LEJEPA FOR HYPERSPECTRAL TREE SPECIES CLASSIFICATION

Miika Koivisto [1] †, Eeli Räsänen [1] †, Eija Honkavaara[2], Anna-Maria Raita[1], Ilkka Pölönen[1]

[1]Spectral Imaging Laboratory, Faculty of Information Technology, University of Jyväskylä, Finland,
[2] Finnish Geospatial Research Institute, National Land Survey of Finland, Finland

ABSTRACT

Tree species classification is important for forest management, carbon accounting, biodiversity monitoring, disturbance detection, and risk assessment, as it reveals which species occupy given areas and how forests change over time. Hyperspectral (HS) imaging provides detailed spectral information, enabling the differentiation of tree species and forest conditions that may appear similar in conventional imagery. The main barrier to the use of such classification techniques is often the lack of annotated training data. Self-Supervised Learning (SSL) has recently transformed computer vision by enabling pre-training on large unlabelled datasets, but its application to HS data remains less explored.

This study presents the first LeJEPA-based approach for tree species classification from UAV HS imagery, employing a single 3D Convolutional Neural Network (CNN) encoder with Sketched Isotropic Gaussian Regularisation (SIGReg) distribution matching and an invariance loss without Exponential Moving Average (EMA), predictor, or stop-gradient. The model was pre-trained on a combination of labelled crowns (species labels withheld) and automatically detected unlabelled crowns from the Vesijako 2014 dataset. The frozen encoder achieved 91.5% linear-probe Overall Accuracy (OA), while the supervised baseline reached 96.2%. Per-class k-Nearest Neighbour (k-NN) accuracy reached 98.4% for pine, 76.4% for spruce, 90.5% for birch and 100.0% for larch.

***Index Terms*—** Hyperspectral imaging, tree species classification, self-supervised learning, joint-embedding predictive architecture, 3D convolutional neural network, sketched isotropic Gaussian regularisation, HSI, SSL, JEPA, LeJEPA, 3D CNN, SIGReg.

1. INTRODUCTION

1.1. Spectral foundational models

We have recently seen a surge of spectral foundation models based on generative pre-training. SpectralGPT [1] introduced a 3D generative transformer pre-trained on more than one million Sentinel-2 images via multi-target reconstruction. S2MAE [2] proposed a spatial-spectral masked autoencoder that uses 90% 3D masking with compact cube tokens to capture local spectral continuity. HyperSIGMA [3] scaled HS representation learning to more than a billion parameters using sparse sampling attention and masked image modelling. These generative approaches operate in pixel space: they reconstruct masked or corrupted input pixels, requiring the model to allocate capacity to low-level texture prediction rather than high-level semantic understanding.

An alternative paradigm is the Joint-Embedding Predictive Architecture (JEPA) [4], which avoids pixel-level reconstruction and instead predicts the representation of one image region from another directly in latent space. By operating in an abstract embedding space, JEPA captures the underlying factors of variation such as object shape, species-specific spectral curves, and crown structure, without allocating capacity to predictable pixel details.

1.2. Main outcomes: tree species in latent space

The LeJEPA formulation [5] operationalises this principle with a single encoder, Sketched Isotropic Gaussian Regularisation (SIGReg) distribution matching, and an invariance loss without an Exponential Moving Average (EMA), a predictor, or a stop-gradient, making it a simpler alternative to both generative pre-training and contrastive Self-Supervised Learning (SSL). Although concurrent work such as CARL [6] has shown that JEPA-based methods can provide effective backbones for sensor-agnostic spectral representation learning, their use in tree species classification has not yet been explored.

This is a particularly relevant gap, as tree species classification involves high sensor diversity across field campaigns and severe label scarcity. By showing that predictive SSL generalises to individual tree crown classification, this work provides domain-specific validation for the broader future development of multi-domain, sensor-agnostic JEPA models. Our main outcome is the first LeJEPA-based approach for HS tree species classification. The model jointly uses spectral reflectance and canopy height data through a predictive embedding objective with a single 3D Convolutional Neural Network (CNN) encoder. We validate the approach with the

† Equal contribution.

