applied sciences

MDPI

Article

Unsupervised Hierarchical Visual Taxonomy of Marble Natural Stone Using Cluster-Aware Self-Supervised Vision Transformers

Margarida Figueiredo [1,*], Carlos M. A. Diogo [1], Gustavo Paneiro [2], Pedro Amaral [3] and António Alves de Campos [1]

[1] CERENA, Centro de Recursos Naturais e Ambiente, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; carlosmadiogo@tecnico.ulisboa.pt (C.M.A.D.); antonio.campos@tecnico.ulisboa.pt (A.A.d.C.)
[2] DER/CERENA, Centro de Recursos Naturais e Ambiente, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; gustavo.paneiro@tecnico.ulisboa.pt
[3] LAETA, IDMEC, Associated Laboratory of Energy, Transports and Aerospace, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal; pedro.amaral@tecnico.ulisboa.pt
* Correspondence: margarida.figueiredo@tecnico.ulisboa.pt

Featured Application

This pipeline enables objective, appearance-based organization of marble inventories, supporting automated stock matching across suppliers, visual quality stratification within commercial varieties, and identification of cost-effective visual alternatives when preferred stones become unavailable.

Abstract

The marble industry relies on proprietary commercial names rather than objective visual categories, creating market inefficiencies for stakeholders who select stones based on appearance. Supervised classification perpetuates this problem by replicating inconsistent commercial labels instead of discovering intrinsic visual structure. We propose an unsupervised pipeline combining a two-stage training strategy: A pure self-supervised pretraining followed by cluster-aware fine-tuning of a DINO Vision Transformer, with empirically selected dimensionality reduction and agglomerative hierarchical clustering. Systematic ablation studies on 1480 marble images spanning 10 commercial varieties validate each design choice: cluster-aware training at k = 10 yields geometrically improved embeddings over the self-supervised baseline (mean Silhouette Score 0.693 ± 0.053 vs. 0.660 ± 0.030; mean Davies–Bouldin Index 0.386 ± 0.075 vs. 0.569 ± 0.012; $N = 9$ independent evaluations across 3 data partitions × 3 training initializations). The resulting taxonomy reveals three phenomena invisible to commercial classification: cross-category merging of visually indistinguishable stones carrying different market names, intra-category splitting of heterogeneous sub-populations within single varieties, and coherent grouping where commercial and visual boundaries coincide, with all three confirmed in every independent run. We further demonstrate that standard extrinsic metrics are misaligned with unsupervised taxonomy objectives when reference labels encode the inconsistencies the method aims to resolve. Validating this methodology across diverse stone types, larger datasets, and varied acquisition conditions represents a natural and necessary next step toward establishing its cross-domain generalizability.

Academic Editor: Pedro Couto

Received: 16 March 2026
Revised: 17 April 2026
Accepted: 18 April 2026
Published: 23 April 2026

Keywords: self-supervised learning; Vision Transformer; DINO; deep clustering; hierarchical clustering; marble classification; unsupervised visual taxonomy

1. Introduction

The natural stone industry is characterized by a persistent reliance on proprietary commercial names rather than systematic visual classification [1,2]. This market-driven naming convention creates a fragmented landscape in which the same visual appearance may carry different commercial designations depending on the supplier, quarry location, or regional tradition, while a single commercial name may encompass significant visual diversity [1,3]. For buyers, designers, and specifiers who select marble primarily based on aesthetic appeal [4], this nomenclature chaos introduces substantial inefficiencies: stones cannot be reliably compared across suppliers, visual alternatives are difficult to identify when a preferred variety becomes unavailable, and the absence of objective grouping criteria obscures the relationship between a stone's appearance and its underlying geological properties [5]. Consequently, the industry lacks a standardized visual taxonomy that would enable transparent, systematic organization of marble varieties based on their intrinsic appearance.

Various classification approaches have been explored to address this challenge [6], but they remain fundamentally misaligned with the objective of discovering natural visual structure. Early methods focused on quality grading and defect detection [7], employing handcrafted feature extraction techniques such as GLCM [8] and Local Binary Patterns [9] to distinguish among quality categories or to identify surface imperfections. While effective for industrial quality control, these systems are not designed to capture the aesthetic variability that defines commercial marble varieties. More recent approaches have employed supervised deep learning to classify stones into predefined commercial categories [10–12], achieving high accuracies on their respective datasets. However, these supervised methods perpetuate the very problem they aim to solve. By learning to replicate commercial labels as ground truth, they encode the inconsistencies and market-driven conventions of existing nomenclature rather than discovering objective visual groupings.

A supervised approach is fundamentally unsuited to our objective. Commercial labels are assigned by traders and quarry operators based on market conventions, regional traditions, and supplier preferences, not on systematic visual criteria. Two visually identical stones may carry different commercial names from different suppliers, while a single commercial name may encompass stones with markedly different veining patterns, base colors, or textural characteristics. Our goal is precisely to transcend these inconsistent labels and identify objective visual groupings based solely on appearance, organized hierarchically to reflect natural relationships, ranging from broad visual families to fine-grained textural distinctions. This requires an unsupervised methodology capable of learning discriminative visual representations without relying on external annotations.

In this paper, we propose an unsupervised pipeline for hierarchical visual taxonomy of marble varieties that combines cluster-aware self-supervised learning with nonlinear dimensionality reduction and agglomerative hierarchical clustering. The methodology is motivated by three converging developments in computer vision and unsupervised learning: the emergence of Vision Transformers (ViTs) as powerful feature extractors for structured visual patterns, the maturation of self-supervised learning as a paradigm for representation learning from unlabeled data, and the integration of clustering objectives directly into representation learning to produce feature spaces optimized for grouping tasks. By synthesizing these approaches, we demonstrate that meaningful visual

https://doi.org/10.3390/app16094137

taxonomies can be automatically discovered and hierarchically organized without any reliance on commercial labels, providing both a practical tool for the marble industry and a methodological template for similar challenges in other material science domains where existing labels are inconsistent or unreliable.

The pipeline integrates three components: (1) cluster-aware self-supervised training using a hybrid Cluster-Aware Distillation with No Labels (CA-DINO) strategy, (2) non-linear dimensionality reduction via UMAP to preserve the local manifold structure of learned embeddings, and (3) agglomerative hierarchical clustering using Ward's linkage to reveal nested relationships in the visual feature space. Each design choice is motivated by specific challenges of marble variety identification and validated through systematic ablation studies.

The main contributions are threefold:

An unsupervised pipeline evaluated through systematic ablation and cross-split replication. We propose and evaluate a complete pipeline for hierarchical grouping of marble varieties based solely on visual features. Unlike supervised approaches that replicate commercial labels, our method discovers objective visual structure, producing taxonomies grouping stones by perceptually salient characteristics such as vein density, base coloration, and textural organization. Comprehensive ablation studies demonstrate the contribution of each component and identify optimal hyperparameter configurations.

Evidence that visual similarity transcends commercial categories. Through quantitative metrics and qualitative dendrogram analysis, we demonstrate that the learned visual hierarchy does not replicate commercial classifications. The model correctly groups visually similar stones under different commercial names while separating visually distinct subpopulations within a single commercial category. This validates the pipeline's ability to discover intrinsic visual structure independent of market-driven naming conventions.

Methodological insights for unsupervised visual taxonomy evaluation. We demonstrate that standard clustering validation metrics, particularly extrinsic measures such as Adjusted Rand Index and Normalized Mutual Information, are misaligned with the task of discovering new visual structure in this class of problem, where the ground truth labels themselves represent the inconsistency the method aims to resolve. Through concrete examples, we show that these metrics penalize correct visual groupings diverging from commercial labels, establishing that rigorous qualitative evaluation of hierarchical coherence is essential for this class of problems. We note that while the individual components employed in this pipeline, self-supervised Vision Transformers, UMAP dimensionality reduction, and agglomerative hierarchical clustering, are established techniques, the contribution lies in their validated integration for a specific applied objective that none of them addresses individually: discovering hierarchical visual structure in a domain where existing labels are unreliable and therefore unsuitable as learning targets. Analogous integration contributions have demonstrated substantial impact in other fields; for instance, retrieval-augmented generation in natural language processing combines retrieval and generation components that are individually well-established, yet their integration unlocks capabilities unavailable to either component in isolation. Critically, the multi-split replication protocol presented in Section 4.7, which evaluates CA-DINO k = 10 across nine independent measurements (3 data partitions × 3 training initializations), provides empirical evidence that this specific combination produces genuine and stable visual structure rather than configuration-dependent artifacts.

Beyond immediate application in the marble industry, this work provides a methodological foundation for similar challenges in materials science, manufacturing,

and other domains in which existing categorical labels are driven by tradition, market convention, or regional variation rather than by systematic feature-based criteria.

The remainder of this paper is organized as follows. Section 2 reviews related work in marble classification, self-supervised learning, deep clustering, and hierarchical visual organization. Section 3 describes the dataset and presents the proposed pipeline, detailing the training procedure, feature extraction, and hierarchical clustering approach. Section 4 presents the results of systematic ablation studies that validate each pipeline component, followed by qualitative and quantitative analyses of the final visual taxonomy. Section 5 discusses broader implications, including the role of evaluation metrics, learned feature priorities, and practical applications. Section 6 concludes with a summary of contributions and directions for future work.

