Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Frozen 2D Foundation-Model Features Organize Tree Species in LiDAR More Coherently Than Geometry or a Frozen 3D Foundation Model: A Source-Controlled Evaluation

Zhenyu Zhou

z427zhou@uwaterloo.ca

University of Waterloo

Research Article

Keywords: Tree species classification, LiDAR point clouds, Vision foundation models, Unsupervised clustering, Multi-view depth rendering

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10238136/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Frozen 2D Foundation-Model Features Organize Tree Species in LiDAR More Coherently Than Geometry or a Frozen 3D Foundation Model: A Source-Controlled Evaluation

Zhenyu Zhou[a,*]

[a]Department of Geography and Environmental Management, University of Waterloo, Waterloo, N2L 3G1, Canada

ARTICLE INFO

Keywords:
Tree species classification
LiDAR point clouds
Vision foundation models
Unsupervised clustering
Multi-view depth rendering

ABSTRACT

Individual tree species mapping from LiDAR underpins forest inventory and biodiversity monitoring, yet remains difficult because supervised classifiers need expensive, geographically biased per-tree labels, and label-free geometric descriptors capture only predefined statistics. We propose a fully unsupervised, matching-free clustering framework on frozen off-the-shelf 2D vision foundation-model features: each tree is rendered into six pose-canonical depth views, encoded by a frozen image encoder, and clustered without species labels or text prompts. Across two independent LiDAR datasets, these features organize crowns by species more coherently than a hand-crafted geometric descriptor (matching-free Normalized Mutual Information (NMI) 0.28–0.35 versus 0.10) and than a frozen 3D point-cloud foundation model given the native geometry (OpenShape/PointBERT, 0.19–0.27), despite no 3D training. Because that 3D model is pre-trained on computer-aided-design (CAD) shapes, this is an availability comparison among frozen models, not an architectural verdict. The advantage holds for every pre-trained backbone, vanishes for a random-initialized control (isolating pre-training), and is stable across four clustering algorithms, fifteen seeds, and an embedding-dimension control. We further identify acquisition-source confounding as a first-class evaluation hazard: on the multi-sensor dataset most of the global gap recovers acquisition source, leaving a small, bootstrap-significant within-source residual of +0.044 NMI (foundation over geometry) on FOR-species20K, versus +0.204 on the single-sensor BioDiv-3DTrees-ULS. Although the uncontrolled gap is permutation-significant ($p < 1e\text{-}4$), we anchor our evidence in this within-source bootstrap, treat the global gap as an upper bound, and scope our claims to feature coherence, not deployable classification.

1. Introduction

Individual tree species classification underpins forest inventory, biodiversity monitoring, and biomass estimation, yet the per-tree labels it depends on remain costly to obtain. Species classification from LiDAR point clouds is an expert task that requires botanical training, is slow at scale, and is inconsistent across annotators and field campaigns. As a result, labeled LiDAR point-cloud datasets are small, geographically clustered, and biased toward a few dominant taxa, which limits the accuracy and transferability of supervised classifiers. This labeling bottleneck motivates unsupervised tree species classification that requires no per-stem labels.

Tree species classification from LiDAR point clouds is approached in two main ways. One hand-crafts geometric descriptors (e.g., Weinmann-style summaries of local point-distribution eigenvalues and height-structure statistics (Weinmann et al., 2015)) and inputs them to a conventional classifier. The other learns representations end to end with deep networks built for LiDAR point clouds (Qi et al., 2017a,b), as in recent supervised species classifiers applied to individual-tree LiDAR point clouds (Liu et al., 2022; Allen et al., 2023). However, both approaches remain fundamentally supervised and therefore suffer from the labeling bottleneck discussed above. Hand-crafted descriptors require no labeled data, but their representational capacity is inherently constrained by predefined geometric statistics. Consequently, they overlook holistic crown morphology and silhouette characteristics that are often critical for classifying tree species. These limitations call for representations that are simultaneously label-free and more expressive than traditional hand-crafted geometry.

Frozen vision-foundation models provide such a representation. Image encoders pre-trained on web-scale data, whether by image-text contrast (Radford et al., 2021) or by self-supervision (Caron et al., 2021; He et al., 2022), yield general visual features that transfer to recognition tasks far outside their training distribution (Dosovitskiy et al., 2021). PointCLIP (Zhang et al., 2022; Zhu et al., 2023) further shows that a point cloud can be presented to such a frozen encoder by rendering it into 2D depth views. We adapt this depth-projection idea to fine-grained tree species. Each tree LiDAR point cloud is rendered into six pose-canonical depth views (a number fixed by the ablation in Section 5.4), encoded by a frozen image foundation model, and grouped by standard unsupervised clustering, with no species label and no text prompt at any stage. This setup frames the central investigation of this work. We examine whether foundation model-based features organize trees by species more faithfully than a strong hand-crafted geometric descriptor, and whether any such advantage is a general property of pre-trained visual features rather than of one particular model. To this end, we evaluate four frozen backbones through an identical pipeline: an image-text model (CLIP (Radford et al., 2021)), a self-supervised model (DINOv2 (Oquab et al., 2024)), a

*Corresponding author
Email address: z427zhou@uwaterloo.ca (Z. Zhou)
ORCID(s):

supervised-ImageNet model, and a random-initialized control. As the key additional label-free comparator, we further evaluate a frozen 3D point-cloud foundation model (OpenShape PointBERT, ShapeNet-pretrained) supplied with the native tree geometry and run through the same clustering harness. The image-pretrained features are thus tested not only against hand-crafted geometry but against an off-the-shelf frozen 3D foundation model from their own target modality. We control the comparison for three confounds that inflate such results, namely feature dimensionality, acquisition source, and the upper-bound nature of post-hoc label matching.

Extensive experiments demonstrate that the proposed framework is substantially more species-coherent than traditional hand-crafted geometry in unsupervised clustering. On two independent, heterogeneously sourced datasets spanning Terrestrial, Mobile, and UAV Laser Scanning across many acquisition campaigns, though predominantly European, foundation model-based features cluster species markedly more coherently than geometry on the matching-free Normalized Mutual Information (NMI). They also cluster species more coherently than a frozen 3D foundation model given the native geometry (OpenShape PointBERT), producing the ordering frozen 2D foundation model > frozen 3D foundation model > hand-crafted geometry on both datasets globally; the same ordering holds within source, although the frozen-3D leg of that within-source comparison rests on a single seed ($n = 1$) and is therefore indicative only, not asserted as firmly as the global leg. Because OpenShape is CAD/ShapeNet-pre-trained and thus out-of-domain for forest LiDAR, part of this gap may reflect domain gap rather than modality, so we read it as an availability comparison among off-the-shelf frozen models, not an architectural verdict. The advantage reproduces for every pre-trained backbone, while the random-initialized control remains at the level of the geometric descriptor. This isolates pre-training, rather than the rendering pipeline, as the driver, and the advantage persists, attenuated, after the dimensionality and acquisition-source controls. Crucially, on the multi-sensor dataset most of this advantage is recovery of acquisition source. The honest species-attributable effect is therefore the within-source residual, a small but bootstrap-significant +0.044 NMI on FOR-species20K, whereas a larger residual (+0.204 NMI) survives only on the single-sensor BioDiv-3DTrees-ULS. We therefore treat the global 2.7–3.5× gap as an upper bound rather than the headline.

In summary, this work makes three main contributions.

- We propose a fully unsupervised, matching-free framework that depth-renders individual-tree LiDAR and encodes it with a frozen off-the-shelf 2D vision foundation model. Its features organize trees by species more coherently than both a hand-crafted geometric descriptor and a frozen 3D point-cloud foundation model given the native geometry, yielding an ordering for tree-species structure in LiDAR (frozen 2D foundation model > frozen 3D foundation model > hand-crafted geometry) in which a model never trained on 3D data leads (interpreted with the domain-gap caveat above; Section 5.6).
- We design a source-controlled evaluation that identifies and quantifies acquisition-source confounding as a first-class hazard in matching-free foundation-model probing on LiDAR. On a multi-sensor benchmark, most of the apparent species advantage is recovery of sensor and acquisition identity, so the honest species-attributable effect is the within-source residual (small but bootstrap-significant, +0.044 NMI) rather than the global advantage, and only on single-sensor data does a larger species residual (+0.204) survive. We present this decomposition as a cautionary result for representation-probing claims in remote sensing.
- We establish these findings through a rigorous, matching-free protocol — dimension-matched, source-controlled, permutation-tested ($p <$ 1e-4), multi-algorithm, and multi-seed — with a random-initialized control that causally isolates pre-training, rather than architecture or dimensionality, as the driver, and with all label-matched scores reported explicitly as upper bounds rather than accuracies.

2. Related Work

2.1. Depth-projection and frozen visual features for point clouds

Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) learns a joint image–text embedding space from large-scale web image–caption pairs, yielding visual representations that transfer to downstream recognition tasks without task-specific fine-tuning. Because CLIP's image encoder is trained on natural photographs, applying it to LiDAR data requires bridging the modality gap between point clouds and 2D images. PointCLIP (Zhang et al., 2022) first showed that this gap can be crossed in three steps. A point cloud is projected onto multiple 2D depth maps, each view is encoded with the frozen CLIP image encoder, and the per-view embeddings are aggregated. This enables zero-shot point-cloud recognition through CLIP's pre-trained knowledge. PointCLIP V2 (Zhu et al., 2023) improved the projection and prompting pipeline to substantially narrow the gap between projected depth maps and the natural images CLIP was trained on. It produces denser, more photorealistic depth renderings and uses prompts generated by a language model.

The depth-projection idea predates CLIP. Multi-view CNNs (Su et al., 2015) first showed that rendering a 3D shape into several 2D views and pooling per-view CNN features yields a strong shape descriptor. RotationNet (Kanezaki et al., 2018) extended this by treating viewpoints as latent variables learned without viewpoint supervision. SimpleView (Goyal et al., 2021) later reported that a minimal depth-projection baseline rivals point-native networks on shape classification. Beyond PointCLIP, ULIP (Xue et al., 2023) aligns point clouds to a frozen image-text space in a backbone-agnostic

manner for 3D understanding. Our pipeline follows this depth-projection approach. Like PointCLIP, it renders each tree LiDAR point cloud into multiple 2D depth views and embeds them with a frozen image encoder. It differs from PointCLIP in two respects. First, it uses the encoder purely as a visual feature extractor and discards the text branch entirely. CLIP's language space encodes no useful prior over Latin or common tree-species names. The embeddings are therefore consumed by unsupervised clustering rather than by zero-shot text matching. Second, the target is fine-grained tree-species structure rather than coarse object categories, a regime in which the species-relevant content lies in crown architecture and branching geometry rather than in object silhouette.

The backbone shared across these methods is the Vision Transformer (ViT) (Dosovitskiy et al., 2021), which showed that large-scale pre-trained transformer features transfer to downstream recognition when image patches are treated as tokens. Beyond image–text CLIP, self-supervised vision foundation models offer a text-free alternative. DINO (Caron et al., 2021) first reported that self-supervised ViT features contain explicit semantic and object structure usable without labels. Masked autoencoders (He et al., 2022) introduced a contrasting masked-image-modeling pre-training objective. Segment Anything (Kirillov et al., 2023) demonstrated strong zero-shot transfer of a promptable frozen vision foundation model. Self-supervised distillation backbones such as DINOv2 (Oquab et al., 2024) learn transferable image features from curated images alone, with no captions and no class labels, yet match or exceed weakly-supervised features on dense recognition tasks. This paradigm has also been adopted in remote sensing. Surveys document a rapidly growing family of Earth-observation foundation models (Xiao et al., 2024), and SatMAE (Cong et al., 2022) showed that masked-autoencoder pre-training on satellite imagery transfers to downstream land-cover tasks. Because such a backbone is never exposed to text or labels of any kind, it allows us to examine whether the depth-feature species-alignment effect requires image–text supervision or is a general property of frozen, pre-trained vision encoders.

Our unsupervised stage builds on deep clustering of pre-trained features. Beyond the classic objectives of Deep Embedded Clustering (DEC) (Xie et al., 2016), SwAV (Caron et al., 2020), and self-labelling (Asano et al., 2020), TEMI (Adaloglou et al., 2023) clusters directly on frozen self-supervised features through a mutual-information objective, the closest precedent for clustering foundation model-based features. For point clouds, unsupervised point-cloud semantic segmentation has been approached by progressive superpoint growing (Zhang et al., 2023) and by distilling multi-view 2D features into the point cloud before super-voxel clustering (Chen et al., 2023), the latter closely paralleling our design.