This work was supported by Research Council of Finland grants (357382, 357380 and 377633).

Vesijako 2014 UAV Hyperspectral (HS) dataset and compare it directly to a supervised 3D-CNN baseline [7] trained with the same data.

2. MATERIAL AND METHODS

2.1. Data

The experiments were performed using the Vesijako 2014 dataset [8]. The Vesijako 2014 dataset comprises nine test sites with 3469 labelled tree crowns: pine (*Pinus sylvestris*; 2249), spruce (*Picea abies*; 584), birch (*Betula pendula*; 519), and larch (*Larix sibirica*; 117). Each labelled crown is represented by a 39-channel multimodal patch: 38 spectral bands and the normalised Digital Surface Model (DSM). In addition to the labelled crowns, 19501 unlabelled crowns were automatically detected from the DSM to expand the self-supervised pre-training dataset.

The unlabelled crowns were detected by applying a fixed-radius ($15\,\text{px} = 1.5\,\text{m}$) local-maximum search in the DSM with an absolute elevation threshold set at the area's first percentile DSM minimum plus $5.0\,\text{m}$. A minimum canopy height of $5.0\,\text{m}$ above ground and a minimum distance of $64\,\text{px}$ ($6.4\,\text{m}$) from any labelled crown centre were enforced to ensure the separation between training and validation data. The detection follows the local-maxima-based methodology [8].

2.2. Data preprocessing

Each tree crown was extracted as a centred 64×64 spatial patch from the bounding box and stacked into a 39-channel tensor comprising 38 spectral bands and the normalised DSM. At extraction time, the DSM was normalised per area by subtracting the minimum elevation of the area's first percentile, following [7]. During training, the spectral channels were L1-normalised to unit sum per spectrum, which preserves spectral curve shape while removing illumination magnitude variation. The DSM channel was additionally standardised via per-batch z-score normalisation. Both normalisations are applied inside the GPU augmentation pipeline. Of the 3469 crowns labelled, 2428 (70%) were used for training and 1041 (30%) for validation, stratified by species and site. The 19501 unlabelled crowns were used exclusively for SSL pre-training, bringing the total training set to 21929 crowns.

2.3. Joint-Embedding Predictive Architecture

JEPAs learn representations by predicting the embedding of one view of an input from another, operating entirely in latent space rather than reconstructing pixels [4]. This predictive objective encourages the encoder to capture semantic factors of variation such as crown shape, species-specific spectral curves, or canopy structure, while avoiding the allocation

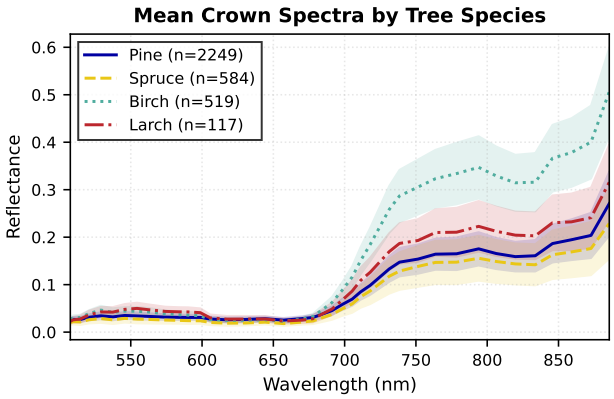


Fig. 1: Mean reflectance spectra $\pm$ 1 standard deviation per tree species, computed from 3469 labelled crowns across all nine test sites. Larch (117) and birch (519) are the most spectrally distinct, while spruce (584) and pine (2249) show greater overlap in the red-edge and NIR regions.

of capacity to low-level pixel details that generative models must reconstruct.

Each tree crown is represented as a 64×64 spatial patch ($6.4 \times 6.4\,\text{m}$ at $0.1\,\text{m}$ GSD) comprising 39 channels: 38 spectral reflectance bands and the normalised DSM.