2. Related Work

2.1. Automated Marble Classification

Automated marble classification has been investigated for over three decades, initially focusing on quality control through defect detection and slab grading. The earliest methods employed classical computer vision techniques: RGB color-space analysis distinguished between broad categories, while morphological operations enabled surface-defect detection [7]. Subsequent work applied Gabor filters and wavelets for multi-scale textural feature extraction [13]. Among handcrafted descriptors, Gray-Level Co-occurrence Matrices (GLCM) and Local Binary Patterns (LBP) proved particularly effective, with LBPs demonstrating continued utility in modern hybrid systems [8,9].

The inherent subjectivity and time-intensive nature of manual marble classification, combined with proprietary data policies in industry settings, motivated a shift toward unsupervised methods [6]. Extracted features were typically fed to clustering algorithms such as k-means [14,15], followed by classification using Support Vector Machines or neural networks [9,16], often with PCA-based dimensionality reduction [13].

The adoption of Convolutional Neural Networks (CNNs) marked a significant advance in classification accuracy [10,12]. However, CNNs face inherent limitations in identifying marble varieties. Their localized receptive fields prioritize local features that may not capture the global continuity essential for distinguishing visually similar varieties. Differentiation among marble types often relies on subtle relationships among veining density, color gradients, and textural organization, features that require both fine-grained local analysis and an understanding of broader spatial context. More fundamentally, supervised methods share a conceptual flaw: if models are trained on commercial labels as ground truth, they will learn to replicate existing market-driven nomenclature rather than to discover intrinsic visual structure.

Recent work has demonstrated that Self-Supervised Learning (SSL) overcomes this limitation in industrial and geological domains. Brondolo and Beaussant [17] applied a second version of the Distillation with No Labels (DINO) methodology to CT-scanned rock-sample analysis, achieving state-of-the-art performance without annotated data. Scabini et al. [18] benchmarked Vision Transformers as texture feature extractors on material databases, demonstrating that DINO-pretrained models achieve superior material-classification performance. Zhu et al. [19] applied DINOv2 to concrete crack detection, outperforming supervised models where precise annotation is impractical. These successes establish SSL as a viable paradigm for domains where labels are scarce, inconsistent, or unreliable.

2.2. Self-Supervised Learning and Vision Transformers

In domains where large-scale expert annotation is infeasible or where labels themselves are inconsistent, Self-Supervised Learning has emerged as a powerful alternative paradigm. SSL methods learn rich visual representations directly from unlabeled data by solving pretext tasks that reveal inherent structure [20], circumventing the reliance on human-provided labels that constrain supervised approaches [21]. The field has evolved from clustering-based methods such as DeepCluster [22], through contrastive frameworks including SimCLR [23] and Momentum Contrast (MoCo) [24], to distillation-based approaches that avoid explicit negative pairs.

Among these methods, DINO [25] has demonstrated particularly strong performance for fine-grained visual tasks. DINO employs a student-teacher distillation framework in which a student Vision Transformer is trained to match the output distribution of a momentum-updated teacher network across multiple augmented views of the same image. Unlike contrastive methods requiring large batches of negative examples, DINO's knowledge distillation produces semantically meaningful features with strong localization properties, as evidenced by the emergence of object-centric attention maps without a supervised signal. This capability is especially relevant for marble classification, where models must simultaneously capture fine local details like vein direction, density, branching patterns, and global textural organization, including color gradients, spatial continuity, and structural layout. DINO's architectural foundation, the Vision Transformer [26], models global relationships among image patches via self-attention mechanisms [27], making it well-suited to capturing the structural continuity characteristic of marble patterns.

The robustness of DINO-learned representations to variations in scale, rotation, and lighting further enhances applicability to industrial image datasets, which often exhibit inconsistent acquisition conditions, scanner artifacts, and variable illumination. These properties position DINO as an ideal foundation for learning visual features without relying on inconsistent commercial labels.

2.3. Deep Clustering

While SSL provides a mechanism for learning discriminative features without labels, our objective requires organizing them into coherent, interpretable clusters. Traditional clustering algorithms perform effectively in low-dimensional spaces but degrade significantly when applied to high-dimensional data, where the curse of dimensionality renders standard distance metrics uninformative [28]. This necessitates learning a compressed representation before clustering [29].

However, a sequential pipeline in which feature learning and clustering are performed independently suffers from a representational disconnect [30]. The objective guiding feature learning, maximizing agreement between augmented views in SSL, is divorced from the downstream goal of forming well-separated clusters. This misalignment often produces feature spaces that are suboptimal for clustering, as learned representations prioritize view invariance over cluster separability.

Deep Clustering addresses this limitation by jointly optimizing feature representations and cluster assignments [30,31]. By integrating a clustering-specific loss directly into deep network training, DC creates a feedback loop: evolving cluster assignments refine the feature space, and improved features yield more accurate clusters. Deep Embedded Clustering (DEC) employs Kullback–Leibler divergence to iteratively sharpen assignments, with gradients backpropagated to produce inherently structured features [29]. Empirical evidence demonstrates this integrated approach substantially outperforms two-stage pipelines [29,30].

Contemporary approaches leverage SSL as the backbone for feature learning. Caron et al. [32] integrate online clustering directly into SSL via learnable prototypes and

optimal-transport assignment between augmented views, providing the foundation for cluster-aware DINO variants. Contrastive Clustering performs instance-level and cluster-level contrastive learning simultaneously, enforcing both feature discrimination and cluster structure [33]. Prototypical Contrastive Learning alternates between discovering cluster prototypes via k-means and optimizing representations to align with assigned prototypes [34]. These methods demonstrate that combining SSL's representational power with explicit clustering objectives produces feature spaces optimized for discovering visual categories without supervision, motivating our use of cluster-aware DINO [35].

2.4. Hierarchical Clustering of Visual Features

Beyond producing discriminative features and encouraging cluster structure, our objective requires organizing the discovered visual groups into an interpretable hierarchy that reveals nested relationships, ranging from broad families to fine-grained distinctions. While flat clustering methods partition data into disjoint categories, hierarchical clustering produces dendrograms that reflect multi-level organization, which is essential for a visual taxonomy that serves diverse stakeholders with varying granularity requirements.

Recent work establishes that agglomerative hierarchical clustering applied to deep visual embeddings produces semantically meaningful taxonomies. Naumov et al. [36] demonstrated that agglomerative methods applied to ResNet features on ImageNet datasets (up to 4.5 million images) yield dendrograms with coherent semantic structure. Yang et al. [37] established that agglomerative clustering and deep feature learning are mutually reinforcing, validating bottom-up hierarchical organization as a natural paradigm for discovering visual categories without supervision. Among linkage criteria, Ward's linkage criterion, with its variance-minimization objective, produces balanced, compact clusters at each hierarchical level, a property particularly desirable for taxonomies intended for practical industrial use, where interpretability and consistency across hierarchical levels are critical [38].

Table 1 consolidates the reviewed works along five axes: material, dataset size, learning paradigm, feature type, and whether a hierarchical output taxonomy is produced, to make explicit the position of the present work relative to the existing literature. The final column is of particular relevance: across all reviewed studies, only Selver et al. [6,13] produce a hierarchical output structure, and no prior work combines unsupervised learning with a ViT-based representation and explicit hierarchical taxonomy construction for natural stone.

Table 1. Summary of representative prior work in automated marble and material classification. Paradigm distinguishes rule-based algorithmic approaches, supervised methods (classical ML and deep transfer learning), self-supervised learning, and unsupervised clustering. Feature Type distinguishes handcrafted descriptors (e.g., GLCM, LBP, wavelet) from deep features automatically extracted by CNNs or Vision Transformers. Hier. indicates whether the study produces a hierarchical output taxonomy (Yes), a flat classification (No), or applies a hierarchical methodology to a flat target (Partial).

Authors (Year)	Material	Dataset Size	Paradigm	Feature Type	Metrics	Hierarchical
Elbehiery et al. (2007) [7]	Ceramic tiles	Not formally reported)	Rule-based/Algorithmic	Handcrafted (morphological, edge detection)	Qualitative only	No
Selver et al. (2009) [13]	Limestone slabs	1158 images	Supervised (Cascaded HRBFN)	Handcrafted (color, SDH, wavelet, morphological)	CCR, Sensitivity, Specificity	Partial

Selver et al. (2011) [6]	Limestone slabs	1158 images	Unsupervised (Hierarchical k-means)	Handcrafted (SDH, wavelet, morphological)	CCR	Yes
Haileslassie et al. (2019) [8]	Marble (3 grades)	180 images	Supervised (KNN, ANN)	Handcrafted (GLCM, color histogram)	Accuracy, Error rate	No
Turan et al. (2021) [9]	Marble (4 types)	200 images	Supervised (ELM, DT, ANN, SVM)	Handcrafted (LBP + histogram)	Accuracy	No
Ouzounis et al. (2021) [10]	Dolomitic marble tiles	489 images †	Supervised (Transfer Learning)	Deep features (DenseNet201 + Grad-CAM)	Accuracy, Precision, Recall, F1	No
Ouzounis et al. (2021) [12]	Dolomitic marble tiles	812 images	Supervised (Regression/Transfer Learning)	Deep features (CNNs)	MAPE	No
Brondolo & Beaussant (2025) [17]	Sandstone & Carbonate (CT)	~4200 image ‡	SSL (DINOv2) + Supervised + Unsupervised	Deep features (ViT); Handcrafted (BFE)	Accuracy, IoU	No
Scabini et al. (2025) [18]	General textures and materials	31,232 images (8 datasets)	Transfer Learning (SSL and supervised ViTs)	Deep features (ViT class token embeddings)	Accuracy, GFLOPs, Throughput	No
Zhu et al. (2026) [19]	Concrete (cracks)	106,057 images (4 datasets)	SSL (DINOv2) + Supervised	Deep features (ViT, CNN)	Precision, Recall, F1, Accuracy	No
Our work	Marble (10 varieties)	1480 images	Unsupervised (CA-DINO + UMAP + AHC)	Deep features (ViT-S/8, fine-tuned)	SS, DB, CH, ARI, NMI, V, CCC	Yes

† Dataset was class-balanced by subsampling from a larger acquisition of 986 images. ‡ Images are 2D slices sampled from 3D CT-scanned cores; 3000 slices for classification, up to 1200 for segmentation. The present work is the only entry combining an unsupervised paradigm, ViT-based deep features, and a hierarchical output taxonomy on a marble dataset.

