017.06.008

[43] Daughtry CST, Walthall CL, Kim MS, Colstoun} EB de, McMurtrey JE. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. Remote Sensing of Environment. 2000;74(2):229-239. doi:https://doi.org/10.1016/S0034-4257(00)00113-9

[44] Gitelson A, Merzlyak MN. Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus Hippocastanum L. and Acer Platanoides L. Leaves. Spectral Features and Relation to Chlorophyll Estimation. Journal of Plant Physiology. 1994;143(3):286-292. doi:https://doi.org/10.1016/S0176-1617(11)81633-0

[45] Jacquemoud S, Baret F. Prospect: A Model of Leaf Optical Properties Spectra. Remote Sensing of Environment. 1990;34(2):75-91. doi:https://doi.org/10.1016/0034-4257(90)90100-Z

[46] Chen X, Ding M, Wang X, et al. Context Autoencoder for Self-Supervised Representation Learning. International journal of computer vision. 2023;132. doi:10.1007/s11263-023-01852-4

[47] Ericsson L, Gouk H, Loy CC, Hospedales TM. Self-Supervised Representation Learning: Introduction, Advances and Challenges. IEEE Signal Processing Magazine. 2022;39(3):42-62. doi:10.1109/MSP.2021.3134634

[48] Chen K, Liu Z, Hong L, Xu H, Li Z, Yeung DY. Mixed Autoencoder for Self-Supervised Visual Representation Learning. Thecvfcom. Published online 2023:22742-22751. Accessed July 29, 2026. https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Mixed_Autoencoder_for_Self-Supervised_Visual_Representation_Learning_CVPR_2023_paper.html

[49] Liu FT, Ting KM, Zhou ZH. Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining. 2008:413-422. doi:10.1109/ICDM.2008.17

[50] Gabdullin N. Latent Space Configuration for Improved Generalization in Supervised Autoencoder Neural Networks. arXiv.org. 2024. Accessed August 2, 2026. https://arxiv.org/abs/2402.08441

[51] Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Machine Learning. 2010;79(1-2):151-175. doi:10.1007/s10994-009-5152-4

[52] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. 2010:3121-3124. doi:10.1109/ICPR.2010.764

[53] tree-guide.com. Green Oak Leaf on White Background. Accessed October 15, 2026. https://www.tree-guide.com/images/styles/600x450-copy_/public/northern-red-oak-leaf.jpg?itok=7ntio_26