The encoder (Figure 2) is a lightweight 3D-CNN (1.7M backbone parameters, Table 1) followed by a 3-layer Multilayer Perceptron (MLP) projector (1.2M parameters, 2.9M total). The first convolution uses a $(1, 3, 3)$ kernel along the (D, H, W) axes that extracts per-band spatial texture without any spectral mixing, followed by three $(3, 3, 3)$ convolutions and asymmetric max-pooling that defer spectral reduction to later layers while progressively compressing spatial dimensions. The padding in the final convolutional stage preserves a 4×4 spatial grid before AdaptiveAvgPool3d reduces it to $3 \times 3 \times 3$. The output is flattened and projected through a 3-layer MLP (72–1024–1024–72) with GELU activation, BatchNorm after the first Linear layer, and no bias on the first linear layer, producing a 72-dimensional embedding.

For each tree, $V = 8$ views are generated through random crops and spectral transformations. $V_g = 2$ global views are produced by random crops of 32–48 px resized to 48×48, and six local views by crops of 8–16 px, also resized to 48×48. Spatial flips and 90° rotations are applied synchronously across all 39 channels. Spectral augmentations include Gaussian noise, brightness scaling, spectral slope perturbation, contiguous band masking, and band dropout.

The LeJEPA loss [5] combines SIGReg, which prevents representational collapse, with a prediction (invariance) term that pulls each view embedding toward the centroid of the $V_g = 2$ global views:

$$\mathcal{L}_{\text{LeJEPA}} = \lambda\, \mathcal{L}_{\text{SIGReg}} + (1 - \lambda)\, \mathcal{L}_{\text{inv}}, \tag{1}$$

with $\lambda = 0.05$. $\mathcal{L}_{\text{SIGReg}}$ is the Epps–Pulley test statistic com-

puted on random projections of the embeddings (see [5], Section 4), and $\mathcal{L}_{\text{inv}}$ is the mean squared error between each view embedding and the centroid of the $V_g = 2$ global views. This simplified loss form follows the official reference code [5]. One departure from the reference is that we compute the centroid from the V_g global views rather than from all V views. This follows Algorithm 2 of [5], where the higher-resolution global views define the prediction target.

SIGReg constrains the embedding distribution towards an isotropic Gaussian, identified as the theoretically optimal distribution for minimising downstream prediction risk. It projects each embedding $z \in \mathbb{R}^D$ onto $K = 1024$ random unit vectors, computes the Empirical Characteristic Function (ECF) of the projected samples using the Epps–Pulley test, and minimises its weighted L^2 distance to the $N(0, 1)$ characteristic function via trapezoidal quadrature (17 knots, $t \in [-5, 5]$). In the distributed setting, projected statistics are synchronised via `all_reduce` across GPUs to maintain a global empirical characteristic function estimate. The invariance term minimises the mean squared error between each view embedding and the centroid of the $V_g = 2$ global views.

Table 1: 3D-CNN backbone architecture. Kernel and output shapes are in $(Depth, Height, Width, Channels)$ order. The backbone contains 1.7M parameters with an embedding dimension of 72; the MLP projector adds 1.2M (not shown).

Layer	Kernel / pool (D,H,W)	Output (D,H,W,C)	Params
Conv3D+BN+ReLU	$(1, 3, 3)$, s= $(1, 2, 2)$	$(39, 23, 23, 64)$	704
Conv3D+BN+ReLU	$(3, 3, 3)$	$(37, 21, 21, 64)$	110 720
MaxPool3D	$(1, 2, 2)$	$(37, 10, 10, 64)$	0
Conv3D+BN+ReLU	$(3, 3, 3)$	$(35, 8, 8, 128)$	221 440
MaxPool3D	$(3, 2, 2)$	$(11, 4, 4, 128)$	0
Conv3D+BN+ReLU	$(3, 3, 3)$	$(11, 4, 4, 256)$	885 248
MaxPool3D	$(2, 1, 1)$	$(5, 4, 4, 256)$	0
AdaptiveAvgPool3D	$(3, 3, 3)$ target	$(3, 3, 3, 256)$	0
Flatten	—	(6912)	0
Dense + ReLU	—	(72)	497 736
Dropout (0.25)	—	(72)	0
Dense	—	(72)	5 256
Total parameters			**1 721 104**