The preceding review reveals three unresolved gaps that motivate the present work. First, existing supervised approaches to marble classification replicate commercially assigned labels without questioning their consistency or granularity, leaving the underlying visual structure of the stone unexplored. Second, self-supervised methods have demonstrated strong representational capacity for textured and geological materials, yet their application to the specific challenge of unsupervised taxonomy construction in natural stone remains unaddressed. Third, standard extrinsic clustering metrics assume ground truth labels are reliable, which is a questionable assumption in a domain where labels are assigned by market convention rather than systematic measurement.

These gaps motivate the following research questions:

(RQ1) Does cluster-aware self-supervised training produce embeddings with measurably improved internal clustering geometry, as assessed by internal compactness and separation metrics on a single industrial dataset, compared to standard self-supervised learning for marble texture organization?

(RQ2) Does the resulting unsupervised hierarchy reveal systematic visual structure that diverges from commercially assigned variety labels?

(RQ3) To what extent are standard extrinsic evaluation metrics appropriate for assessing unsupervised taxonomies when reference labels are themselves inconsistent?

RQ1 is addressed in Section 4.5, RQ2 in Section 4.6, and RQ3 in Section 5.2.

3. Materials and Methods

3.1. Dataset

The dataset comprises 1480 digital images of marble slab surfaces distributed across 10 commercial varieties: Estremoz Creme (59 images), Dante Grey (63), Branco Carrara (7), Arabescato Brown (81), Calacata (22), Bardiglio (260), Irish Black (10), Branco Peletigre (179), Exotic Ambar (327), and Ruivina (472). Figure 1 displays a representative sample from each variety.

Acquisition protocol. Images were acquired from marble slabs on active industrial production lines across multiple factory sites in Portugal. All sites use the same manufacturer and model of factory-calibrated line-scan camera system, operating under controlled and reproducible illumination conditions. Each physical slab measures between 0.5 and 2.5 m in its longest dimension and is scanned at a mean resolution of 7185 × 4166 pixels. For model input, each full-resolution scan was resized to 512 × 512 pixels, which constitutes the image resolution used throughout this study. The standardized acquisition setup, i.e., fixed camera geometry, consistent lighting, and uniform scanning speed, ensures that appearance variation in the dataset reflects genuine material heterogeneity rather than artifacts of the imaging process. Minor residual scanner-induced distortions were intentionally retained to assess the pipeline's tolerance to realistic imperfections. Because all images were collected under controlled industrial conditions using the same equipment family, the pipeline's robustness to acquisition variability, such as different scanner types, uncontrolled lighting, or variable viewing angles, remains an open question and warrants investigation in future work.

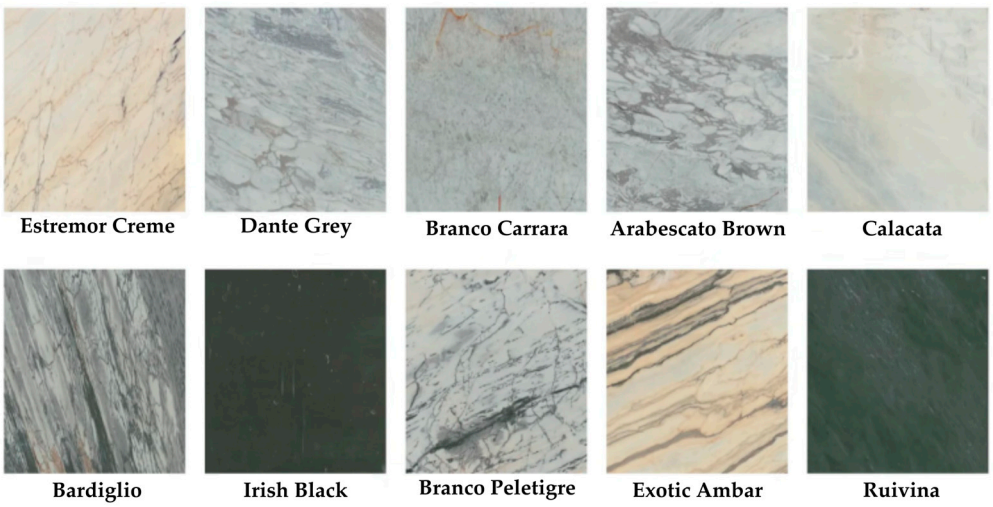

Figure 1. Representative sample image for each of the 10 marble varieties.

Dataset properties. The dataset is particularly well-suited for evaluating an unsupervised visual taxonomy because it exhibits variability along two complementary axes. First, it exhibits high intra-class variability: varieties such as Ruivina, Exotic Ambar, and Bardiglio show considerable diversity in veining density, base color, and textural organization within the same commercial category (Figure 2). Second, it exhibits low inter-class variability in certain pairs, with varieties such as Branco Peletigre and Dante Grey sharing a gray base tone and fine veining structure despite carrying distinct commercial designations. The dataset also presents significant class imbalance, ranging from 7 images for Branco Carrara to 472 for Ruivina. These properties together create a challenging setting in which an appearance-based taxonomy is expected to both diverge from and occasionally coincide with commercial boundaries.

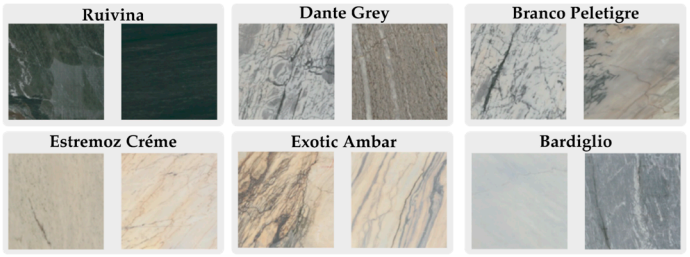

Figure 2. Examples of intra-class visual variability are present in the dataset. Images within the same commercial variety exhibit substantial differences in veining density, base color, and textural organization.

Commercial variety labels are used exclusively for post hoc evaluation and are never provided to the model during training. They serve as a reference for computing extrinsic validation metrics and for contextualizing the learned visual groupings; they do not constitute the learning target. No offline preprocessing or data augmentation was applied to the source images prior to the training pipeline, ensuring that all evaluation is performed on original acquisition data.

Data partitioning. The dataset is divided into three non-overlapping subsets following a nested 80/20 scheme. The full dataset is first split into a training-related pool (80%, 1175 images) and a held-out test set (20%, 305 images). The training-related pool is further divided into a training set (80%, 944 images), used exclusively for model weight updates, and a validation set (20%, 231 images), used exclusively for all post-training hyperparameter selection, epoch checkpoint, UMAP configuration, and linkage method. The test set is reserved for final metric reporting and is not accessed until the pipeline is fully frozen; no configuration decision is made based on test-set performance. To assess the reproducibility of results across different data partitions, this procedure is repeated three times with different random seeds, yielding three independent data splits, each with its own disjoint train, validation, and test populations. Table 2 reports the per-variety counts of images. Distributions are consistent across iterations to within sampling variability. All 10 commercial varieties are represented in every test set.

Table 2. Per-variety image counts for all data partitions. The training set is used for model weight updates; the validation set is used for all hyperparameter selection decisions; the test set is used exclusively for final evaluation. Varieties are ordered by total image count (descending).

Variety	Total	Train	Validation	Test
Ruivina	472	302	75	95
Exotic Ambar	327	209	52	66
Bardiglio	260	166	41	53
Branco Peletigre	179	114	28	37
Arabescato Brown	81	52	12	17
Dante Grey	63	40	9	14
Estremoz Creme	59	37	9	13
Calacata	22	14	3	5
Irish Black	10	6	1	3
Branco Carrara	7	4	1	2
Total	1480	944	231	305

3.2. Proposed Pipeline

We propose a three-component pipeline for unsupervised hierarchical visual taxonomy of marble varieties (Figure 3). The pipeline transforms raw slab images into an interpretable visual hierarchy through three sequential stages: (1) representation learning

via a hybrid self-supervised training strategy using CA-DINO, which produces a discriminative, clustering-optimized feature space from unlabeled images; (2) non-linear dimensionality reduction via UMAP, which compresses the high-dimensional embeddings while preserving local manifold structure; and (3) agglomerative hierarchical clustering, which organizes the reduced embeddings into a nested visual hierarchy ranging from broad visual families to fine-grained textural distinctions.

A distinctive feature of this pipeline is that the specific configuration of each post-processing component, e.g., the UMAP hyperparameters (target dimensionality, neighborhood size, minimum distance) and the agglomerative linkage criterion, is not assumed a priori but selected empirically through a joint optimization on a held-out validation set, using internal clustering metrics that require no reference to commercial labels. This design ensures that all configuration decisions are made independently of the test data on which the pipeline is ultimately evaluated and that the selection process is consistent with the work's unsupervised premise.