2.2. Tree species classification

Classifying tree species from remote-sensing data is a long-standing problem. The review by Fassnacht et al. (2016) surveys the sensors (multispectral, hyperspectral, and LiDAR), feature families, and classifiers used in the field, and documents the field's heavy reliance on supervised learning with expensive, expert-collected reference labels. This annotation cost is precisely the bottleneck that motivates an unsupervised approach. Airborne Laser Scanning established individual-tree analysis by detecting and measuring single trees from canopy structure (Persson et al., 2002; Popescu and Wynne, 2004), and deep learning now also drives the upstream individual-tree segmentation step on dense LiDAR point clouds (Ma et al., 2025). On the structural side, a common LiDAR-based representation is the hand-crafted geometric (e.g., Weinmann-style (Weinmann et al., 2015)) descriptor, which summarizes local point-distribution and height-structure statistics. This work adopts a traditional (size-invariant) variant of this descriptor as its baseline to isolate the contribution of learned visual features from that of explicit geometry. The supervised alternative learns representations end to end with deep point-cloud backbones. PointNet (Qi et al., 2017a) consumes raw point sets through a permutation-invariant max-pooling operator, and PointNet++ (Qi et al., 2017b) adds hierarchical local feature aggregation. DGCNN (Wang et al., 2019) builds dynamic graphs in feature space. KPConv (Thomas et al., 2019) defines continuous kernel-point convolutions in Euclidean space. Attention-based variants such as Point Transformer (Zhao et al., 2021) and the hybrid convolution-transformer 3DCTN (Lu et al., 2022) combine local aggregation with global context. Several studies apply these backbones to individual-tree species classification. Liu et al. (2022) benchmark PointNet-family networks on ground-based LiDAR stems. 2D renderings of single-tree LiDAR point clouds have also been classified with CNNs (Seidel et al., 2021) and with deep networks over multiple rendered depth views per stem (Allen et al., 2023). All of these methods are fully supervised and depend on field-collected species labels.

Recent large and openly released benchmarks have made cross-study evaluation feasible. FOR-species20K (Puliti et al., 2025) assembles roughly 20,000 individual-tree LiDAR point clouds from many source studies into a single species-classification benchmark and reports strong supervised deep-learning baselines. We use it as one of our two evaluation datasets. Because our protocol is unsupervised rather than supervised like the benchmark's leaderboard, we do not rank our method against the published accuracies (see Section 6.3).

Beyond structure, the spectral branch of this literature is also dominated by supervised deep learning. 3D CNNs classify species from airborne hyperspectral imagery (Zhang et al., 2020), and fusion approaches combine multispectral imagery with LiDAR for individual urban-tree classification (Hartling et al., 2019; Zhang and Qiu, 2012). A recurring weakness across these supervised pipelines is limited transfer. The IDTReeS cross-site competition (Graves et al., 2023) on NEON airborne data reported that species-classification accuracy drops substantially when labeled models are applied to unseen sites, underscoring the dependence on site-specific labels that a label-free approach seeks to avoid.

The most closely related label-free studies are recent. A training-free pipeline clusters individual trees by graph community detection over hand-crafted geometric shape descriptors (Kohek et al., 2026). This corresponds to the geometry arm of our comparison, which foundation model-based features outperform in unsupervised cluster coherence. In a parallel modality, Abuhani et al. (2026) cluster frozen visual embeddings of street-level imagery for label-free genus-level urban tree diversity, whereas we resolve per-tree species from LiDAR depth views. Concurrent label-efficient work reduces species-label demand through self-supervised pretraining on forest LiDAR point clouds (Rizaldy et al., 2025) but still fine-tunes a supervised species head, unlike our fully frozen, label-free encoder.

3. Method

We cast unsupervised tree-species classification as clustering of per-tree embeddings, contrasting a frozen vision-foundation image encoder against a traditional hand-crafted geometric descriptor. The pipeline comprises four stages (Figure 1), namely (1) multi-view depth rendering of each LiDAR point cloud, (2) feature extraction with a frozen vision-foundation image encoder, (3) L2-normalized multi-view fusion, and (4) KMeans clustering with Hungarian matching used only for evaluation. Only the frozen image encoder changes between backbones, while the rest of the pipeline is identical. No component is trained or fine-tuned on the target data, which means the embeddings carry no species supervision. We detail each stage below, using CLIP ViT-B/16 as the representative encoder before turning to the additional backbone variants, the geometric baseline, and the clustering and evaluation protocol.

3.1. LiDAR point-cloud preprocessing

Every individual-tree LiDAR point cloud is first voxel-downsampled to a common resolution before rendering. We apply voxel downsampling with a target of $n_{\text{target}} = 4096$ points and a voxel size of 0.1 m, uniformly across all clouds. Applying identical downsampling to all clouds caps local point density across the heterogeneous Airborne, Terrestrial, Mobile, and UAV Laser Scanning (ALS, TLS, MLS, ULS) acquisitions, so that the encoder is less able to exploit raw local-density differences as a spurious cue. It does not equalize platform-level differences in coverage and scan completeness, which remain a route by which acquisition source stays partly recoverable (Section 6.1).

3.2. Multi-view PCA depth rendering

Each downsampled cloud is rendered into six depth views at 224×224 resolution. The in-plane principal (XY) axis is found by Principal Component Analysis (PCA) and rotated about the vertical (Z) axis through six evenly spaced angles in $[0, \pi)$. Concretely, each cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$ is rendered at the six azimuths

$$I_k = \mathcal{R}(\mathbf{P}; \theta_k), \qquad \theta_k = \tfrac{(k-1)\pi}{6}, \quad k = 1, \dots, 6, \quad (1)$$

where $\mathcal{R}$ is the PCA-aligned, orthographic, z-buffered depth-rendering operator that projects onto a 224×224 grid and keeps, at each pixel, the nearest surface along the view ray. Each rotated axis defines the horizontal projection direction, with tree height as the vertical, so the renders are pose-consistent rather than dependent on an arbitrary global coordinate system. Within each view, the horizontal (projected in-plane) and vertical (height) coordinates are each independently min–max normalized to span the full 224×224 canvas, and the retained per-pixel depth scalar is likewise min–max normalized per tree to the 8-bit $[0, 255]$ intensity range. Scaling every tree to fill the frame means the rendering itself discards absolute size, so the depth views are size-invariant in the same sense as the geometric baseline (Section 3.5). What makes the representation approximately rotation-stable is the six densely-sampled in-plane angles (Section 5.4) rather than PCA canonicalization alone, since PCA alignment is only approximate for near-symmetric crowns. The two-view special case (the major and minor PCA axes) is the $n = 2$ instance of this scheme. We fix the number of views to six by the ablation in Section 5.4. It shows that moving from two to six views raises cluster quality on the denser-scan dataset and is neutral on the other, with no gain beyond six. The 224×224 size matches the native input resolution of the CLIP image encoder, so the renders are consumed without resizing artifacts. Figure 2 illustrates this rendering on a real *Pinus sylvestris* LiDAR point cloud.

3.3. Frozen CLIP image encoder

Each of the six depth views is passed through a frozen OpenCLIP ViT-B/16 image encoder (Radford et al., 2021; Ilharco et al., 2021), producing a 512-dimensional embedding per view. Each single-channel depth view is replicated to three identical channels (no colormap is applied) and passed through the encoder's standard image preprocessing — normalization by the CLIP per-channel mean and standard deviation — before it enters the frozen tower. We denote this frozen encoder by ϕ, so that a depth view I_k maps to a per-view feature $\phi(I_k) \in \mathbb{R}^d$. The encoder weights are never updated. There is thus no training, fine-tuning, or prompt tuning on either target dataset, which keeps the downstream clustering fully unsupervised with respect to species labels.

3.4. L2-normalized multi-view fusion

The per-tree embedding $\mathbf{f} \in \mathbb{R}^d$ is the L2-normalized mean of the V per-view features,

$$\mathbf{f} = \frac{1}{V} \sum_{k=1}^{V} \frac{\phi(I_k)}{\|\phi(I_k)\|_2} \in \mathbb{R}^d, \quad (2)$$

with $V = 6$ views and feature dimension $d = 512$ for CLIP ($d = 768$ for the DINOv2 ViT-B/14, supervised-ImageNet

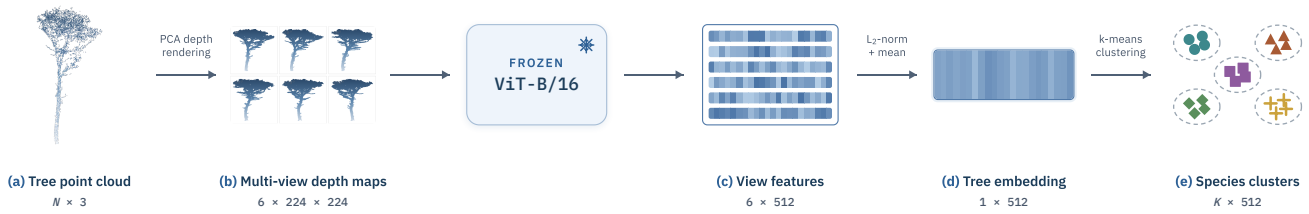

Fig. 1. Frozen vision-foundation pipeline (unsupervised). Each tree LiDAR point cloud is rendered from six viewpoints into PCA-rotated depth views, encoded by a frozen image tower (the snowflake marks frozen weights), L2-normalized and mean-fused into one per-tree embedding, then clustered by KMeans with the number of clusters set to the species count. Only the frozen encoder changes between backbones (CLIP, DINOv2, supervised-ImageNet, and a random-init control), and no stage is trained on the target data.

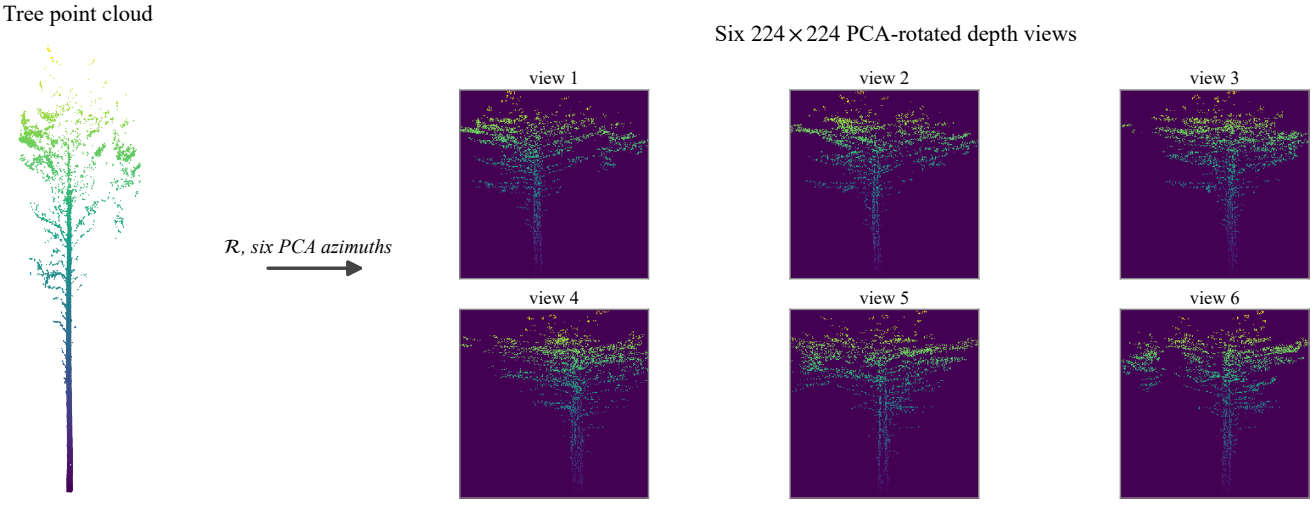

Fig. 2. Multi-view depth rendering of a single tree. A real *Pinus sylvestris* LiDAR point cloud (FOR-species20K, treeID 2573) is rendered into six PCA-rotated 224×224 depth views (Eq. 1), the in-plane principal axis rotated about the vertical through six evenly spaced angles. These pose-canonical views are the input to the frozen image encoder.

ViT-B/16, and random-initialized ViT-B/16 backbones). Normalizing each view to the unit hypersphere before averaging makes the fusion a direction-preserving mean rather than one dominated by view-specific magnitude differences. This fused vector $\mathbf{f}$ is the single per-tree descriptor used throughout, the input to clustering. For the dimension-capacity control we additionally PCA-reduce it to 11 dimensions to match the geometric descriptor.

Backbone variants. The pipeline is backbone-agnostic, with the frozen image encoder the only component that changes between backbones. Besides CLIP ViT-B/16, we evaluate a frozen DINOv2 ViT-B/14 (self-supervised), a frozen supervised-ImageNet ViT-B/16, and, as a control, a random-initialized (untrained) ViT-B/16. Each consumes the same six depth views and produces an L2-normalized mean embedding (512-d for CLIP, 768-d for the DINOv2 ViT-B/14, supervised-ImageNet ViT-B/16, and random-initialized ViT-B/16 models), reduced to geometry's 11 dimensions by PCA for the matched comparison. No backbone is fine-tuned or exposed to species labels.