2.4. Training details

The model was trained using the $AdamW$ optimiser with a peak learning rate of 5×10^{-4}, linear warm-up over 10 epochs, cosine decay to 10^{-6}, and weight decay 5×10^{-4} (excluded from BatchNorm parameters and biases). The total batch size across GPUs was 128 (64 per GPU) with gradient clipping at 5.0. Automatic Mixed Precision (AMP) was used via `torch.amp.GradScaler` and `autocast`. Pretraining ran for 200 epochs on 2 NVIDIA A4000 16GB GPUs using DistributedDataParallel, completing in approximately 7.5 hours with peak VRAM usage of 15.7 GB per GPU.

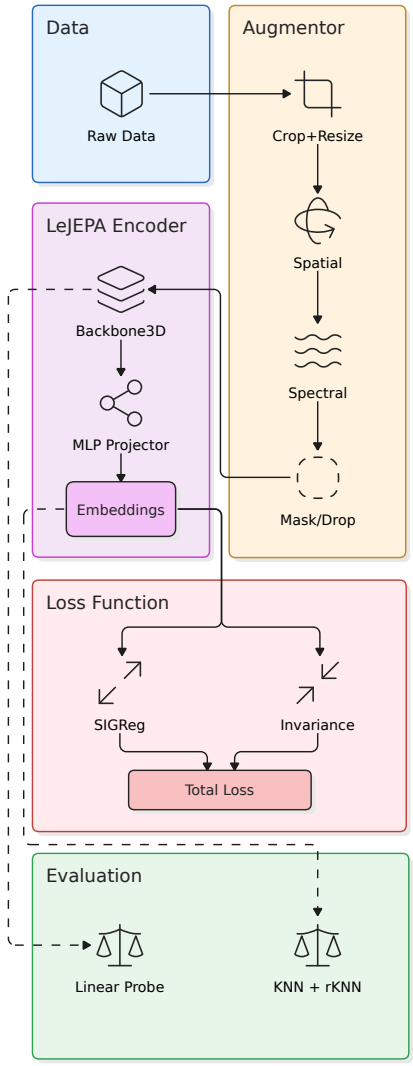


Fig. 2: Architecture overview. A single Backbone3D encoder processes all $V = 8$ views; the LeJEPA loss combines SIGReg and the invariance term.

2.5. Evaluation metrics

The frozen pre-trained encoder was evaluated on tree species classification using three protocols: a k-Nearest Neighbour (k-NN) classifier ($k = 20$, cosine distance) in the 72-dimensional projector space, an adaptive-radius variant (rKNN; radius at the fifth percentile of training pairwise distances), and a Linear Probe (LP) (logistic regression) fitted on the backbone output features. The k-NN variants operate in the projector space because SIGReg constrains it toward an isotropic Gaussian, shown to be the theoretically optimal distribution for nearest-neighbour retrieval [5], while the LP

evaluates the raw backbone representations that carry the full discriminative signal before any invariance modifications.

The approach was compared to the fully supervised 3D-CNN architecture [7], which achieved 96.2% Overall Accuracy (OA) on the Vesijako 2014 dataset. That study used the same field sites and tree crowns as ours but trained end-to-end on three species only (pine, spruce and birch), excluding larch due to its limited occurrence (117 crowns in a single site). The objective is to assess how closely a predictive self-supervised approach can approach this supervised upper bound without using any labels during pre-training.

OA for three and four species, and per-class accuracy are reported. The 19,501 unlabelled crowns detected by the DSM are excluded from supervised evaluation; only the 3,469 labelled crowns are used for evaluation. All metrics are reported on a held-out validation set consisting of a random 70/30 split of labelled data stratified by species and area. The embedding space was visualised using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) [9] projections coloured by species to qualitatively assess class separation.

3. RESULTS

3.1. Comparison Against Supervised Baseline

The performance of frozen projector embeddings was measured using methods described in Section 2.5. These results are summarised in Table 2.