The pipeline is evaluated through a structured series of analyses organized in two tiers. Tier 1 comprises a stepwise ablation on the primary data partition, proceeding in four stages: (i) epoch selection, in which the training checkpoint is determined by representational convergence criteria, specifically, the stabilization of self-attention maps across consecutive epoch pairs, measured by structural similarity (SSIM) and cosine similarity, independently of any downstream pipeline configuration; (ii) joint post-processing optimization, in which UMAP dimensionality, neighborhood parameters, and linkage criterion are jointly selected on the validation set via internal clustering quality metrics; (iii) model comparison, in which all trained model variants (Pure DINO and CA-DINO at $k \in \{5, 8, 10, 12, 15\}$) are evaluated on the held-out test set using the frozen pipeline configuration; and (iv) qualitative taxonomy analysis, in which the emergent hierarchical structure is examined through dendrogram decomposition and per-variety clustering quality. Tier 2 provides cross-split replication by applying the fully frozen pipeline, with no configuration changes, to CA-DINO K10 embeddings from all three independent data partitions, each trained with three independent weight initializations, yielding nine independent measurements in total (3 data partitions × 3 training initializations per partition). Tier 2 evaluates three complementary aspects of reproducibility: (i) cross-partition metric consistency, assessing whether internal, external, and hierarchical metrics remain stable across data partitions and training runs; (ii) within-partition cluster label stability, quantifying agreement between flat $k = 10$ partition assignments across the three runs within each data partition via pairwise Adjusted Rand Index; and (iii) taxonomy phenomena persistence, testing whether the three key structural findings identified in the qualitative taxonomy analysis, namely the cross-category merging, intra-category splitting, and coherent pure family formation, are recovered in every independent run.

All quantitative evaluations follow a strict data-split protocol: the training set (944 images) is used exclusively for model training, the validation set (231 images) for all hyperparameter and configuration decisions, and the test set (305 images) for final metric reporting. This protocol is applied independently within each of the three data partitions.

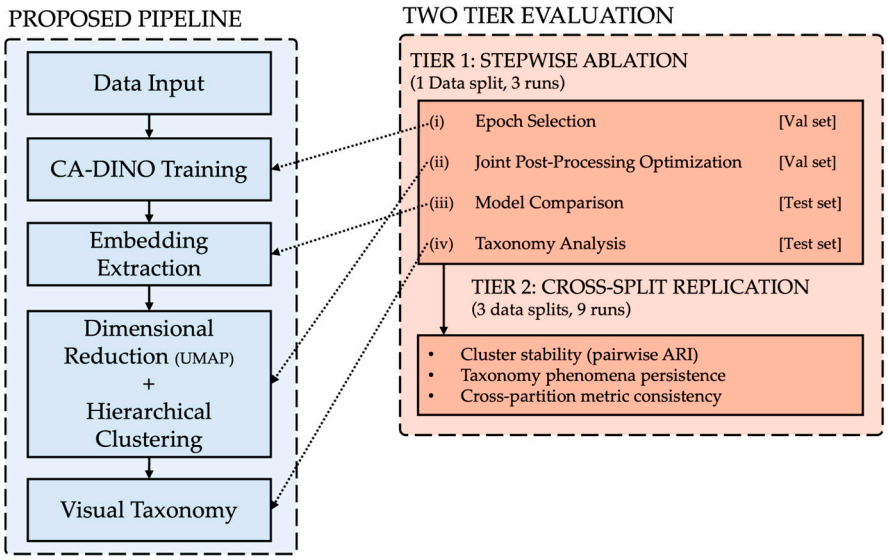

Figure 3. Overview of the proposed pipeline (**left**) and the two-tier evaluation studies (**right**).

3.3. Model Architecture and Training

3.3.1. Architecture

The model is built upon a Vision Transformer (ViT) backbone, specifically the Small variant with a patch size of 8 × 8 pixels (ViT-S/8). We use a patch-size-8 visual transformer model trained using the DINO method [25], pre-trained on ImageNet, followed by a 3-layer Multilayer Perceptron (MLP) projection head with Gaussian Error Linear Unit (GELU) activations and a final weight-normalized linear layer that yields 2048-dimensional output embeddings. The backbone's weights were not frozen and were fine-tuned on the marble dataset throughout training.

3.3.2. Two-Stage Training Strategy

Training proceeds in two sequential stages designed to provide a stable representational foundation before the clustering objective is introduced. The transition point at epoch 90 follows the training schedule adopted in the original CA-DINO framework [35], which demonstrated that dedicating approximately 45% of the standard 200-epoch training schedule to pure self-supervised pretraining provides a sufficiently stable representational foundation for the subsequent cluster-aware phase. The convergence of the DINO loss to a low-variance regime by epoch 90 (Section 4.1) confirms that this schedule is appropriate for the present dataset. All training was performed using the AdamW optimizer [39] with a base learning rate of 5×10^{-4} scaled linearly by batch size following the formula $lr = lr_{base} \times \frac{2 \times batch\_size}{256}$, yielding an effective learning rate of approximately 7.8×10^{-6}, with weight decay of 0.04. A Cosine Annealing scheduler [40] was applied over 200 epochs for all model variants; for CA-DINO k = 10, training was extended to epoch 400 with a fixed minimum learning rate to enable post-convergence monitoring, and the teacher network weights were updated via Exponential Moving Average with momentum following a cosine schedule from 0.996 to 1.0.

Stage 1—Pure DINO (epochs 1–90). The model is pre-trained using the standard DINO objective [25]. A multi-crop augmentation strategy is used: the teacher receives two global views (80–100% scale), while the student receives eight views: the same two global views plus six additional local crops (20–80% scale), all resized to 224 × 224 and subject to random horizontal flips (50% probability) and rotations (up to 10°). Following the original DINO formulation [25], the loss is computed over all cross-view pairs in which the teacher and student process different views; same-view pairs are excluded. The teacher receives

$N_t = 2$ global views, and the student receives $N_s = 8$ views (2 global + 6 local crops), yielding a total of $(N_t \times N_s) - N_t = 14$ unique cross-view pairs per image. The Stage 1 loss is thus the cross-entropy between the student and teacher output distributions over these cross-view pairs:

$$\mathcal{L}_{\text{DINO}} = -\frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{N_s}\sum_{j=1}^{N_s}\text{softmax}\left(\frac{t_i - c}{\tau_t}\right)\cdot \log\text{softmax}\left(\frac{s_j}{\tau_s}\right) \quad (1)$$

where t_i and s_j are the teacher and student outputs over N_t and N_s views, with student temperature $\tau_s = 0.1$ and teacher temperature $\tau_t = 0.04$. Training collapse is prevented via a global center c updated by EMA with momentum $m_c = 0.9$:

$$c_{new} = c_{old} \times m_c + x_{batch} \times (1 - m_c) \quad (2)$$

Stage 2—CA-DINO (epochs 91–200; CA-DINO k = 10 additionally trained to epoch 400 for convergence analysis). At each epoch, student embeddings are clustered via k-means to generate pseudo-labels. The CA-DINO phase requires a target cluster count k as a hyperparameter; to assess the pipeline's sensitivity to this choice, models were trained with $k \in \{5,8,10,12,15\}$, spanning a range from under- to over-segmentation of the visual space. A Dynamic Loss Gate (DLG) classifies samples as reliable or unreliable by fitting a two-component Gaussian Mixture Model to the per-sample loss distribution. The gate threshold is set to the minimum of the two fitted component means, placing the boundary between the low-loss (reliable) and high-loss (unreliable) sample populations. For reliable samples, a standard cross-entropy loss is computed against the pseudo-labels:

$$\mathcal{L}_{ce} = \frac{1}{|\, I_{\text{reliable}} \,|}\sum_{i \in I_{\text{reliable}}}\text{CrossEntropy}(\text{logits}_i, p_i) \quad (3)$$

For unreliable samples whose prediction confidence exceeds the gate condition, with $\tau_2 = 0.5$:

$$\left(\mathcal{L}_{ce,i} > \text{gate}_{\text{threshold}}\right) \wedge \left(\max(\text{softmax}(\text{logits}_i)_c) > \tau_2\right) \quad (4)$$

a sharpened target distribution is generated using sharpening temperature $\epsilon_c = 0.1$:

$$q_i = \text{softmax}\left(\frac{\text{logits}_i}{\epsilon_c}\right) \quad (5)$$

The Label Correction loss is the Kullback–Leibler divergence between this sharpened distribution and the prediction from a new augmented view $p_{\text{aug},i}$:

$$\mathcal{L}_{LC} = \frac{1}{|\, I_{\text{unreliable}} \,|}\sum_{i \in I_{\text{unreliable}}} D_{KL}\left(q_i^{(\text{detached})} \,\|\, p_{\text{aug},i}\right) \quad (6)$$

$$p_{\text{aug},i} = \text{softmax}(\text{logits}_{\text{aug},i}) \quad (7)$$

The total loss for Stage 2 combines all three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{ce} + \mathcal{L}_{LC} \quad (8)$$

https://doi.org/10.3390/app16094137

3.4. Feature Extraction and Post-Processing Pipeline

3.4.1. Embedding Extraction

After training, the model is set to evaluation mode, disabling stochastic elements such as dropout. Images are passed through the ViT backbone and MLP projection head, yielding a 2048-dimensional embedding vector from the MLP projection head (DINOHead). These high-dimensional embeddings form the input to the subsequent dimensionality reduction and clustering stages.