3.5. Geometric baseline

The competing baseline is a traditional, size-invariant 11-dimensional geometric descriptor built from Weinmann-type covariance features computed at the whole-tree scale — a single global covariance decomposition per tree, rather than the per-point, multi-scale local-neighbourhood eigenfeatures of the original Weinmann formulation (Weinmann et al., 2015). We therefore describe it as a whole-tree single-scale covariance descriptor, and confirm against a true local multi-scale variant in Section 6.3. Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of the per-tree coordinate covariance, with $s = \lambda_1 + \lambda_2 + \lambda_3$, and let z_q be the within-tree height percentiles normalized by tree height h. From these, the descriptor stacks six dimensionless covariance ratios and five height fractions,

$$\mathbf{g} = \left[\frac{\lambda_1-\lambda_2}{\lambda_1}, \frac{\lambda_2-\lambda_3}{\lambda_1}, \frac{\lambda_3}{\lambda_1}, \frac{\lambda_1-\lambda_3}{\lambda_1}, -\sum_i \frac{\lambda_i}{s} \ln \frac{\lambda_i}{s}, \frac{\lambda_3}{s}, \frac{z_q}{h} \right] \in \mathbb{R}^{11}, \tag{3}$$

whose six ratios are the linearity, planarity, sphericity, anisotropy, eigenentropy, and change-of-curvature, with $q \in \{10, 25, 50, 75, 90\}$. Every entry is a ratio, so $\mathbf{g}$ is

invariant to isotropic scaling. It deliberately discards absolute tree size, isolating shape and structure as a fair, dimension-matched (11-d) reference point for the PCA-reduced CLIP control.

3.6. KMeans clustering and Hungarian evaluation

Given the per-tree features $\{\mathbf{x}_i\}_{i=1}^{n}$ (either the fused embedding $\mathbf{f}$ of Eq. 2 or the geometric descriptor $\mathbf{g}$ of Eq. 3), KMeans partitions them into K clusters $\{C_j\}$ by minimizing the within-cluster variance

$$\min_{\{C_j\}} \sum_{j=1}^{K} \sum_{\mathbf{x}_i \in C_j} \left\| \mathbf{x}_i - \boldsymbol{\mu}_j \right\|_2^2, \qquad \boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i, \quad (4)$$

over $n_{\text{init}} = 10$ restarts. We fix K to the number of species in each analyzable subset ($K = 15$ for FOR-species20K, $K = 6$ for BioDiv-3DTrees-ULS), the known-K (oracle-K) convention standard in the deep-clustering literature, following DEC (Xie et al., 2016) and SCAN (Van Gansbeke et al., 2020). Robustness to mis-specified K and to estimate-K alternatives is reported in Section 5.3. To evaluate cluster quality we solve an optimal cluster-to-species assignment with the Hungarian algorithm and then report NMI, macro-F1, Overall Accuracy (OA), and mean Intersection-over-Union (mIoU). NMI is matching-free and is therefore our primary metric (the dimension-capacity confound is handled separately by the control that PCA-reduces the embedding to 11 dimensions, not by NMI itself). The upper-bound nature of Hungarian matching is detailed in Section 4.3.

To assess stability, the clustering is repeated over 15 KMeans seeds and we report the mean ± standard deviation, evaluating only the analyzable subset of trees (species with at least 30 trees spanning at least 2 spatial blocks, with the over-represented *Fagus* class in BioDiv-3DTrees (Griese et al., 2025) capped to 39% of its trees).

3.7. Frozen 3D foundation-model baseline

To place the depth-projection pipeline against a competitor from its own modality, we add an off-the-shelf frozen 3D point-cloud foundation model and evaluate it through the identical clustering harness. We use OpenShape (the PointBERT variant, canonical checkpoint openshape-pointbert-shapenet, scaling factor 1), ShapeNet-pretrained and kept fully frozen with no fine-tuning and no probing head, so that it consumes the native tree geometry directly rather than rendered depth views. Each tree cloud is resampled to $N = 1024$ points and encoded from its xyz coordinates plus a constant color channel (OpenShape's no-color convention), using pure-PyTorch farthest-point sampling and ball-query grouping so that no custom CUDA operator is required. We take the primary 3D comparator to be the 256-dimensional pre-projection pooled token, because OpenShape trains the 1280-dimensional projected token to align with CLIP's image–text embedding space, so that projected token shares representational structure with the very 2D encoders it is meant to challenge, whereas the pre-projection pooled token is the cleaner, modality-native 3D readout. We report the 1280-dimensional CLIP-aligned projected token alongside it to avoid cherry-picking a single readout. The checkpoint loaded with zero missing and zero unexpected parameters, and its configuration was verified line-by-line against the official OpenShape repository, so the baseline is off-the-shelf and not inadvertently crippled. Model choice, preprocessing, and reporting rules were fixed in a pre-registration before any 3D result was seen (reports/PREREG_3d_baseline.md), including the commitment to report this head-to-head comparison in the main text regardless of its outcome. The resulting embedding is passed through exactly the same matched-dimension clustering harness as the image backbones (StandardScaler, PCA to 11 dimensions, KMeans at the true species count) and scored with the same matching-free NMI and within-source control. This backbone is thus an off-the-shelf frozen 3D competitor evaluated under an identical protocol. Its results appear in Section 5.6.

4. Datasets and Evaluation Protocol

4.1. Datasets

We evaluate on four independently collected, publicly released tree point-cloud datasets that together span three acquisition platforms (Airborne, Terrestrial and Mobile, and UAV Laser Scanning) and a range of species compositions. The two datasets used for our headline unsupervised comparison (FOR-species20K and BioDiv-3DTrees) are independent of each other in sensor and acquisition campaign, though both sample European temperate forests with overlapping dominant genera (*Fagus*, *Picea*, *Pinus*). FOR-species20K's analyzable subset pools 21 source studies that are predominantly European; its only non-European source is a single North American study (UBC, Canada; 40 trees, 0.3% of the subset), and BioDiv-3DTrees is entirely German. A consistent result across their differing sensors is therefore not attributable to a single acquisition idiosyncrasy, but neither dataset probes cross-continental generalization or a substantially different species pool, which we flag as limitations (Section 6.4). The remaining two datasets carry no headline number and play supporting roles only: Weiser/SYSSIFOSS (Weiser et al., 2022) is a single-biome anchor used in the supervised context only, and NEON-MLBS is a boundary-case probe used only for the depth-versus-RGB comparison (Section 6.3). Table 1 summarizes the analyzable subsets actually entering our experiments and marks each dataset's role.

For BioDiv-3DTrees we use the leave-region ULS subset with the dominant *Fagus* class down-sampled, yielding an analyzable set of 2,250 trees drawn from six species across three regions, as listed in Table 1.

All four datasets are publicly released. FOR-species20K (Puliti et al., 2025) is a cross-study compilation of individual-tree LiDAR point clouds, predominantly Terrestrial Laser Scanning, pooled from many acquisition campaigns. We

Table 1
Datasets (analyzable subsets) used in this study. FOR-species20K and BioDiv-3DTrees are the two independent datasets of the unsupervised headline comparison. Weiser/SYSSIFOSS is a single-biome supervised anchor only, and NEON-MLBS is a single-site depth-versus-RGB boundary probe only (Section 6.3).

Dataset	Sensor and platform	Region	Trees (n)	Species	Blocks (unit)
FOR-species20K	Mostly TLS (pooled with MLS and ULS)	Cross-study (Europe and N. America)	13,497	15	21 (source studies)
BioDiv-3DTrees (ULS)	UAV Laser Scanning	Germany	2,250	6	3 (regions)
Weiser/SYSSIFOSS	ALS (multi-platform, ALS subset)	Germany	1,093	7	6 (plots)
NEON-MLBS	ALS	Mountain Lake, Virginia, USA	543	6	45 (spatial blocks)

retain the source studies whose species satisfy the coverage criteria defined below, which leaves 21 analyzable blocks. BioDiv-3DTrees (Griese et al., 2025) is a combined Terrestrial and UAV Laser Scanning acquisition in Germany. We use only its UAV Laser Scanning per-tree clouds, so that this part of the comparison remains airborne and distinct from the predominantly terrestrial FOR-species20K pool. Weiser/SYSSIFOSS (Weiser et al., 2022) is a multi-platform laser scanning dataset from Germany, of which we use only the Airborne Laser Scanning subset; it serves as a single-biome anchor in the supervised context only and enters no headline unsupervised number. NEON-MLBS (Graves et al., 2023) is a NEON/IDTReeS airborne site at Mountain Lake, Virginia, USA; it appears only as the depth-versus-RGB boundary-case probe of Section 6.3 and likewise enters no headline result. The repository identifiers, licenses, and reference publications for these datasets are listed in the References.

Because FOR-species20K pools clouds acquired by different platforms, the unsupervised structure recovered on it may partly reflect sensor or acquisition campaign rather than species alone. This confound is controlled and quantified in Section 6.1 and bounded in Section 6.4.

4.2. Evaluation protocol

All evaluation is organized around blocks (plots for Weiser, regions for BioDiv-3DTrees, and source studies, i.e. acquisition campaigns, for FOR-species20K) to support leakage-aware analysis (Ploton et al., 2020; Roberts et al., 2017). Trees from the same block are never split in a way that lets near-duplicate or co-located trees inform both sides of a comparison. For FOR-species20K, where the block is a source study rather than a spatial plot, this control is source-level (acquisition-disjoint) rather than strictly spatial. From each dataset we retain an analyzable subset defined by two rules applied per species. A species must contribute at least 30 trees and must be present in at least 2 spatial blocks. Species failing either rule are dropped, since clusters cannot be meaningfully attributed to a class that is too rare or confined to a single block. This filtering yields the subset sizes in Table 1 (FOR n=13,497 across 21 blocks; BioDiv n=2,250 across 3 regions; Weiser n=1,093 across 6 plots).

BioDiv class-balance cap. In its raw composition BioDiv is dominated by *Fagus*. To prevent a single class from trivially driving cluster purity, *Fagus* is capped to 39% of the analyzable BioDiv subset. After this cap, *Picea* becomes the majority class, so we describe the procedure as "Fagus capped to 39%" rather than claiming that no class exceeds 40%.

Scope caveat. This unsupervised evaluation is global (KMeans is fit on the whole analyzable pool and scored against held-out labels for cluster quality) and is not a spatially held-out generalization test. Accordingly, our claim is that CLIP-derived clusters are more species-coherent than geometry-derived clusters within these pools. It does not extrapolate to cross-site or cross-sensor generalization. A closed-book spatial cross-validation (spatial-CV) deployability probe (Section 6.3) quantifies this boundary directly.

4.3. Evaluation metrics

We report four metrics. To compute label-based scores (macro-F1, OA, and mIoU) for unsupervised clusters, clusters are mapped to species by Hungarian (optimal) assignment against the test labels. This optimal cluster-to-class assignment is the standard evaluation protocol for unsupervised clustering. It was introduced by DEC (Xie et al., 2016) and is used throughout the deep-clustering literature, including IIC (Ji et al., 2019), SCAN (Van Gansbeke et al., 2020), and recent point-cloud unsupervised segmentation. In that literature, "clustering accuracy" is, by convention, the Hungarian-matched accuracy when the number of clusters equals the number of classes. We follow this established convention so that our numbers are directly comparable to that body of work. We additionally adopt the matching-free NMI as the primary metric, which requires no cluster-to-species assignment.

Our primary metric is the Normalized Mutual Information between the cluster assignment $\mathbf{c}$ and the ground-truth species $\mathbf{y}$,

$$\mathrm{NMI}(\mathbf{c},\mathbf{y}) = \frac{I(\mathbf{c};\mathbf{y})}{\sqrt{H(\mathbf{c})\,H(\mathbf{y})}}, \tag{5}$$

which requires no cluster-to-species assignment. The label-based scores instead require such an assignment. We take the optimal assignment. The Hungarian algorithm chooses, among all cluster-to-species permutations, the one that best agrees with the test labels post hoc. macro-F1 and mIoU are matched the same way. The resulting Overall Accuracy is therefore an upper bound on label-free accuracy in the

deployable one-to-one sense. It is the best score any *one-to-one* cluster-to-species assignment could achieve. A cluster-purity mapping that assigns several clusters to one species could score higher still, but is not a usable label-free classifier. This is why we lead with the matching-free NMI of Eq. 5.

- **NMI (Normalized Mutual Information).** Our primary metric. NMI is matching-free (it requires no cluster-to-label assignment). NMI does not by itself control for feature dimensionality, because a higher-capacity feature can still raise mutual information at fixed cluster count. We therefore address the capacity confound empirically, with a control that PCA-reduces the embedding to 11 dimensions rather than relying on NMI's properties. A random clustering yields NMI ≈ 0.003 on FOR (0.0034 on BioDiv).
- **macro-F1.** Unweighted mean per-class F1 after Hungarian matching, robust to class imbalance.
- **OA (Overall Accuracy).** Fraction of trees correctly labeled after Hungarian matching.
- **AMI (Adjusted Mutual Information).** A chance-corrected variant of NMI that is also matching-free, reported in the cluster-count sensitivity analysis (Section 5.3), where comparing across different cluster counts requires correcting for NMI's upward bias at large cluster counts.
- **mIoU (mean Intersection-over-Union).** Mean per-class IoU after Hungarian matching, reported as the 15-seed mean on the analyzable subset.