Table 2: Classification accuracy (%) on the Vesijako 2014 dataset. The supervised baseline [7] uses a similar 3D-CNN, trained end-to-end on three species with larch excluded. k-NN operates on the 72-d projector embedding space; LP is fitted on the backbone output features.

	OA (4)	OA (3)	Pine	Spruce	Birch	Larch
Supervised	—	96.2	96.2	86.6	98.2	—
k-NN	93.5	93.3	98.4	76.4	90.5	100.0
rKNN	93.5	93.1	98.1	76.4	90.4	100.0
LP	91.5	91.2	92.5	84.8	93.0	100.0

The k-NN classifier achieved 93.5% OA on four species. When larch is excluded for a direct 3-species comparison, k-NN reached 93.3%. LP achieved 91.5% OA. The k-NN accuracy Per-class is 98.4% (pine), 76.4% (spruce), 90.5% (birch), and 100.0% (larch). The adaptive-radius variant (rKNN) produced nearly identical results (93.5% OA). LP fell short of the k-NN classifiers, but got a significantly better result in the hardest class, spruce.

3.2. Embedding Space Analysis

To understand what the encoder has learnt, we project the frozen embedding space onto three dimensions using PCA and UMAP [9], coloured by species (Figure 3, 4). In both projections, birch is the only species that forms a distinct, dedicated cluster. The UMAP projection reveals several other clusters, but they contain a mixture of two or more species. The PCA projection shows less clearly defined cluster boundaries, as expected from a linear projection, but overall separation is still visible between the four classes.

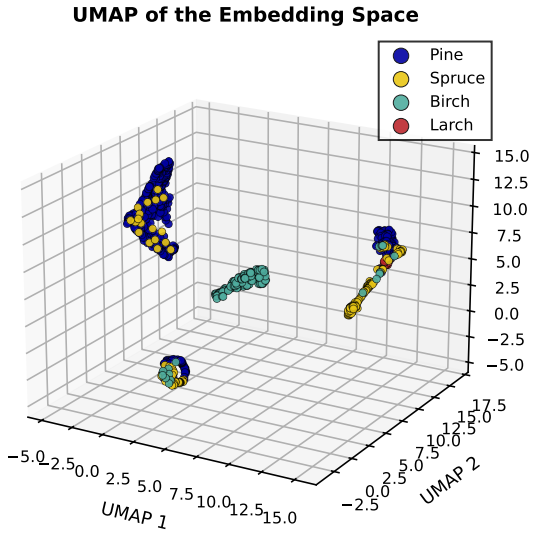


Fig. 3: UMAP projection of the frozen embedding space (validation set only), coloured by species. Birch is the only one of the four species that has a dedicated cluster for itself. The other clusters in the projection contain a mixture of two or more species.

4. DISCUSSION AND CONCLUSIONS

In this study, we investigated whether the minimal LeJEPA formulation with a single 3D-CNN encoder, SIGReg, and an invariance term without any EMA, predictor, or stop-gradient, can learn discriminative species representations from HS and canopy height data. Despite the simplicity, frozen projector embeddings achieved 93.5% OA via k-NN, 93.5% via adaptive-radius k-NN, and 91.5% via LP on four species. For a three-species comparison excluding larch (the baseline [7] uses a closely related 3D-CNN across the same field sites and excluded larch due to its scarcity), the k-NN classifier achieves 93.3% OA. Despite being entirely unsupervised during pre-training, the predictive SSL approach retains a substantial fraction of the supervised baseline's discriminative power with minimal architectural modifications to the downstream classifier.

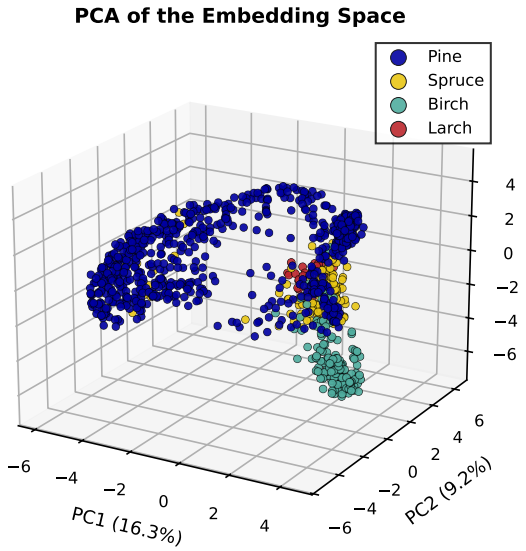


Fig. 4: PCA projection of the frozen embedding space (validation set only), coloured by species. As expected from a linear projection, cluster boundaries are less clear than in the UMAP, but species separation is still visible.