3.4.2. Dimensionality Reduction

The 2048-dimensional embeddings are susceptible to the curse of dimensionality, which degrades the reliability of distance metrics used in clustering. We apply UMAP, a non-linear dimensionality reduction technique that preserves local neighborhood structure in the data manifold, to project the embeddings into a lower-dimensional space. UMAP was selected because the visual differences between marble varieties are encoded non-linearly by the Vision Transformer; a linear projection, such as PCA, could discard the manifold structure critical for distinguishing fine-grained textural patterns.

The optimal UMAP configuration is not assumed a priori but determined empirically as part of the joint post-processing optimization described in Section 3.2. The three UMAP hyperparameters, namely target dimensionality, neighborhood size, and minimum distance, are optimized jointly with the agglomerative linkage criterion (Section 3.4.3) in a single grid search over the hyperparameter space shown in Table 3, yielding 480 candidate configurations in total. Each configuration is evaluated end-to-end on the validation set (n = 231) using internal clustering metrics, namely Silhouette Score and Davies-Bouldin Index, that require no reference to commercial labels, consistent with the unsupervised premise of this work. The configuration achieving the highest mean Silhouette Score across the three training runs is selected and held fixed for all subsequent test-set evaluations. The ablation in Section 4.3 examines the effect of target dimensionality as the dominant factor, with all other hyperparameters fixed at the jointly selected values.

Table 3. UMAP hyperparameter search space. All combinations are evaluated jointly with the four linkage criteria (Section 3.4.3), yielding 480 candidate configurations in total. Selection is performed on the validation set (n = 231) using internal clustering metrics.

Parameter	Range
Output Dimensions	[5, 10, 15, 30, 50, 75]
Number of Neighbors	[5, 10, 15, 30, 50]
Minimum Distance	[0.0, 0.1, 0.25, 0.5]
Distance Metric	Euclidean
Linkage (co-optimized)	Ward, Complete, Average, Single

3.4.3. Hierarchical Clustering

Agglomerative hierarchical clustering is applied to the UMAP-reduced embeddings, producing a dendrogram that reveals nested visual relationships from broad visual families to fine-grained textural distinctions [38]. Prior to clustering, the UMAP-reduced embeddings are L2-normalized to ensure that distance computations are scale-invariant across output dimensions.

Four linkage criteria are evaluated as candidate clustering strategies (Table 4). No linkage method is assumed a priori; the optimal criterion is selected jointly with the UMAP configuration through the validation-set grid search described in Sections 3.2 and 3.4.2. All linkage methods are evaluated using Euclidean distance: this is a formal

requirement for Ward's variance-minimization criterion and is applied consistently across all four methods for comparability. The selected linkage method and its empirical justification are reported in Section 4.4.

Table 4. Linkage criteria evaluated for agglomerative hierarchical clustering.

Linkage Criteria	Description
Ward	Minimizes the increase in total within-cluster variance at each merge
Complete	Distance between the two furthest points across clusters
Average	Mean pairwise distance between all points across clusters
Single	Distance between the two closest points across clusters

3.5. Evaluation Framework

3.5.1. Qualitative Evaluation

Qualitative analysis of the generated dendrograms serves as an essential evaluation tool in this study, complementing the quantitative metrics described below. Two criteria guide the assessment. First, hierarchical coherence: an optimal dendrogram exhibits a general-to-specific organization, where the highest-level splits separate data into broad, visually distinct families and subsequent splits capture progressively finer distinctions. Second, structural integrity: the dendrogram must avoid common artifacts such as chaining (a staircase-like pattern in which points merge sequentially rather than in balanced groups) and early singleton formation (in which outliers detach prematurely from the main hierarchy). Figure 4 provides a conceptual illustration of an ideal dendrogram structure.

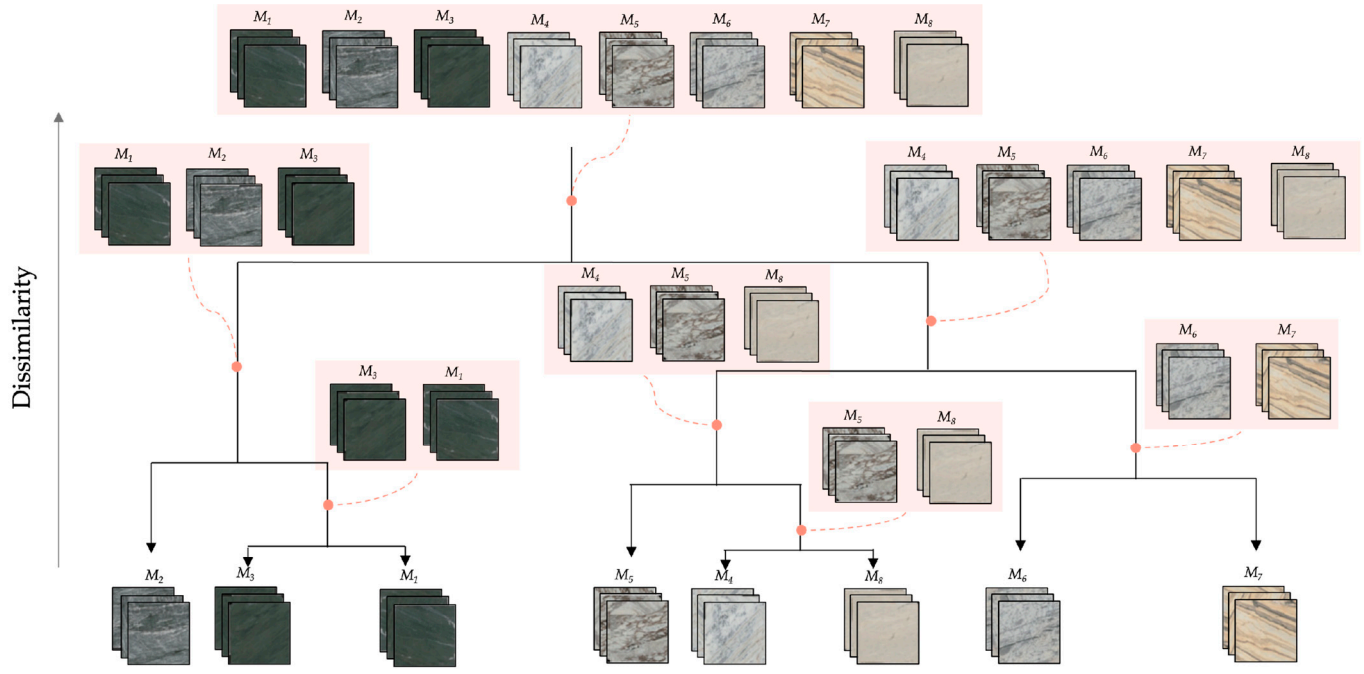

Figure 4. Conceptual illustration of an ideal dendrogram structure, showing clear hierarchical separation and the corresponding general-to-specific visual organization of clusters.

This qualitative analysis is indispensable because it can identify well-formed clusters containing visually similar marble images from different commercial varieties—a correct outcome for an appearance-based taxonomy that extrinsic metrics would incorrectly penalize.

3.5.2. Quantitative Evaluation

The clustering results are assessed using a suite of standard validation indices organized into three categories. Table 5 summarizes each metric's interpretation, range, and optimal direction.

Table 5. Summary of clustering validation indices used in this study. Internal metrics assess cluster quality from the data structure alone; external metrics compare cluster assignments against commercial variety labels (used for evaluation only, never for training); the hierarchical metric assesses dendrogram fidelity. Standard optimal directions are indicated (↑ higher is better; ↓ lower is better.), with the caveat that lower external metric scores may reflect desirable behavior in this study.

Metric	Abbreviation	Category	Range	Optimal	What It Measures
Silhouette Score	SS	Internal	[−1, +1]	↑ Higher	Mean ratio of intra-cluster cohesion to inter-cluster separation for each sample
Davies-Bouldin Index	DB	Internal		↓ Lower	Average similarity between each cluster and its most similar neighbor; penalizes loose, overlapping clusters
Calinski-Harabasz Index	CH	Internal		↑ Higher	Ratio of between-cluster dispersion to within-cluster dispersion; favors compact, well-separated clusters
Adjusted Rand Index	ARI	External	[−1, +1]	↑ Higher	Chance-corrected agreement between predicted cluster assignments and commercial variety labels
Normalized Mutual Information	NMI	External	[0, +1]	↑ Higher	Mutual dependence between predicted cluster assignments and commercial variety labels, normalized by entropy
V-measure	V	External	[0, +1]	↑ Higher	Harmonic mean of homogeneity (each cluster contains only one label) and completeness (all samples of a label are in one cluster)
Cophenetic Correlation Coefficient	CCC	Hierarchical	[−1, +1]	↑ Higher	Pearson correlation between pairwise distances in the original feature space and the distances implied by the dendrogram structure

Internal measures. Silhouette Score (SS) [41], Davies-Bouldin Index (DB) [42], and Calinski-Harabasz Index (CH) [43] assess cluster compactness and separation based solely on the data's inherent structure, without reference to external labels. These metrics serve as the primary selection criterion during validation-set optimization and as the primary quality indicator in all ablation comparisons.

External measures. Adjusted Rand Index (ARI) [44], Normalized Mutual Information (NMI) [45], and V-measure [46] compare cluster assignments to the commercial labels. These are reported as descriptive reference metrics characterizing the relationship between the discovered visual structure and the existing commercial taxonomy, not as measures of pipeline correctness (see Section 3.5.3).