Hungarian matching gives an upper bound, not label-free accuracy. It uses the held-out test labels post hoc to pick the best cluster-to-species assignment, so the resulting macro-F1, OA, and mIoU measure cluster quality as an upper bound, not accuracy attainable without labels. This is exactly why we make matching-free NMI the primary metric. It is also why we frame every result as relative cluster coherence, more species-aligned than the geometric descriptor, rather than as a species classifier whose accuracy a practitioner could reproduce without labels.

The headline NMI values these metrics produce on the analyzable subsets are reported in Section 5.1.

5. Results

Implementation details. We instantiate the pipeline of Section 3 with a single fixed configuration shared by every backbone and both datasets. Each individual-tree LiDAR point cloud is voxel-downsampled with a voxel size of 0.1 m and a target of $n_{\text{target}} = 4096$ points. It is then rendered into six pose-canonical depth views at 224×224 resolution, the in-plane Principal Component Analysis axis rotated about the vertical through six evenly spaced angles in $[0, \pi)$. The representative encoder is a frozen OpenCLIP ViT-B/16 image tower producing a 512-dimensional per-view embedding, with three further frozen backbones (DINOv2 ViT-B/14 at 768 dimensions, a supervised-ImageNet ViT-B/16, and a random-initialized ViT-B/16 control) swapped in through the identical pipeline. The six per-view features are L2-normalized and mean-fused into one per-tree embedding, and for the dimension-capacity control this vector is PCA-reduced to 11 dimensions to match the hand-crafted geometric descriptor. Clustering uses KMeans, with the number of clusters set to the number of species and 10 random initializations per fit, repeated over 15 seeds. We report the mean and standard deviation over those seeds on the analyzable subset. Hungarian matching is applied only at evaluation time, so the matched scores are an upper bound and we report the matching-free Normalized Mutual Information as primary. Significance is assessed by a fixed-data, fixed-clustering label-permutation null over 10,000 permutations. Frozen-encoder inference runs on a single NVIDIA GeForce RTX 4080 GPU. We report no wall-clock runtime. Absolute compute is not part of any claim, and the result is framed as a relative cluster-coherence advantage over geometry rather than an absolute classification accuracy.

5.1. Unsupervised species alignment

Across both leakage-aware datasets, unsupervised K-means clustering of the frozen multi-view CLIP embedding (Eq. 2) is substantially more species-aligned than clustering of the whole-tree single-scale geometric descriptor (a Weinmann-type covariance summary (Weinmann et al., 2015); Eq. 3), on all four metrics (Table 2). All values are 15-seed means on the analyzable subset (species with at least 30 trees in at least 2 spatial blocks, the dominant *Fagus* class in BioDiv capped to 39%).

Frozen CLIP features lead the geometric descriptor by +0.176 NMI on FOR-species20K and +0.254 on BioDiv-3DTrees-ULS, positive on every one of the 15 seeds (Table 2). On the matching-free NMI this is a large relative gap. CLIP reaches 0.278 against geometry's 0.102 on FOR ($K = 15$, $n = 13{,}497$) and 0.355 against 0.101 on BioDiv ($K = 6$, $n = 2{,}250$), roughly 2.7× and 3.5× higher. This ratio is measured against the whole-tree single-scale descriptor. Against a stronger local multi-scale Weinmann baseline the gap narrows to +0.048 NMI on FOR-species20K and +0.158 on BioDiv-3DTrees-ULS (Section 6.3). The global gap is an upper bound on the species effect rather than our headline number. Because acquisition source is partly confounded with species (Section 6.1), the 2.7–3.5× ratio is source-inflated. The honest species-attributable estimate is instead the within-source residual, which we treat as the primary estimate throughout. That residual is a small but bootstrap-significant +0.044 NMI on the multi-sensor FOR-species20K, and a larger +0.204 on the single-sensor BioDiv-3DTrees-ULS. A random clustering scores only NMI ≈ 0.003, so geometry itself is well above chance and this is a gain over an informative baseline rather than over noise. We therefore read the result as a relative cluster-coherence advantage over geometry rather than accurate species classification, with the absolute levels assessed in Section 6.4.

Table 2
Unsupervised clustering quality (15-seed mean, with ± s.d. shown for CLIP-512, CLIP-11d, and geo-11d) on the analyzable subsets, all at matched 11-d unless noted. Frozen vision-foundation features (image-text CLIP, self-supervised DINOv2, and supervised-ImageNet) are more species-aligned than the traditional hand-crafted geometric descriptor (geo-11d) on both datasets, whereas a random-initialized ViT through the identical pipeline is not.

Dataset	Feature	NMI	macro-F1	OA	mIoU
FOR-species20K (K=15, n=13497)	CLIP-512	**0.278** ± 0.001	0.209 ± 0.002	**0.275** ± 0.003	0.126 ± 0.001
	CLIP-11d (PCA)	0.275 ± 0.001	**0.211** ± 0.002	0.273 ± 0.004	**0.127** ± 0.001
	DINOv2-11d	0.273	0.196	0.241	0.116
	ImageNet-11d	0.207	0.167	0.215	0.096
	random-11d	0.108	0.123	0.175	0.070
	geo-11d	0.102 ± 0.002	0.132 ± 0.003	0.198 ± 0.005	0.074 ± 0.002
BioDiv-3DTrees-ULS (K=6, n=2250)	CLIP-512	**0.355** ± 0.001	**0.287** ± 0.001	**0.420** ± 0.002	**0.191** ± 0.001
	CLIP-11d (PCA)	0.350 ± 0.001	**0.287** ± 0.001	0.419 ± 0.004	**0.191** ± 0.001
	DINOv2-11d	0.298	0.265	0.414	0.178
	ImageNet-11d	0.209	0.263	0.360	0.169
	random-11d	0.059	0.194	0.276	0.115
	geo-11d	0.101 ± 0.0003	0.204 ± 0.0001	0.326 ± 0.0001	0.124 ± 0.000

Figure 3 visualizes these NMI values alongside the source-controlled (within-source) results discussed in Section 6.1, making the global advantage and its source attenuation directly comparable.

Backbone generality and the pre-training control. The advantage is a property of pre-trained features in general, not of CLIP specifically. Repeating the comparison with three further frozen backbones under the identical pipeline (a self-supervised DINOv2, a supervised-ImageNet ViT, and a random-initialized ViT), every pre-trained backbone clusters species more coherently than geometry on both datasets, in the order CLIP $\gtrsim$ DINOv2 > ImageNet > geometry (Table 2). Their matched-dimension NMI gains over geometry span +0.10 to +0.25, all positive across the 15 seeds and significant against a label-permutation null ($p < 1e$-4, permutation-limited, since each observed backbone gap exceeds all 10,000 permuted nulls). Holm–Bonferroni correction across the six pre-trained-backbone tests (three backbones × two datasets) leaves every test significant at the 0.05 level (all adjusted p = 6e-4). This family excludes the random-initialized control, which is a negative control rather than a confirmatory hypothesis. Chance-corrected AMI corroborates the NMI ordering rather than reflecting cluster-count inflation, with CLIP exceeding geometry by +0.174 AMI on FOR-species20K and +0.250 on BioDiv-3DTrees-ULS. The random-initialized control, by contrast, stays at geometry's level, marginally above it on FOR-species20K and below it on BioDiv. The gain therefore comes from what the encoder learned in pre-training, not from the pipeline that renders, encodes, and clusters, which the untrained network also shares. The three pre-trained backbones also survive the within-source control of Section 6.1, whereas the random control does not.

5.2. Capacity control

A natural objection is that CLIP's 512 dimensions give K-means more degrees of freedom to separate classes than geometry's 11. To control for this, we PCA-reduce the CLIP embedding to the descriptor's dimensionality (CLIP-11d) and re-cluster. The advantage is preserved almost in full.

Compressing CLIP to geometry's 11 dimensions incurs almost no loss (CLIP-11d stays within 0.005 NMI of the full 512-d feature — 0.003 on FOR-species20K and 0.005 on BioDiv-3DTrees), and the matched-dimension advantage over geometry holds at +0.173 NMI on FOR-species20K and +0.249 on BioDiv-3DTrees-ULS, positive on all 15 seeds (Table 2). That PCA compression to 11 dimensions does not degrade performance indicates K-means does not exploit the extra degrees of freedom, since the advantage is a property of feature content at low dimension, not of capacity. (The separate acquisition-source confound is addressed in Section 6.1.)

A full PCA dimension sweep confirms this is not specific to the 11-dimensional operating point. Across reductions from 4 to 512 dimensions, CLIP's NMI saturates by about 11 to 16 dimensions and is essentially flat thereafter, exceeding the geometric descriptor at every dimension tested. Even a 4-dimensional CLIP embedding outperforms the full 11-dimensional descriptor (Figure 4). The same flatness holds for DINOv2 between its native 768 dimensions and the matched 11, so the dimension-insensitivity is a property of frozen pre-trained features generally, not of CLIP alone.

This contrasts with the supervised case, which we report with explicit caveats because it remains underpowered (Section 6.3). Against the matched 11-d size-invariant geometric baseline the supervised CLIP macro-F1 advantage is small as a descriptive point estimate, only +0.025 on FOR-species20K and +0.035 on BioDiv; these are size-invariant descriptive numbers, not significance-tested effects. The significance test we can run is on a different, dimension- and size-matched pair

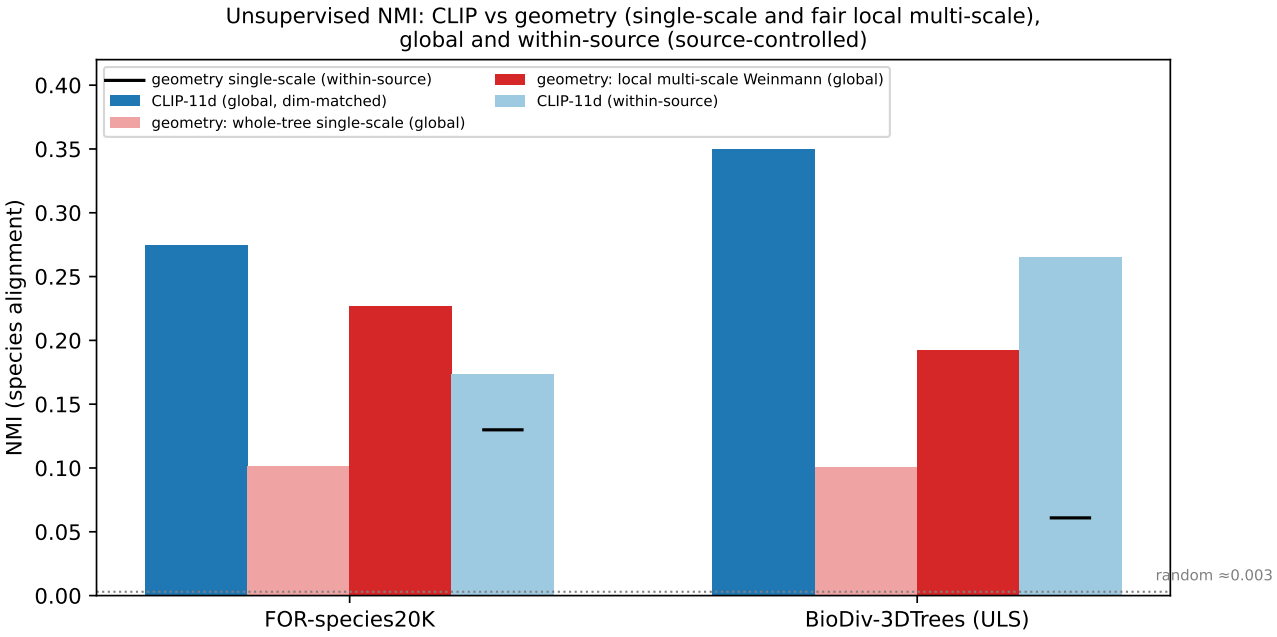

Fig. 3. Unsupervised matching-free NMI, CLIP versus geometry, global and within-source. Global bars compare the dimension-matched CLIP-11d embedding against two geometric baselines: the whole-tree single-scale covariance descriptor (light red) and a fair local multi-scale Weinmann variant (dark red; per-point eigenfeatures at k=10/20/40, per-tree mean; Section 6.3). CLIP leads the single-scale descriptor by a large margin ($\approx$ 2.7–3.5×) but leads the stronger local multi-scale baseline by only +0.048 (FOR-species20K) and +0.158 (BioDiv-3DTrees) NMI. Hatched bars (CLIP) and black markers (single-scale geometry) show the within-source (source-controlled) NMI, where the apparent advantage attenuates further (Section 6.1). The dotted line marks the random-clustering baseline (NMI $\approx$ 0.003).