The practical implication is that predictive SSL can be added as a pre-training step to any existing supervised 3D-CNN classifier with minimal overhead: only multi-view augmentation and the LeJEPA loss are required. No decoder, no contrastive memory bank, and no second network is needed. This minimalism is not just an implementation convenience but the property that makes LeJEPA an ideal candidate for sensor-agnostic foundation models, where the pre-training architecture must remain simple enough to accommodate arbitrarily varying input channel counts and spectral resolutions.

Several directions remain open for building on this work. First, a direct comparison with generative SSL (e.g. 3D-MAE) on the same backbone and data would quantify the parameter-efficiency advantage of JEPA style predictive pre-training. Second, the LeJEPA formulation's architectural minimalism with a single encoder without EMA, a predictor or a decoder, makes it amenable to multi-domain, sensor-agnostic scaling. CARL [6] recently demonstrated that JEPA-based methods outperform generative models for cross-sensor spectral pre-training across medical imaging, satellite remote sensing, and autonomous driving. Extending this principle to forestry would require combining the LeJEPA pre-training objective with a sensor-agnostic architecture capable of handling heterogeneous spectral inputs from diverse domains. Third, the fixed-size bounding box (64×64 px) is larger than some tree crowns, so a single patch may capture multiple crowns of different species. A spatially-adaptive version with crown-size-dependent bounding boxes or a segmentation-aware pooling step would avoid this mixed-signal issue. Lastly, a supervised fine-tuning stage after SSL pre-training could close the remaining gap to the fully supervised baseline.

To conclude, we have presented the first LeJEPA-based approach for HS tree species classification. The results demonstrate that predictive embedding-space learning is a viable, simpler alternative to generative pre-training for spectral remote sensing, and that the LeJEPA formulation can approach supervised accuracy without any labels during pre-training. By establishing this proof of concept in a challenging domain-specific application, this work contributes to the broader effort toward multi-domain, sensor-agnostic JEPA foundation models.

5. REFERENCES

[1] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al., "Spectralgpt: Spectral remote sensing foundation model," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5227–5244, 2024.

[2] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot, "S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24088–24097.

[3] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, et al., "Hypersigma: Hyperspectral intelligence comprehension foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," *arXiv preprint arXiv:2301.08243*, 2023.

[5] Randall Balestriero and Yann LeCun, "Lejepa: Provable and scalable self-supervised learning without the heuristics," *arXiv preprint arXiv:2511.08544*, 2025.

[6] Alexander Baumann, Leonardo Ayala, Silvia Seidlitz, Jan Sellner, Alexander Studier-Fischer, Berkin Özdemir, Lena Maier-Hein, and Slobodan Ilic, "Carl: Camera-agnostic representation learning for spectral image analysis," 2026.

[7] Ilkka Pölönen, Leevi Annala, Samuli Rahkonen, Olli Nevalainen, Eija Honkavaara, Sakari Tuominen, Niko Viljanen, and Teemu Hakala, "Tree species identification using 3d spectral data and 3d convolutional neural network," in *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2018, pp. 1–5.

[8] Olli Nevalainen, Eija Honkavaara, Sakari Tuominen, Niko Viljanen, Teemu Hakala, Xiaowei Yu, Juha Hyyppä, Heikki Saari, Ilkka Pölönen, Nilton N. Imai, and Antonio M. G. Tommaselli, "Individual tree detection and classification with uav-based photogrammetric point clouds and hyperspectral imaging," *Remote Sensing*, vol. 9, no. 3, 2017.

[9] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.