Hierarchical measures**.** The Cophenetic Correlation Coefficient (CCC) [47] evaluates how faithfully the dendrogram structure represents the original pairwise distances in the feature space, providing a direct measure of hierarchical fidelity independent of any flat cluster cut.

3.5.3. Metric Interpretation and Evaluation Strategy

Quantitative validation indices serve two complementary roles in this study. Internal measures (SS, DB, CH) provide an objective assessment of cluster compactness and separation in the learned embedding space, independently of any labeling convention.

External measures (ARI, NMI, V-measure) quantify the degree of correspondence between the discovered groupings and the existing commercial classification, a relationship that is itself informative: high alignment would suggest the model recovers the commercial taxonomy; low alignment, combined with high internal scores and coherent dendrograms, suggests the model has identified a more perceptually consistent organization than the commercial labels impose.

One important caveat applies to the interpretation of external metrics. Because commercial labels were assigned by market convention rather than systematic visual criteria, penalization for deviating from them is not equivalent to penalization for clustering error. For example, grouping Branco Peletigre and Dante Grey together based on their shared gray base tone and fine veining pattern constitutes a visually correct decision that all extrinsic metrics would score as incorrect. External metrics are therefore reported as a descriptive reference to characterize the relationship between the learned visual structure and the existing commercial taxonomy, rather than as a primary measure of pipeline quality. Qualitative dendrogram analysis and internal indices serve as the primary evaluation criteria.

To assess the robustness of all reported metrics to training stochasticity and data partition variability, results are reported as mean ± standard deviation across three independent training runs. Cross-partition consistency across all three data splits is evaluated separately in Section 4.7.

3.6. Implementation Details

All models were implemented in PyTorch (v2.4.1+cu21) and trained on a single NVIDIA GeForce RTX 2080 Ti GPU (12 GB). The ViT-S/8 backbone was initialized from publicly available DINO ImageNet pre-trained weights. Training was performed with a batch size of 32 images, a gradient clipping norm of 3.0, and mixed-precision (fp16) forward passes with fp32 parameter updates.

The training experiment comprises 54 model instances in total (3 data partitions × 6 model variants × 3 runs per variant per partition). However, the training scope is asymmetric by design: CA-DINO k = 10, the configuration identified as optimal in Section 4.5, is trained across all three data partitions with three independent runs each, yielding nine measurements and enabling the cross-split replication analysis in Section 4.7. All remaining model variants (Pure DINO and CA-DINO at $k \in \{5, 8, 12, 15\}$) are trained with three independent runs on the primary data partition (Data split 1) only, as their role is to support the model comparison ablation in Section 4.5 rather than cross-partition reproducibility analysis. CA-DINO k = 10 was additionally trained to epoch 400 on all three partitions to support the convergence analysis in Section 4.2; all other variants use the standard 200-epoch schedule. Each run uses a different random seed for weight initialization and data shuffling.

Post-processing steps, namely UMAP projection, L2 normalization, agglomerative clustering, and metric computation, were implemented in Python (v3.8.20) using the umap-learn [48], scikit-learn [49], and scipy libraries [50]. All UMAP projections use random_state = 42 to ensure deterministic output given fixed input embeddings. The full post-processing pipeline is encapsulated in a single parameterized function applied identically across all analyses, ensuring that the configuration selected on the validation set is propagated without modification to all test-set evaluations.

4. Results

This section presents the experimental results of the proposed CA-DINO pipeline for unsupervised hierarchical visual taxonomy of marble varieties, organized according to the three-tier validation structure introduced in Section 3.2. All quantitative results are

reported as mean ± standard deviation across three independent training runs unless stated otherwise; test-set evaluations use n = 305 images per data partition with the pipeline configuration fully frozen prior to any test-set access.

Tier 1 comprises a stepwise ablation on the primary data partition. Section 4.1 reports the training dynamics of CA-DINO k = 10. Section 4.2 presents the multi-criterion convergence analysis based on structural and cosine similarities of consecutive-epoch attention maps that justifies the selection of the epoch 200 checkpoint and addresses the relationship between post-convergence attention behavior and representational stability. Sections 4.3 and 4.4 present the ablation studies on dimensionality reduction and linkage method, respectively, both evaluated on the held-out test set using the configuration jointly selected on the validation set. Section 4.5 reports the main quantitative clustering results across all six model variants (Pure DINO and CA-DINO at k ∈ {5, 8, 10, 12, 15}), including paired statistical significance tests comparing CA-DINO k = 10 against the Pure DINO baseline. Section 4.6 presents the qualitative analysis of the emergent hierarchical structure at k = 10, including dendrogram decomposition, composition heatmap, and per-variety clustering quality.

Tier 2 results are reported in Section 4.7, which provides cross-split replication by applying the frozen pipeline to CA-DINO k = 10 embeddings from all three independent data partitions, each with three training runs, yielding nine independent measurements. This tier reports cross-partition metric consistency, within-split cluster stability via pairwise Adjusted Rand Index, and the persistence of key taxonomy phenomena across all nine runs.

4.1. Training Dynamics

The training of CA-DINO proceeds in two stages: a pure self-supervised DINO pretraining phase (epochs 1–90) and a cluster-aware fine-tuning phase (epochs 91–200). The evolution of both loss components and the Dynamic Loss Gate threshold across training is shown in Figure 5 for a representative run of CA-DINO k = 10 from the primary data partition. Training dynamics were qualitatively consistent across all three independent training runs of the primary data partition; multi-run aggregated loss curves were not available for this analysis.

During Stage 1, the DINO loss decreased from an initial value of 7.623 at epoch 1 to 1.743 at epoch 90, representing a total reduction of 77.1%. The descent was non-monotonic, exhibiting 38 upward perturbations consistent with the stochastic gradient dynamics of self-supervised ViT training with multi-crop augmentation, but converged to a stable low-variance regime by the end of Stage 1, with a windowed mean of 1.744 ± 0.055 at epoch 90. At the onset of Stage 2 (epoch 91), the classification loss activated immediately, exhibiting high initial variance and peaking at 108.524 at epoch 95. These early perturbations are consistent with the known instability of pseudo-label assignment during the first cluster centroid updates, when the GMM-based Dynamic Loss Gate has not yet converged. Concurrently, the DINO loss exhibited two secondary spikes at epochs 96 (2.318) and 109–110 (2.794), reflecting the temporary disruption to the self-supervised objective caused by the emerging cluster-alignment pressure before the joint optimization stabilized. Importantly, the DINO loss recovered below its Stage 1 minimum from epoch 106 onward, reaching an overall minimum of 1.365 at epoch 192, confirming that the self-supervised objective continued to improve throughout Stage 2 despite the added cluster-alignment pressure. The classification loss stabilized to a low-variance regime by epoch 150, with windowed means of 5.771 ± 0.398 at epoch 150 and 6.222 ± 0.290 at epoch 200, indicating that pseudo-label quality and cluster assignments had converged well before the end of training.

The Dynamic Loss Gate threshold reached a maximum of 76.154 in the early epochs of Stage 2 before decaying to a final value of 1.560 by epoch 200, representing a 23.6% reduction from its first active value. This trajectory confirms that the proportion of samples classified as unreliable pseudo-labels decreased substantially over the course of Stage 2, as cluster centroids stabilized and the classifier became progressively more confident in its assignments.

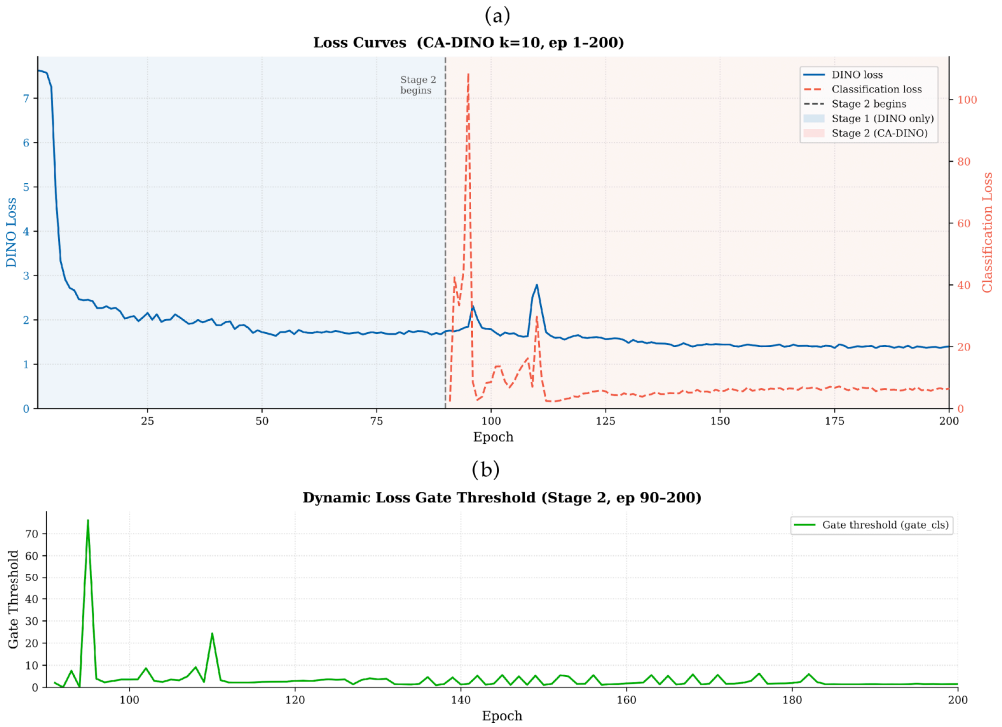

Figure 5. Training dynamics for CA-DINO k = 10, representative run from the primary data partition. (**a**): DINO loss (left axis, epochs 1–200) and classification loss (right axis, Stage 2 only, epochs 91–200), with the Stage 1/Stage 2 transition marked at epoch 90. (**b**): Dynamic Loss Gate threshold across Stage 2 (epochs 91–200). Training dynamics were qualitatively consistent across all three independent training runs.