(17-d CLIP against the size-aware best-geometry variant), and there the FOR-species20K point estimate is in fact slightly negative, about −0.007 (CLIP marginally behind), and is not statistically significant (Wilcoxon $p = 0.35$). A significant difference appears only in the unmatched full-dimensional comparison ($p = 0.0036$), which is itself flagged "INCONCLUSIVE (underpowered)" by its own learning-curve saturation gate. On BioDiv, significance is unreachable over only three spatial folds. The supervised arm also conflates classifier with descriptor, since CLIP uses a linear logistic-regression head, whereas geometry uses a Random Forest or a Support Vector Machine (SVM). We therefore draw no conclusion from it. The robust statement is the unsupervised one. At matched dimension CLIP depth features carry more species-coherent structure than hand-crafted geometry, surviving both the dimension-matching control and, attenuated, the within-source control.

5.3. Robustness of the clustering stage

Clustering algorithm. The advantage is not specific to K-means. We re-cluster the matched 11-dimensional features with four standard fixed-K algorithms: K-means, a Gaussian mixture (GMM), spectral clustering, and Ward agglomerative clustering, all with the number of clusters set to the number of species. Both CLIP and DINOv2 are more species-aligned than geometry on every algorithm and both datasets. This is an 8-of-8 NMI sweep for each backbone, as Table 3 reports. Spectral and agglomerative clustering are $O(n^2)$ and are run on a fixed 4,000-tree subsample for FOR-species20K. Agglomerative clustering is deterministic and so carries no seed variance (Figure 5).

We also tested two algorithms that estimate the cluster count rather than fixing it, and neither recovers the species number. A Dirichlet-process GMM infers $\hat{K} = 38$ on FOR-species20K (true $K = 15$) and $\hat{K} = 18$ on BioDiv-3DTrees (true $K = 6$). HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) collapses to $\hat{K} = 4$ on FOR and $\hat{K} = 2$ on BioDiv, leaving most points unclustered as noise. On the geometric descriptor it degenerates entirely ($\hat{K} = 0$, every point noise). Because estimate-K methods substantially misestimate the true species count in both directions, we adopt the known-K convention standard in deep clustering for the headline comparison and report robustness to K below.

Cluster count. To confirm the advantage does not hinge on knowing K exactly, we sweep K from half to five times the true species count (CLIP-11d, K-means). NMI is robust to moderate mis-specification. On FOR-species20K it is nearly constant from $K = 12$ to 18 and only drifts upward under heavy over-clustering, and on BioDiv-3DTrees it is similarly flat around the true $K = 6$. We do not claim the true species count maximises NMI. On BioDiv it is in fact highest at $K = 3$ (NMI 0.387), so we read this only as evidence that the comparison of CLIP against geometry is stable to reasonable

Table 3
Clustering-algorithm robustness. Unsupervised NMI (matched 11-d, fixed-K) for four standard algorithms. Both CLIP and DINOv2 exceed the traditional geometric descriptor on every algorithm and both datasets.

	FOR-species20K			BioDiv-3DTrees		
Algorithm	CLIP	DINOv2	geo	CLIP	DINOv2	geo
K-means	0.275	0.272	0.103	0.350	0.301	**0.101**
GMM	0.289	0.281	0.092	0.351	0.323	0.069
Spectral	**0.312**	**0.292**	**0.136**	**0.389**	**0.344**	0.093
Agglomerative	0.304	0.282	0.122	0.333	0.340	0.074

errors in K. Chance-corrected AMI penalises the upward NMI drift under over-clustering, as expected, and likewise does not single out the true count.

5.4. View-count ablation

The pipeline renders six depth views per tree (Section 3.2). We justify this choice rather than the two orthogonal views used in early prototypes by a view-count sweep. We vary the number of PCA-rotated views $n \in \{2, 4, 6, 8, 12\}$ and re-evaluate the matched 11-d CLIP NMI under the identical protocol (15 KMeans seeds, analyzable subset, BioDiv Fagus capped to 39%), on the full FOR-species20K and BioDiv-3DTrees sets. The $n = 2$ case exactly reproduces the rendering from two orthogonal views (it is the major and minor PCA pair), so the sweep measures the effect of adding views on top of that baseline.

On FOR-species20K, moving from two to six views raises NMI by +0.014 (Table 4), roughly 14 times the seed standard deviation, with the gain essentially complete by six. On BioDiv-3DTrees the view count has a negligible effect, staying within seed noise across the sweep, consistent with its sparser ULS clouds carrying less view-dependent silhouette detail than the predominantly terrestrial FOR scans. Six views is therefore the saturation point. It captures the full FOR gain, sits at or near the best on BioDiv, and requires half the computational cost of twelve views. We adopt six views throughout, as used for every headline result in Section 5.1.

5.5. Qualitative cluster structure

Figure 6 visualizes the cluster structure underlying these scores. A t-SNE projection of the CLIP-11d features colored by species shows partially separated, species-coherent regions on both datasets. The same projection colored by acquisition source reveals that part of this structure aligns with source, consistent with the source-confound analysis (Section 6.1). The row-normalized Hungarian-matched confusion matrices (Figure 7) show moderate per-species recovery, in line with the OA values in Section 5.1. The off-diagonal mass concentrates among a few structurally similar species rather than scattering uniformly, consistent with crown architecture driving the hard confusions (we do not, however, have saliency evidence to separate structural similarity from taxonomic relatedness). (The FOR-species20K t-SNE uses a 3,000-tree subsample for legibility.)

Table 4
Unsupervised CLIP NMI (matched 11-d, 15-seed mean) versus the number of PCA-rotated depth views, on the full analyzable sets. Bold marks the per-dataset maximum, though the spread beyond six views stays within seed noise (Section 5.4) and six views is the adopted operating point.

Views (n)	2	4	6	8	12
FOR-species20K	0.261	0.271	0.275	0.275	**0.276**
BioDiv-ULS	0.350	0.345	0.350	**0.357**	0.343

(The corresponding DINOv2 projections show the same partially-separated, source-aligned structure. CLIP applied to RGB renders, rather than depth, is examined as a boundary case in Section 6.3.)

Figure 8 shows representative real LiDAR point clouds for four species spanning the dataset's conifer and broadleaf crown forms.

5.6. Comparison with a frozen 3D foundation model

The comparisons so far compare image-pretrained features with hand-crafted geometry. A sharper test is whether a foundation model built for the native modality — a frozen 3D point-cloud encoder given the geometry directly — organizes species at least as well. We therefore run the frozen OpenShape PointBERT of Section 3.7 through the identical matched-dimension clustering harness. Table 5 reports its global NMI, chance-corrected AMI, and within-source NMI beside the frozen 2D CLIP encoder and the geometric descriptor.

On both datasets, and on both the global and the within-source estimates, the ordering is frozen 2D foundation model (CLIP) > frozen 3D foundation model (OpenShape) > hand-crafted geometry. Leading with OpenShape's stronger pooled-token readout, on FOR-species20K the 3D model reaches 0.204 (pooled) and 0.194 (projected) global NMI, against CLIP's 0.275 and geometry's 0.102. On BioDiv-3DTrees-ULS it reaches 0.272 (pooled) and 0.266 (projected), against 0.350 and 0.101. The chance-corrected AMI and the within-source NMI preserve this ranking. The frozen 2D encoder therefore organizes tree crowns by species more coherently than the 3D encoder, even though it never saw a 3D point cloud during pre-training and receives each tree only as rendered depth views. The 3D model, by contrast, is given the native geometry directly.

We read this as an availability comparison among off-the-shelf frozen models rather than an architectural verdict: OpenShape is CAD/ShapeNet-pre-trained and hence out-of-domain for forest LiDAR, so part of the gap may reflect this domain gap rather than the 2D-versus-3D modality. A domain-matched 3D foundation model pre-trained on real outdoor or vegetation scans, once available, could narrow or reverse the ordering, which we leave to future work (Section 6.4). The ordering is also not an artifact of the pre-registered 1024-point input to the 3D model. Re-encoding

each tree at OpenShape's native 10,000-point budget raises the 3D-FM NMI (to 0.215/0.209 on FOR-species20K and 0.292/0.290 on BioDiv-3DTrees-ULS, pooled/projected) yet leaves it below CLIP on both datasets, so the ordering survives at the 3D model's own input scale.

6. Discussion

6.1. Mechanism and source confound

We offer the following interpretation as an untested hypothesis. This study measures cluster quality against a baseline and provides no saliency, feature-attribution, or morphology-ablation evidence for why the image-pretrained features cluster more coherently. By construction the geometric descriptor sees only second-order point covariance and the vertical distribution of returns, so it encodes nothing beyond local covariance ratios and height fractions. The frozen CLIP encoder, trained on web-scale natural images, instead carries broad visual priors over shape, silhouette, and texture. Applied to multi-view depth renders, those priors may capture crown- and canopy-level cues that a small set of geometric ratios cannot summarize. The magnitudes that motivate this reading (the matched-dimension advantage and its capacity control) are reported in Section 5.1 and Section 5.2 and are not re-quoted here.

That the advantage exceeds chance label–cluster association is established by a label-permutation null. Holding the clustering fixed and permuting the species labels over 10,000 permutations, the observed matched 11-d NMI gap (CLIP minus geometry) is 0.172 on FOR and 0.249 on BioDiv, whereas the 95th-percentile null gap is only 0.0006 and 0.0021 respectively. No permutation reaches the observed value, giving $p < 1e\text{-}4$ (permutation-limited) on both datasets — the observed gap exceeds all 10,000 permuted nulls. This is a fixed-data, fixed-clustering null rather than a data-resampling or cross-site test, and it complements the 15-seed analysis, which separately establishes robustness to clustering initialization.

We foreground this acquisition-source confound as a primary contribution, in the form of a reusable, quantitative source-decomposition protocol rather than a single headline number. The protocol has three components: a within-source residual that reclusters separately inside each source block, an explicit removal of the between-source linear subspace, and a within-species source-recovery probe that measures how strongly each representation encodes acquisition identity. Sensor and domain confounding in remote-sensing model evaluation is not itself new — spatially and structurally blocked validation is precisely the concern raised by Ploton et al. (2020) and Roberts et al. (2017), on which our block-structured protocol builds — so we claim not the discovery of the hazard but a concrete decomposition that separates a source-inflated upper bound from a source-controlled residual for matching-free foundation-model probing on LiDAR. The global advantage of Section 5.1 is an upper bound on the species effect, while the source-controlled within-source residual is the honest species-attributable estimate. This reading must be weighed against a substantive competing interpretation. Because the pooled clouds span many acquisition platforms and source studies, species and acquisition source are partly confounded, and the pre-trained features predict source more strongly than geometry does. A within-source control clusters separately inside each source block, so between-source structure cannot contribute. Crucially, each block is reclustered with K set to that block's own species count, so the within-source figures sit at a different (smaller) K scale than the global K=15/6 clustering; the operative quantity is therefore not a decay of CLIP's global NMI (0.275) to its within-source value (0.174), which mixes two clustering scales, but the within-block CLIP-minus-geometry delta with both representations clustered at the same per-block K. That delta is positive on every seed: a small +0.044 NMI on the multi-sensor FOR-species20K and a larger +0.204 on the single-sensor BioDiv-3DTrees-ULS. (A second pre-specified control, removal of the between-source subspace below, is uninformative on BioDiv-3DTrees-ULS, where only three regions make its rank-2 removal over-remove species-correlated variance, so the within-source residual is the operative source control on that dataset.) The label-permutation null of Section 5.1 only rejects chance label–cluster association and is weak, since even the random-initialized control clears it. We therefore treat the data-level bootstrap below, rather than permutation significance, as the operative evidence for these residuals. The within-source NMI values, the per-backbone residual gaps, and the source-recovery probe accuracies are reported in Table 6. A data-level bootstrap that resamples trees within each source block (fixed clustering, 1,000 replicates) gives 95% confidence intervals of $[0.034, 0.058]$ on FOR-species20K and $[0.162, 0.230]$ on BioDiv-3DTrees-ULS for this residual. Both intervals exclude zero and are positive on every replicate, so the residual is significant at the level of the data rather than only across clustering initializations. This within-source residual is a size-weighted mean of those per-block NMIs, each block clustered at its own species count. On BioDiv-3DTrees-ULS the bootstrap resamples trees within only three fixed regions, so its interval reflects within-region rather than between-region variance and should be read together with the limited-external-validity caveat of Section 6.4.