4.2. Epoch Convergence Analysis

The selection of epoch 200 as the evaluation checkpoint rests on three independent and complementary convergence criteria, grounded in the analysis of self-attention maps across checkpoints at epochs 90, 100, 200, 300, and 400, conducted on the CA-DINO k = 10 training run extended to 400 epochs. Since the convergence behavior of self-attention maps reflects model-intrinsic representational properties shared across training runs and data partitions, the findings are taken as representative of the full experimental configuration. Aggregated SSIM and cosine similarity (CS) between consecutive epoch-pair attention maps were computed over the test set. The convergence curves are shown in Figure 6. Figure 7 shows the evolution of self-attention maps across all five checkpoints for three purposefully selected test images, each illustrating a distinct convergence behavior: one image whose attended regions stabilize rapidly after epoch 100 (SSIM = 0.591 at the 100→200 transition, above the global mean of 0.431), one exhibiting a local non-monotonic spatial disruption between epochs 200 and 300 (SSIM dropping from 0.680 at 100→200 to 0.543 at 200→300, below the global mean of 0.607), and one displaying an atypical pattern of early spatial stability followed by a transient decrease at the 100→200 transition (SSIM = 0.463) before consolidating.

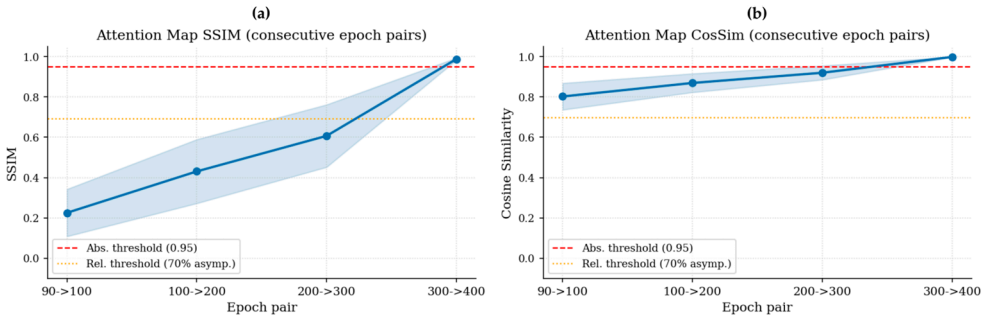

Figure 6. Convergence of attention map representations across training epochs, measured as the mean Structural Similarity Index (SSIM; (**a**)) and cosine similarity (**b**) between attention maps of consecutive epoch pairs, averaged over the set of test images. Shaded bands indicate one standard deviation, and dashed lines mark the convergence reference thresholds.

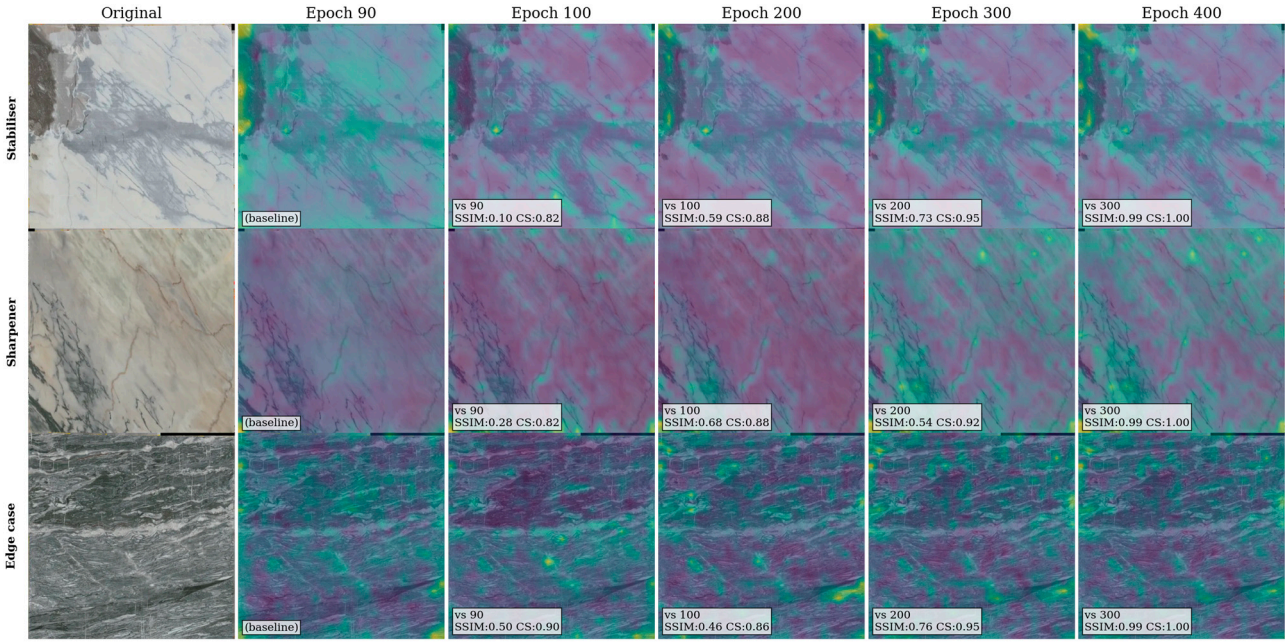

Figure 7. Self-attention map evolution for three representative test images illustrating contrasting convergence behaviors. Each row shows the original image followed by attention maps at epochs 90, 100, 200, 300, and 400, with per-cell SSIM and cosine similarity reported relative to the previous checkpoint. Top row (Stabilizer): SSIM rises sharply from 0.101 at the 90–100 transition to 0.591 at 100–200—well above the global mean of 0.431—indicating rapid spatial stabilization; the attended region geometry is effectively fixed by epoch 200. Middle row (Sharpener): SSIM peaks at 0.680 at 100–200 then dips to 0.543 at 200–300 (below the global mean of 0.607), revealing a localized spatial reorganization in the post-200 window before full consolidation at 300–400 (SSIM = 0.991). Bottom row (Edge case): unusually high SSIM at the 90–100 transition (0.498 vs. global mean 0.226) indicates early spatial convergence, followed by a transient decrease at 100–200 (SSIM = 0.463) consistent with a secondary perturbation driven by EMA teacher drift, before re-stabilization through the remaining training.

Primary criterion—rate-of-change inflection in spatial attention. The most operationally informative convergence signal is the per-epoch rate of change in SSIM, which directly quantifies how rapidly the spatial structure of self-attention maps reorganizes between consecutive checkpoints. At the 90–100 transition, mean SSIM was 0.2257 ± 0.1169 (range −0.012 to 0.508, n = 305), with a per-epoch rate of 0.02257, reflecting active and highly variable spatial reorganization during the early Stage 2 period. Between epochs 100 and 200, this rate dropped by 90.9% to 0.002052 per epoch, with mean SSIM

rising to 0.4309 ± 0.1582, marking the most abrupt inflection in the entire training trajectory and confirming that the primary phase of spatial reorganization had concluded by the 100–200 interval. Beyond this inflection, the per-epoch rate remained low through the 200–300 interval (0.001760 per epoch, a further 14.2% reduction), indicating a slow, sustained consolidation phase rather than continued rapid representational learning.

Secondary criterion—late-epoch spatial fixation. The 300–400 transition exhibited a qualitatively distinct behavior: mean SSIM reached 0.9876 ± 0.0050, with all 305 test images falling within the narrow range 0.972 – 0.994. This near-perfect inter-checkpoint similarity, with a standard deviation 29-fold lower than the average of the three preceding intervals (σ = 0.0050 vs. mean σ = 0.143), is a marker of spatial fixation rather than meaningful continued learning (the self-attention maps have become essentially frozen between epochs 300 and 400). The per-epoch SSIM rate at 300–400 (0.003807) is higher than at both 100–200 (0.002052) and 200–300 (0.001760); this apparent acceleration is driven by the large absolute SSIM gain over this interval and does not reflect faster spatial learning, but rather the terminal collapse of inter-checkpoint variability as the model enters a regime of fixated attention. Selecting a checkpoint within this fixated regime risks encoding training-set-specific spatial patterns that do not generalize to the broader test population. This risk is concretely illustrated by the sharpener image in Figure 7, which shows a local spatial reorganization at 200–300 (SSIM = 0.543, below the global mean of 0.607), confirming that even for images that appear broadly stable by epoch 200, late-stage training can introduce secondary spatial perturbations of non-negligible magnitude.