As a stronger, pre-specified control we also remove the between-source linear subspace (a single variant, rank $= n_{\text{source}} - 1$) from the embeddings before re-clustering. On FOR-species20K, whose 21 sources give the removal ample rank, CLIP species-NMI falls from 0.275 to 0.130 yet stays above geometry's raw 0.102, while the all-tree source-recovery balanced accuracy — predicting acquisition source across all trees, at chance $1/n_{\text{source}} = 0.048$ for 21 sources — collapses from 0.395 to 0.026. This all-tree probe is distinct from the within-species source recovery of Table 6 (0.73 for CLIP), which predicts source among trees of a single species and whose chance is correspondingly the higher 0.26. A species residual therefore survives explicit removal of the source directions. On BioDiv-3DTrees-ULS

Table 5
Comparison with a frozen 3D foundation model (matched-dimension clustering harness, 15-seed means). OpenShape PointBERT (ShapeNet-pretrained, frozen) is given the native geometry directly through the same StandardScaler / PCA-11-d / KMeans pipeline as the image backbones; its primary 1280-d projected token (proj) and 256-d pooled token (pooled) are both reported. Global NMI is the source-inflated upper bound and the within-source NMI is the source-controlled estimate. The ordering frozen 2D FM > frozen 3D FM > geometry holds on both datasets and on both estimates, despite the 2D encoder never seeing 3D data. OpenShape's within-source NMI was recorded from a single seed ($n = 1$) on FOR-species20K (0.154); its BioDiv proj/pooled within-source NMI values are 0.167/0.189. These frozen-3D-FM within-source figures are therefore read as indicative rather than seed-averaged, and we do not lean on the exact 3D within-source ordering. CLIP and geometry within-source NMI are the dedicated source-control-harness values of Table 6 (source_control_multi_*.json); the frozen-3D-FM harness independently reproduces the CLIP within-source figures to within 0.002.

Dataset	Feature	Global NMI	AMI	Within-source NMI
FOR-species20K	CLIP (2D)	**0.275**	**0.273**	**0.174**
(K=15, n=13,497)	OpenShape (3D, proj)	0.194	0.191	0.154
	OpenShape (3D, pooled)	0.204	0.202	0.154
	geometry	0.102	0.099	0.130
BioDiv-3DTrees-ULS	CLIP (2D)	**0.350**	**0.348**	**0.265**
(K=6, n=2,250)	OpenShape (3D, proj)	0.266	0.263	0.167
	OpenShape (3D, pooled)	0.272	0.269	0.189
	geometry	0.101	0.097	0.061

the same procedure drives CLIP species-NMI to 0.055 and source recovery to 0.333 (= chance for three regions). With only three regions the rank-2 removal over-removes species-correlated variance, because species and region are strongly correlated there. This linear control is thus uninformative on BioDiv, and the appropriate source control is instead the within-source analysis (+0.204). We therefore report the subspace-removal figure with this caveat rather than reading it either way. The residual-source signal is symmetric across pre-trained backbones rather than a CLIP idiosyncrasy, since DINOv2 recovers within-species source above chance just as CLIP does. The random-initialized control, meanwhile, sits below geometry. The residual-source caveat therefore applies to every pre-trained backbone, not to CLIP specifically. We thus claim a source-attenuated residual advantage rather than a species-morphology effect. Fully disentangling morphology from acquisition is left to future work.

6.2. Supervision-free comparison

A potential objection is that CLIP is not strictly unsupervised, since its pre-training uses image-text pairs as a supervisory signal, which is why some clustering work excludes it. We address this concern empirically. A purely self-supervised DINOv2 (no text, no labels of any kind) reproduces the species-alignment advantage under the identical pipeline (Section 5.1, Table 6), significant against the label-permutation null at $p < 1e\text{-}4$ (permutation-limited) on both datasets. The effect is therefore not an artifact of image-text supervision. The random-initialized control isolates pre-training rather than the pipeline that renders, encodes, and clusters (Section 5.1), staying at geometry's level, marginally above on FOR-species20K and below geometry on BioDiv. With DINOv2 the comparison is thus supervision-free with respect to the species task. The remaining open question is narrower, namely which pre-training objective transfers best. Among our paradigms the ranking is image-text $\gtrsim$ self-supervised > supervised-classification.

6.3. Scope and boundary cases

Supervised contrast. Placed beside the supervised, leakage-free spatial-CV experiments, the unsupervised result is the robust one. Against the same matched 11-d geometric baseline the unsupervised CLIP margin is large and seed-robust while the supervised margin against that identical baseline is small and underpowered (Section 5.2), so we draw no conclusion from the supervised arm.

Deployability probe. The coherence advantage reported here is a property of the pooled feature geometry and does not, on its own, yield a deployable classifier. We test this under a closed-book spatial cross-validation protocol that fits K-means on the training blocks, labels each cluster by its training-set majority species, and applies that frozen map to spatially held-out blocks. The unsupervised macro-F1 never rises to a usable level relative to a per-block majority baseline. On FOR-species20K it reaches only 0.062 against a 0.026 majority, barely above a near-zero floor, and on BioDiv-3DTrees-ULS 0.080, below the 0.103 majority, with the geometric descriptor behaving the same way (0.039 and 0.065). This is expected and consistent with the scope we claim throughout, namely a matching-free representation-coherence diagnostic rather than a turnkey species classifier. It also reinforces the source-confound reading of Section 6.1. Spatial hold-out removes exactly the between-block acquisition structure that inflates the global comparison, so no deployable species signal survives for either representation.

Size-aware geometry. The geometric baseline is deliberately size-invariant (Section 3.5), raising the question of whether the advantage would shrink against a descriptor that retains absolute size. Re-running the identical harness on a

Table 6
Source control (15-seed means, matched 11-d), over the 21 acquisition blocks of FOR-species20K and the 3 regions of BioDiv-3DTrees-ULS. Within-source NMI clusters separately inside each block so that between-source structure cannot contribute, and the CLIP−geometry delta, positive on all 15 seeds, is the source-controlled advantage. Within-species source recovery is the balanced accuracy of predicting acquisition source from an embedding among trees of a single species, averaged over species; its chance level (0.26 on FOR-species20K, 0.46 on BioDiv-3DTrees-ULS) is the within-species balanced-accuracy chance, i.e. the mean over species of one over the number of sources that species spans, and is therefore *not* $1/n_{\text{source}}$ (which would be 0.048 and 0.333). This within-species probe also differs from the all-tree source-recovery probe of Section 6.1 (chance $1/n_{\text{source}} = 0.048$ on FOR-species20K), which predicts source across all trees rather than within one species. The advantage survives source control but attenuates, and the random-initialized control sits below geometry on both datasets (random-init ViT within-source NMI minus geometry is −0.039 on FOR-species20K and −0.005 on BioDiv-3DTrees-ULS).

	Within-source NMI				Within-species source recovery (bal. acc.)			
Dataset	CLIP	DINOv2	geo	CLIP−geo	CLIP	DINOv2	geo	chance
FOR-species20K (n=21 blocks)	0.174	0.187	**0.130**	+0.044	**0.73**	**0.73**	0.50	0.26
BioDiv-3DTrees-ULS (n=3 blocks)	**0.265**	**0.241**	0.061	**+0.204**	0.61	0.59	**0.52**	**0.46**

size-aware 17-dimensional variant (absolute-size dimensions added to the Weinmann descriptor) narrows but does not close the gap. Frozen CLIP still leads on both datasets, globally (0.275 versus 0.216 on FOR-species20K, 0.350 versus 0.173 on BioDiv-3DTrees-ULS) and within source (0.174 versus 0.143, and 0.265 versus 0.064). Absolute size does carry species information, lifting the FOR-species20K geometric NMI from 0.102 to 0.216, but we retain the size-invariant descriptor as the primary baseline. Absolute size is partly an acquisition and normalization artifact across heterogeneous scanners, and the frozen-feature advantage survives either choice.

Local multi-scale geometry. A sharper test replaces the whole-tree single-scale covariance descriptor with a true local multi-scale Weinmann descriptor (Weinmann et al., 2015). For each tree we compute per-point eigenfeatures (linearity, planarity, sphericity, anisotropy, eigenentropy, and change-of-curvature) in $k = 10$, 20, and 40 nearest-neighbour neighbourhoods, average them over the tree, and run the result through the identical harness. This local descriptor is substantially stronger than the single-scale baseline. It more than doubles geometric NMI on FOR-species20K (0.227 versus 0.102) and lifts it on BioDiv-3DTrees-ULS (0.192 versus 0.101). Frozen CLIP still leads on both datasets, but by a much narrower margin than against the single-scale descriptor, namely +0.048 NMI on FOR-species20K (0.275 versus 0.227) and +0.158 on BioDiv-3DTrees-ULS (0.350 versus 0.192). The large global ratio over single-scale geometry (Section 5.1) therefore partly reflects the weakness of that baseline. Within source, however — where acquisition structure can no longer inflate either representation — the CLIP advantage over even this stronger baseline is larger still. The local descriptor's high global NMI on FOR-species20K (0.227) is largely acquisition-source structure: reclustered inside each source block it falls to a within-source NMI of only 0.083, whereas frozen CLIP holds at 0.175. The within-source CLIP-minus-local-Weinmann residual is thus +0.092 on FOR-species20K — roughly double the +0.048 global margin — because most of the local descriptor's global coherence is source rather than species. On BioDiv-3DTrees-ULS the same within-source residual is +0.104 (CLIP 0.263, local multi-scale Weinmann 0.160). Measured against a properly local, multi-scale Weinmann descriptor the frozen-feature advantage therefore survives, modest at the global level on the multi-sensor FOR-species20K but widening once the baseline's own source inflation is removed. We read the headline geometry gap accordingly, as an advantage over a standard whole-tree descriptor rather than over the strongest hand-crafted geometry available.

Depth versus RGB inputs. Because depth renders discard color, a natural question is whether radiometric input would improve performance where structure is ambiguous. On the reliable NEON-MLBS site — an airborne NEON/IDTReeS plot at the Mountain Lake Biological Station, Virginia, USA ($n = 543$ trees, six species), distinct from the three LiDAR datasets of Section 4.1 and used only for this depth-versus-RGB boundary case — the ranking is metric-dependent. CLIP on RGB renders leads on macro-F1, geometry leads on absolute Overall Accuracy, and the depth render is best on neither (Table 7). No representation is clearly best, and with this sample ($n = 543$, six species) no significance is available. We therefore treat the choice between depth and RGB inputs as an open question for future multimodal work rather than one this study resolves.

Relation to the FOR-species20K leaderboard. We do not rank against the published FOR-species20K accuracies. (For external context, Puliti et al. (2025) report 2D and 3D family means of 0.77 and 0.72 Overall Accuracy, with Detail-View highest at 79.5%.) That protocol differs fundamentally, using random splitting, supervised end-to-end training, and the full 33-species set. Within the stated bounds (Hungarian upper bound, global rather than spatial evaluation), the result holds across the two datasets at matched dimension. Frozen image-pretrained features cluster trees by species more coherently than the hand-crafted geometric descriptor.

6.4. Limitations

Several boundaries qualify how these results should be read. The Hungarian-matched macro-F1, Overall Accuracy,

Table 7
Depth versus RGB on the reliable NEON-MLBS site (n = 543, 6 species, K = 6). Unsupervised Hungarian-matched macro-F1 and Overall Accuracy for frozen CLIP on RGB renders, frozen CLIP on PCA-rotated depth renders, and the traditional geometric descriptor. CLIP-RGB leads on macro-F1, the geometric descriptor on Overall Accuracy, and the depth render on neither, and the differences are not significant at this sample size.

Method	macro-F1	Overall Accuracy
CLIP-RGB	**0.226**	0.278
CLIP-struct (depth)	0.206	0.263
geo	0.189	**0.298**

and mIoU consult the test labels post hoc to select the most favorable cluster-to-species permutation. They are therefore an upper bound on cluster quality rather than label-free accuracy. This is why matching-free Normalized Mutual Information is our primary metric (Section 4.3). The absolute levels remain modest even at that upper bound (Normalized Mutual Information 0.278 to 0.355), so the defensible claim is strictly relative, that frozen depth features are more species-aligned than the geometric descriptor, not that clustering resolves species into well-separated clusters. The evaluation is global over the full analyzable pool rather than a spatially blocked split, so it licenses no cross-site or cross-region generalization claim. Because FOR-species20K pools heterogeneous TLS, MLS, and ULS clouds from 21 source studies, acquisition source is partly confounded with species, a confound that the within-source control of Section 6.1 attenuates but does not remove. On BioDiv-3DTrees the dominant *Fagus* class was capped (after which *Picea* is the majority), so we claim only that *Fagus* was capped, not that the classes are balanced. Finally, the contribution is an unsupervised structure-alignment finding only, since the supervised arm against the identical 11-d geometric baseline is underpowered and inconclusive (Section 5.2). Two further boundaries follow from the new comparisons. First, the larger surviving-species residual (+0.204 within-source NMI) rests substantially on one small, single-sensor, single-country dataset (BioDiv-3DTrees-ULS: n = 2,250, six species, three German regions), so its external validity is limited. The multi-sensor FOR-species20K residual is by contrast small (+0.044), and the random-initialized control's edge over geometry is negligible (+0.006 NMI on FOR-species20K, global at matched dimension, and below geometry within source on both datasets), which is what isolates pre-training rather than the pipeline as the cause. Second, the frozen 3D foundation-model comparison of Section 5.6 is an availability comparison rather than an architectural verdict, since the OpenShape baseline is CAD/ShapeNet-pre-trained and therefore out-of-distribution for forest LiDAR. The features are also depth-only, so RGB or hyperspectral fusion remains future work.

7. Conclusion

Species-level interpretation of LiDAR tree point clouds is constrained by a labeling bottleneck, since expert per-stem annotation is expensive, slow, and geographically biased. The label-free alternative, a hand-crafted geometric descriptor (a Weinmann-type covariance summary (Weinmann et al., 2015)), is confined to a small set of predefined covariance and height statistics. To relieve this bottleneck, we propose an unsupervised framework that reads tree structure through frozen vision-foundation features rather than through learned point-cloud supervision or hand-crafted geometry. We render each LiDAR point cloud as six PCA-rotated depth views, encode them with a frozen image tower, fuse the views into a single per-tree embedding, and recover species by standard clustering, with no species label or text prompt at any stage. To test this framework fairly, we design a matching-free evaluation protocol that is dimension-matched, source-controlled, permutation-tested, and applied identically to every backbone. We then establish that the resulting advantage is a property of pre-training in general rather than of one model.

The central finding is a claim about representation rather than an enumeration of benchmark scores. At matched dimension, frozen depth features organize trees by species markedly more coherently than a whole-tree single-scale covariance descriptor. This relative cluster-coherence advantage is roughly 2.7 to 3.5× on NMI across two independent datasets (Section 5.1); against a fair local multi-scale Weinmann baseline the margin is far smaller (+0.048 and +0.158 NMI, Section 6.3). That global ratio is a source-inflated upper bound, whose source-controlled counterpart is the smaller within-source residual (Section 6.1). The advantage is preserved almost in full when the embedding is compressed to geometry's dimensionality (Section 5.2), which rules out capacity as the explanation and places it in feature content. A random-initialized control stays at geometry's level. This isolates pre-training as the cause, not the pipeline that renders, encodes, and clusters. The gap is also significant against a label-permutation null. Because the within-source control attenuates the advantage to a positive residual, this is a source-attenuated structure-alignment finding rather than accurate species classification. For practitioners, the operational implication is direct: unsupervised foundation-model species clusters obtained from pooled multi-sensor LiDAR should not be read as species maps, and acquisition source should be stratified out before the clusters are probed for species structure. Future work should disentangle species morphology from acquisition source and fuse the structural depth view with co-registered RGB and hyperspectral signals to raise absolute coherence. A further direction is to train a foundation model purpose-built for tree LiDAR point clouds, which could raise unsupervised tree-species accuracy and remove the inefficiency of manual species labels.

8. Statements and Declarations

Declaration of generative AI in the writing process. During the preparation of this work the author used generative AI tools to assist with language editing and the organization of the manuscript. After using these tools the author reviewed and edited the content as needed and takes full responsibility for the content of the publication. Generative AI was not used to produce, analyze, or interpret the scientific results, which derive solely from the described experiments.

Data availability. All four datasets are publicly available: FOR-species20K (Zenodo, https://doi.org/10.5281/zenodo.13255198), BioDiv-3DTrees (GRO.data, https://doi.org/10.25625/8PB1IF), Weiser/SYSSIFOSS (PANGAEA, https://doi.org/10.1594/PANGAEA.942856), and NEON-MLBS, from the NEON/IDTReeS airborne benchmark (Graves et al., 2023).

Code availability. Code to reproduce the analysis has been prepared for release; it is provided to the editors as an anonymized artifact for peer review and will be released publicly with a citable archive (Zenodo DOI) upon acceptance.

Competing interests. The author declares no competing interests.

Funding. This research received no external funding.

CRediT authorship contribution statement. The sole author is responsible for all aspects of the work: conceptualization, methodology, software, formal analysis, investigation, data curation, writing (original draft and revision), and visualization.

References

Abuhani, D.A., Seccaroni, M., Mazzarello, M., Zualkernan, I., Duarte, F., Ratti, C., 2026. Unsupervised mapping of urban tree diversity using spatially-aware visual clustering. Scientific Reports 16, 10440. doi:10.1038/s41598-026-37043-7.

Adaloglou, N., Michels, F., Kalisch, H., Kollmann, M., 2023. Exploring the limits of deep image clustering using pretrained models, in: British Machine Vision Conference (BMVC). ArXiv:2303.17896.

Allen, M.J., Grieve, S.W.D., Owen, H.J.F., Lines, E.R., 2023. Tree species classification from complex laser scanning data in Mediterranean forests using deep learning. Methods in Ecology and Evolution doi:10.1111/2041-210X.13981.

Asano, Y.M., Rupprecht, C., Vedaldi, A., 2020. Self-labelling via simultaneous clustering and representation learning, in: International Conference on Learning Representations (ICLR). ArXiv:1911.05371.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 9912–9924. ArXiv:2006.09882.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9650–9660. ArXiv:2104.14294.

Chen, Z., Xu, H., Chen, W., Zhou, Z., Xiao, H., Sun, B., Xie, X., Kang, W., 2023. PointDC: Unsupervised semantic segmentation of 3D point clouds via cross-modal distillation and super-voxel clustering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14290–14299. ArXiv:2304.08965.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S., 2022. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery, in: Advances in Neural Information Processing Systems (NeurIPS). ArXiv:2207.08051.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR). ArXiv:2010.11929.

Fassnacht, F.E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L.T., Straub, C., Ghosh, A., 2016. Review of studies on tree species classification from remotely sensed data. Remote Sensing of Environment 186, 64–87. doi:10.1016/j.rse.2016.08.013.

Goyal, A., Law, H., Liu, B., Newell, A., Deng, J., 2021. Revisiting point cloud shape classification with a simple and effective baseline, in: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 3809–3820. ArXiv:2106.05304.

Graves, S.J., Marconi, S., Stewart, D., Harmon, I., Weinstein, B.G., Kanazawa, Y., Scholl, V.M., Joseph, M.B., McGlinchy, J., Browne, L., Allen, M.K., Bohlman, S.A., et al., 2023. Data science competition for cross-site individual tree species identification from airborne remote sensing data. PeerJ 11, e16578. doi:10.7717/peerj.16578.

Griese, N., Ritzert, M., Nölke, N., 2025. A large dataset of labelled single tree point clouds, QSMs and tree graphs. Scientific Data 12, 1953. doi:10.1038/s41597-025-06421-7.

Hartling, S., Sagan, V., Sidike, P., Maimaitijiang, M., Carron, J., 2019. Urban tree species classification using a WorldView-2/3 and LiDAR data fusion approach and deep learning. Sensors 19, 1284. doi:10.3390/s19061284.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009. ArXiv:2111.06377.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L., 2021. OpenCLIP. https://github.com/mlfoundations/open_clip. doi:10.5281/zenodo.5143773.

Ji, X., Henriques, J.F., Vedaldi, A., 2019. Invariant information clustering for unsupervised image classification and segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9864–9873. doi:10.1109/ICCV.2019.00996.

Kanezaki, A., Matsushita, Y., Nishida, Y., 2018. RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5010–5019. ArXiv:1603.06208.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R., 2023. Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026. ArXiv:2304.02643.

Kohek, Š., Žalik, B., Mongus, D., Strnad, D., 2026. Community detection framework based on 3D shape descriptors for tree species classification in point cloud data. Scientific Reports 16, 12091. doi:10.1038/s41598-026-42392-4.

Liu, B., Huang, H., Su, Y., Chen, S., Li, Z., Chen, E., Tian, X., 2022. Tree species classification using ground-based LiDAR data by various point cloud deep learning methods. Remote Sensing 14, 5733. doi:10.3390/rs14225733.

Lu, D., Xie, Q., Gao, K., Xu, L., Li, J., 2022. 3DCTN: 3D convolution-transformer network for point cloud classification. IEEE Transactions on Intelligent Transportation Systems 23, 24854–24865. doi:10.1109/TITS.2022.3198836.

Ma, H., Zhang, F., Chen, S., Yu, J., 2025. Individual tree segmentation using deep learning and climbing algorithm: A method for achieving high-precision single-tree segmentation in high-density forests under complex environments. Photogrammetric Engineering & Remote Sensing 91, 101–110. doi:10.14358/PERS.24-00083R2.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J.,

Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research ArXiv:2304.07193.

Persson, Å., Holmgren, J., Söderman, U., 2002. Detecting and measuring individual trees using an airborne laser scanner. Photogrammetric Engineering & Remote Sensing 68, 925–932.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Bénedet, F., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11, 4540. doi:10.1038/s41467-020-18321-y.

Popescu, S.C., Wynne, R.H., 2004. Seeing the trees in the forest: Using LiDAR and multispectral data fusion with local filtering and variable window size for estimating tree height. Photogrammetric Engineering & Remote Sensing 70, 589–604.

Puliti, S., Lines, E.R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M.J., Winiwarter, L., Rehush, N., Hristova, H., Murray, B., Calders, K., Coops, N., Höfle, B., Irwin, L., Junttila, S., Krůček, M., Krok, G., Král, K., Levick, S.R., Luck, L., Missarov, A., Mokroš, M., Owen, H.J.F., Stereńczak, K., Pitkänen, T.P., Puletti, N., Saarinen, N., Hopkinson, C., Terryn, L., Torresan, C., Tomelleri, E., Weiser, H., Astrup, R., 2025. Benchmarking tree species classification from proximally sensed laser scanning data: Introducing the FOR-species20K dataset. Methods in Ecology and Evolution 16, 801–818. doi:10.1111/2041-210X.14503.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 652–660. ArXiv:1612.00593.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space, in: Advances in Neural Information Processing Systems (NeurIPS). ArXiv:1706.02413.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 8748–8763.

Rizaldy, A., Fassnacht, F.E., Afifi, A.J., Jiang, H., Gloaguen, R., Ghamisi, P., 2025. Label-efficient 3D forest mapping: Self-supervised and transfer learning for instance segmentation, semantic segmentation, and species classification. arXiv preprint ArXiv:2511.06331.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. doi:10.1111/ecog.02881.

Seidel, D., Annighöfer, P., Thielman, A., Seifert, Q.E., Thauer, J.H., Glatthorn, J., Ehbrecht, M., Kneib, T., Ammer, C., 2021. Predicting tree species from 3D laser scanning point clouds using deep learning. Frontiers in Plant Science 12, 635440. doi:10.3389/fpls.2021.635440.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3D shape recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 945–953. ArXiv:1505.00880.

Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J., 2019. KPConv: Flexible and deformable convolution for point clouds, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6411–6420. ArXiv:1904.08889.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L., 2020. SCAN: Learning to classify images without labels, in: Computer Vision – ECCV 2020, pp. 268–285. doi:10.1007/978-3-030-58607-2_16.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics (TOG) 38, 1–12. doi:10.1145/3326362.

Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. ISPRS Journal of Photogrammetry and Remote Sensing 105, 286–304. doi:10.1016/j.isprsjprs.2015.01.016.

Weiser, H., Schäfer, J., Winiwarter, L., Krašovec, N., Fassnacht, F.E., Höfle, B., 2022. Individual tree point clouds and tree measurements from multi-platform laser scanning in German forests. Earth System Science Data 14, 2989–3012. doi:10.5194/essd-14-2989-2022.

Xiao, A., Xuan, W., Wang, J., Huang, J., Tao, D., Lu, S., Yokoya, N., 2024. Foundation models for remote sensing and earth observation: A survey. arXiv preprint ArXiv:2410.16602.

Xie, J., Girshick, R., Farhadi, A., 2016. Unsupervised deep embedding for clustering analysis, in: Proceedings of the 33rd International Conference on Machine Learning (ICML), pp. 478–487.

Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S., 2023. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1179–1189. ArXiv:2212.05171.

Zhang, B., Zhao, L., Zhang, X., 2020. Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images. Remote Sensing of Environment 247, 111938. doi:10.1016/j.rse.2020.111938.

Zhang, C., Qiu, F., 2012. Mapping individual tree species in an urban forest using airborne LiDAR data and hyperspectral imagery. Photogrammetric Engineering & Remote Sensing 78, 1079–1087. doi:10.14358/PERS.78.10.1079.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H., 2022. PointCLIP: Point cloud understanding by CLIP, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8542–8552. doi:10.1109/CVPR52688.2022.00836.

Zhang, Z., Yang, B., Wang, B., Li, B., 2023. GrowSP: Unsupervised semantic segmentation of 3D point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17619–17629. doi:10.1109/CVPR52729.2023.01690.

Zhao, H., Jiang, L., Jia, J., Torr, P.H.S., Koltun, V., 2021. Point transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 16259–16268. ArXiv:2012.09164.

Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P., 2023. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2639–2650. doi:10.1109/ICCV51070.2023.00249.

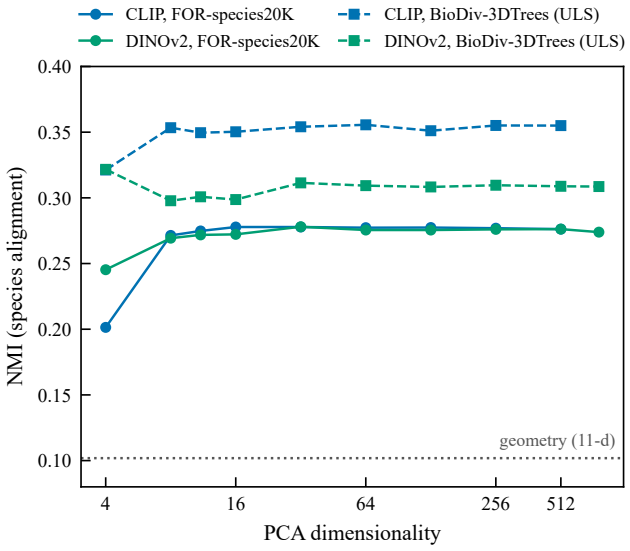

Fig. 4. PCA dimension sweep. The unsupervised NMI of CLIP (solid) and DINOv2 (dashed) is essentially flat from ≈11 dimensions up to each backbone's native width (512 for CLIP, 768 for DINOv2) on both datasets and exceeds the 11-dimensional geometric descriptor (dotted) at every dimension. Even a 4-dimensional embedding outperforms 11-dimensional geometry.

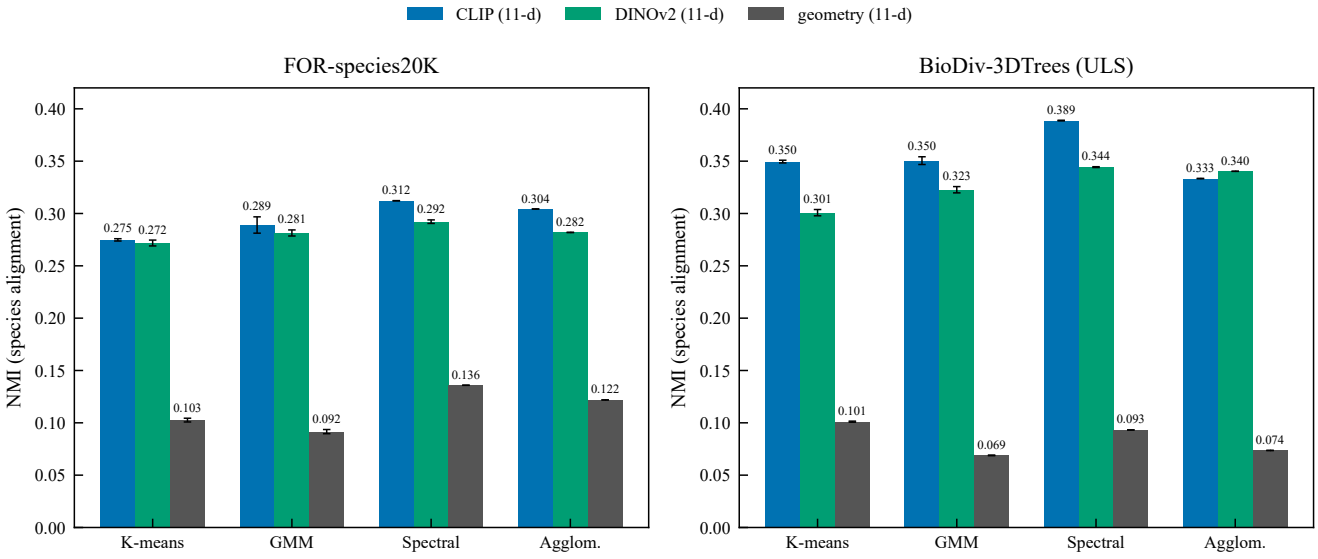

Fig. 5. Clustering-algorithm robustness. Unsupervised NMI at matched 11-d for four fixed-K algorithms. CLIP and DINOv2 each exceed geometry on all four methods and both datasets.

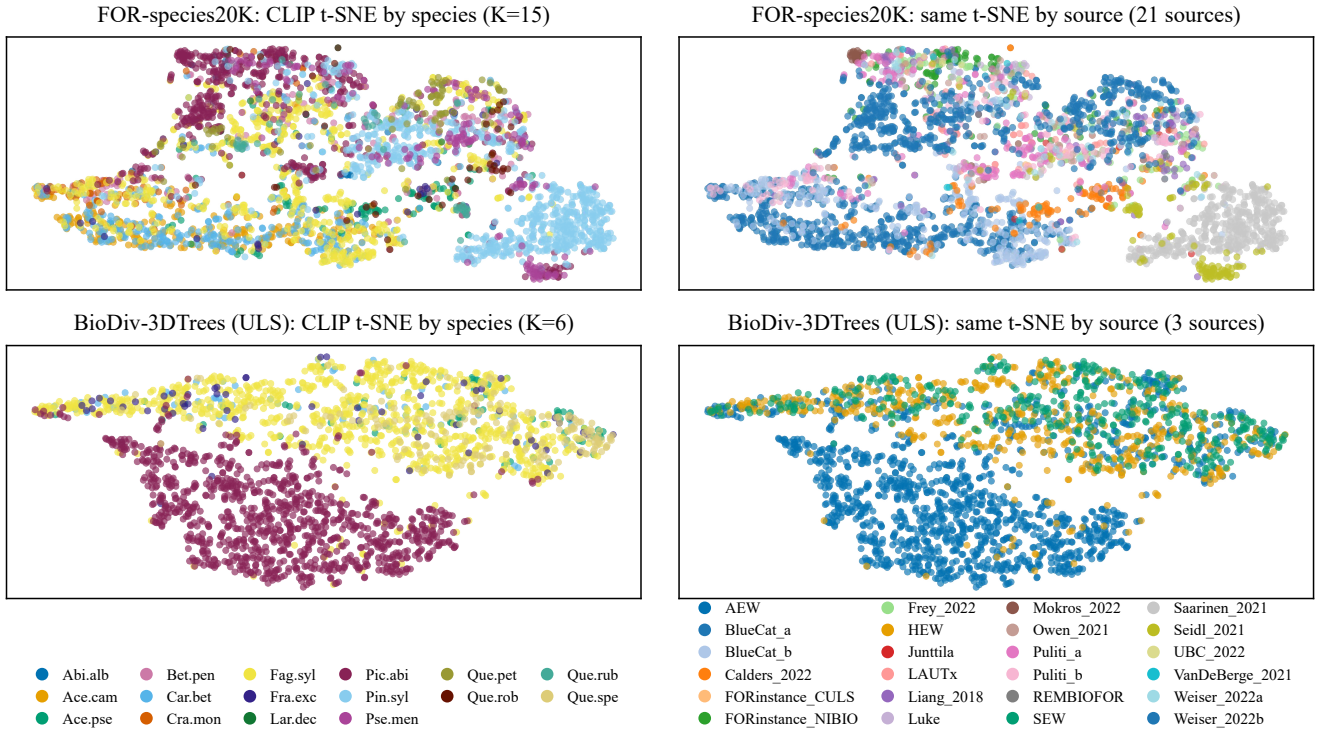

Fig. 6. Qualitative cluster structure (CLIP-11d), t-SNE projections. For each dataset, with FOR-species20K in the top row and BioDiv-3DTrees in the bottom row, the projection is coloured by species (left) and by acquisition source or region (right). Species-coherent structure is visible but partially aligns with source. The FOR-species20K projection uses a 3,000-tree subsample for legibility.

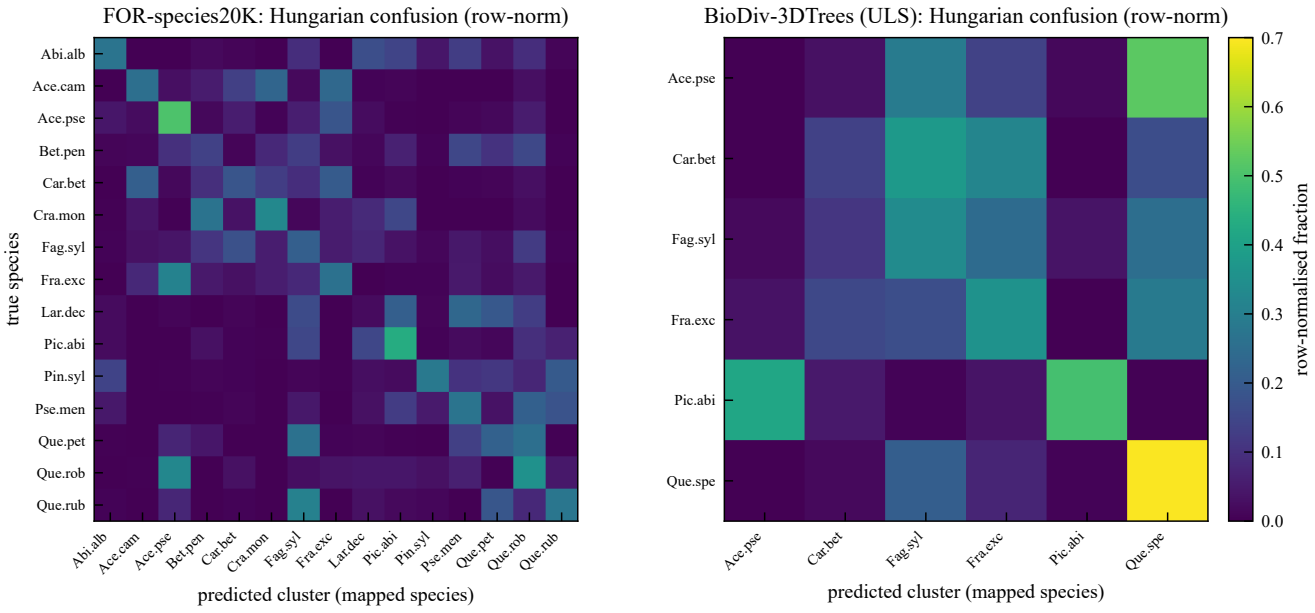

Fig. 7. Row-normalised Hungarian-matched confusion matrices (CLIP-11d) for FOR-species20K (left) and BioDiv-3DTrees (right), with abbreviated species codes on both axes. Per-species recovery is moderate, and the off-diagonal mass concentrates among structurally similar species rather than scattering uniformly.

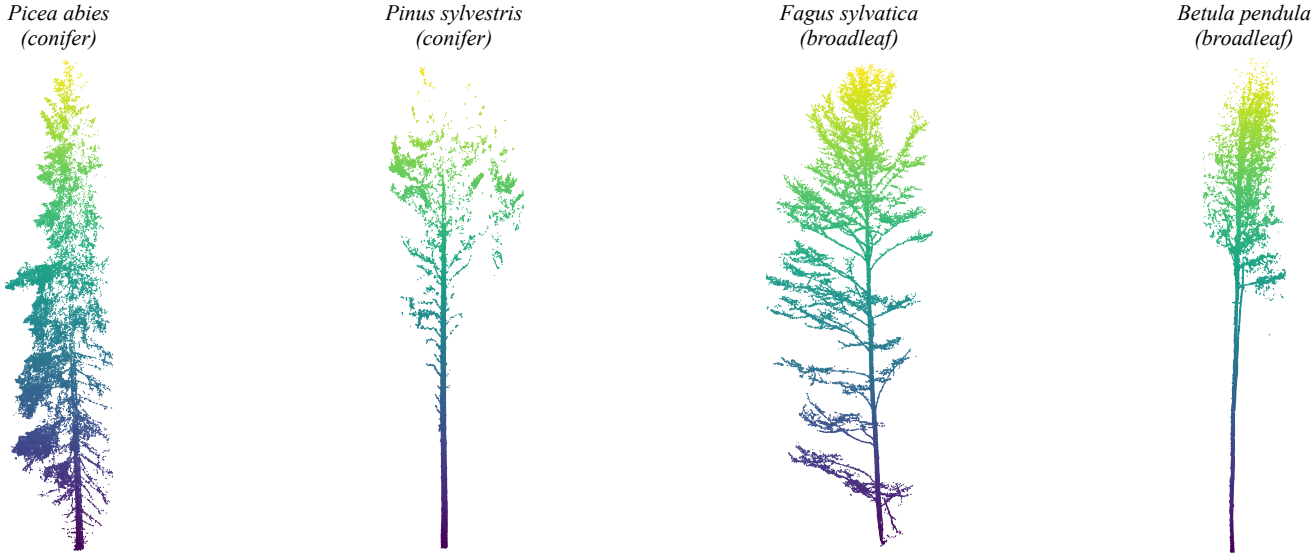

Fig. 8. Representative real LiDAR point clouds of four FOR-species20K species, two conifers (*Picea abies*, *Pinus sylvestris*) and two broadleaves (*Fagus sylvatica*, *Betula pendula*). The hard confusions in Figure 7 concentrate among structurally similar species rather than among same-genus pairs.