Tertiary criterion—directional convergence of embeddings. Since downstream clustering operates via cosine distance in the embedding space, cosine similarity between consecutive checkpoint attention maps provides a complementary directional convergence signal. CS increased monotonically from 0.8022 ± 0.0659 at 90–100 (80.4% of the asymptotic value of 0.9982 observed at epoch 400) to 0.8692 ± 0.0463 at 100–200 (87.1%), 0.9198 ± 0.0345 at 200–300 (92.1%), and 0.9982 ± 0.0006 at 300–400 (100.0%). The incremental CS gain decreased from 0.0670 at 100–200 to 0.0506 at 200–300, indicating a tapering rate of directional reorganization consistent with the SSIM rate-of-change inflection. While the embedding directions have not fully reached their asymptotic values by epoch 200 (87.1% of asymptote), the marginal directional gain available between epochs 200 and 300 (0.0506) must be weighed against the spatial fixation and secondary perturbation risk identified above; this gain is insufficient to override the convergence signal provided by the rate inflection criterion.

Contextual criterion—SSL generalization on constrained datasets**.** Consistent with established behavior in self-supervised ViT training, the optimal downstream evaluation checkpoint is not necessarily the terminal one. In the DINO framework, k-NN accuracy on downstream tasks does not increase monotonically with the number of epochs [25], and intermediate checkpoints have been shown to generalize better when the downstream dataset is small or domain-shifted relative to the pretraining distribution. In the present setting, a constrained natural stone dataset of 305 test samples, continuing beyond epoch 200, risks the model's self-attention heads progressively encoding training-set-specific spatial patterns that reduce embedding transferability to the full test population, as evidenced by the late-epoch spatial fixation described above and by the secondary perturbation visible in the sharpener image of Figure 7. Based on these three criteria jointly, epoch 200 was selected as the evaluation checkpoint: it sits at the post-inflection plateau where the per-epoch spatial reorganization rate has dropped by 90.9% from the early Stage 2 value, well before the spatial fixation regime that characterizes epochs 300–400, while retaining 87.1% of the asymptotic embedding directional similarity.

The possibility of representational overfitting after epoch 200 warrants explicit consideration. In supervised settings, overfitting manifests as degraded generalization on

held-out data; in self-supervised learning, the analogous risk is representational collapse or drift, i.e., the progressive loss of discriminative structure in the embedding space. The near-perfect SSIM at 300–400 (0.9876 ± 0.0050, cross-image range 0.972–0.994) is consistent with such drift: rather than continuing to refine discriminative spatial representations, the model has entered a regime in which attention patterns are replicated with near-perfect fidelity between successive checkpoints, and the residual per-epoch variation has effectively vanished. This interpretation is further supported by the DINO loss trajectory reported in Section 4.1, which reached its minimum near epoch 192 and remained consistently low thereafter, indicating that the self-supervised objective had converged and that continued training beyond epoch 200 provided no meaningful objective-level signal to drive further representational improvement.

4.3. Dimensionality Reduction Ablation

To determine the operating configuration for dimensionality reduction, a joint two-phase protocol was applied. In Phase 1, a UMAP hyperparameter grid search was conducted on the validation set (n = 231 per split) across 480 configurations spanning six target dimensionalities (5, 10, 15, 30, 50, 75), five neighborhood sizes (5, 10, 15, 30, 50), four minimum-distance values (0.0, 0.1, 0.25, 0.5), and four linkage methods (Ward, Complete, Average, Single), totaling 1440 evaluations across the three (split, run) pairs of the first data split. Configuration selection was based exclusively on the internal Silhouette Score; commercial variety labels were not used at any point during Phase 1 to prevent label leakage into the hyperparameter selection process. In Phase 2, seven preprocessing configurations were evaluated on the held-out test set (n = 305 per split) using the pipeline frozen in Phase 1: the raw 2048-dimensional embeddings and UMAP projections to 5, 10, 15, 30, 50, and 75 dimensions, all followed by L2 normalization and hierarchical clustering at k = 10 with the linkage method selected in Phase 1. All metrics are reported as mean ± standard deviation across nine independent measurements (3 data splits × 3 training runs). Results are presented in Table 6; two-dimensional UMAP projections for a representative run are shown in Figure 8, and the dimensionality metric curve is shown in Figure 9.

Phase 1—Joint configuration selection. All nine validation runs independently selected n_neighbors = 5 and min_dist ∈ {0.0, 0.1}, with 0.0 preferred in eight of nine runs, confirming a consistent preference for tight local neighborhood structure across all data partitions and training initializations. Linkage method selection was more distributed: Ward was selected in four of nine runs, Average in three, and Complete and Single in one each. Because no single (dimensionality, linkage) pair was selected in more than two runs, the modal configuration across all nine run-level optima was (50D, n_neighbors = 5, min_dist = 0.0, Average linkage), appearing twice. On the aggregated validation-set mean SS, the configuration (30D, n_neighbors = 5, min_dist = 0.0, Average) ranked first (SS = 0.731 ± 0.057) and the modal winner ranked second (SS = 0.723 ± 0.066), a margin of Δ = 0.008 that falls within the inter-run standard deviation of both estimates. UMAP 50D with n_neighbors = 5, min_dist = 0.0, and Average linkage was therefore adopted as the operating configuration and fixed for all Phase 2 evaluations and all subsequent downstream analyses.

Phase 2—Test-set results. Raw 2048-dimensional embeddings produced highly unstable results, with SS = 0.372 ± 0.238 and a Calinski-Harabasz index of 1170 ± 2199 (coefficient of variation, CV = 188%), confirming that high-dimensional geometry is an unreliable basis for clustering across independent runs. All six UMAP configurations substantially improved both geometric quality and inter-run stability relative to the raw baseline, with SS above 0.677 and DB below 0.399 across all tested dimensionalities.

Internal metrics followed a consistent pattern along the dimensionality axis (Figure 8). Silhouette Score rose from 5D (SS = 0.677 ± 0.039) to a peak at 30D (SS = 0.711 ± 0.039), declined slightly at 50D (SS = 0.693 ± 0.053), and partially recovered at 75D (SS = 0.706 ± 0.034). The Davies-Bouldin index reached its minimum at 10D (DB = 0.349 ± 0.051), with 30D achieving the lowest variance across all UMAP configurations (DB = 0.368 ± 0.041, CV = 11.1%). Label-alignment scores were broadly stable across the dimensionality range. NMI peaked at 30D (0.515 ± 0.108) and ARI also reached its maximum at 30D (0.367 ± 0.158), with both metrics declining marginally at higher dimensionalities. The cophenetic correlation coefficient exhibited a monotonically increasing trend with dimensionality, rising from 0.921 ± 0.032 at 5D to 0.953 ± 0.010 at 75D, indicating progressively better preservation of dendrogram geometry at higher embedding dimensions under Average linkage. The selected configuration (UMAP 50D) achieved SS = 0.693 ± 0.053 and DB = 0.386 ± 0.075 on the test set, values within one standard deviation of the 30D peak in both metrics. The overall dimensionality effect on internal geometry is modest: the SS range across all six UMAP configurations spans 0.034 points (0.677 to 0.711) and the ARI range spans 0.037 points (0.330 to 0.367), with all pairwise UMAP-to-UMAP differences falling within one standard deviation.

For practitioners applying this pipeline to similar fine-grained visual domains, the results suggest that UMAP target dimensionality in the range of 10–50 dimensions provides robust clustering geometry, with the specific optimum depending on the intrinsic complexity of the embedding manifold. The overall dimensionality effect on internal geometry is modest: the SS range across all six UMAP configurations spans only 0.034 points (Table 6), suggesting that the pipeline is not brittle to this hyperparameter within the tested range. We recommend selecting dimensionality via validation-set grid search using internal clustering metrics, as described in this study, rather than adopting a fixed value across domains.

Table 6. Dimensionality reduction ablation: clustering metrics at k = 10 (Average linkage, L2 normalization, CA-DINO K10, epoch 200). Values are mean ± standard deviation across 3 data splits × 3 training runs (N = 9). UMAP hyperparameters selected on the validation set (n = 231 per split); all metrics computed on the held-out test set (n = 305 per split). Bold indicates the best value per metric among UMAP configurations. ↑ Higher is better; ↓ Lower is better.

Configuration	SS ↑	DB ↓	CH ↑	ARI ↑	NMI ↑	V-Measure ↑	CCC ↑
Raw (2048D)	0.372 ± 0.238	1.157 ± 0.632	**1170 ± 2199**	0.302 ± 0.212	0.421 ± 0.179	0.421 ± 0.179	0.886 ± 0.109
UMAP 5D	0.677 ± 0.039	0.399 ± 0.057	810 ± 229	0.335 ± 0.157	0.484 ± 0.104	0.484 ± 0.104	0.921 ± 0.032
UMAP 10D	0.698 ± 0.053	**0.349 ± 0.051**	763 ± 339	0.351 ± 0.149	0.500 ± 0.104	0.500 ± 0.104	0.936 ± 0.025
UMAP 15D	0.708 ± 0.021	0.370 ± 0.054	673 ± 301	0.336 ± 0.147	0.500 ± 0.108	0.500 ± 0.108	0.951 ± 0.016
UMAP 30D	**0.711 ± 0.039**	0.368 ± 0.041	652 ± 237	**0.367 ± 0.158**	**0.515 ± 0.108**	**0.515 ± 0.108**	0.948 ± 0.008
UMAP 50D	0.693 ± 0.053	0.386 ± 0.075	588 ± 280	0.352 ± 0.144	0.500 ± 0.105	0.500 ± 0.105	0.952 ± 0.015
UMAP 75D	0.706 ± 0.034	0.386 ± 0.101	582 ± 315	0.330 ± 0.152	0.499 ± 0.106	0.499 ± 0.106	**0.953 ± 0.010**

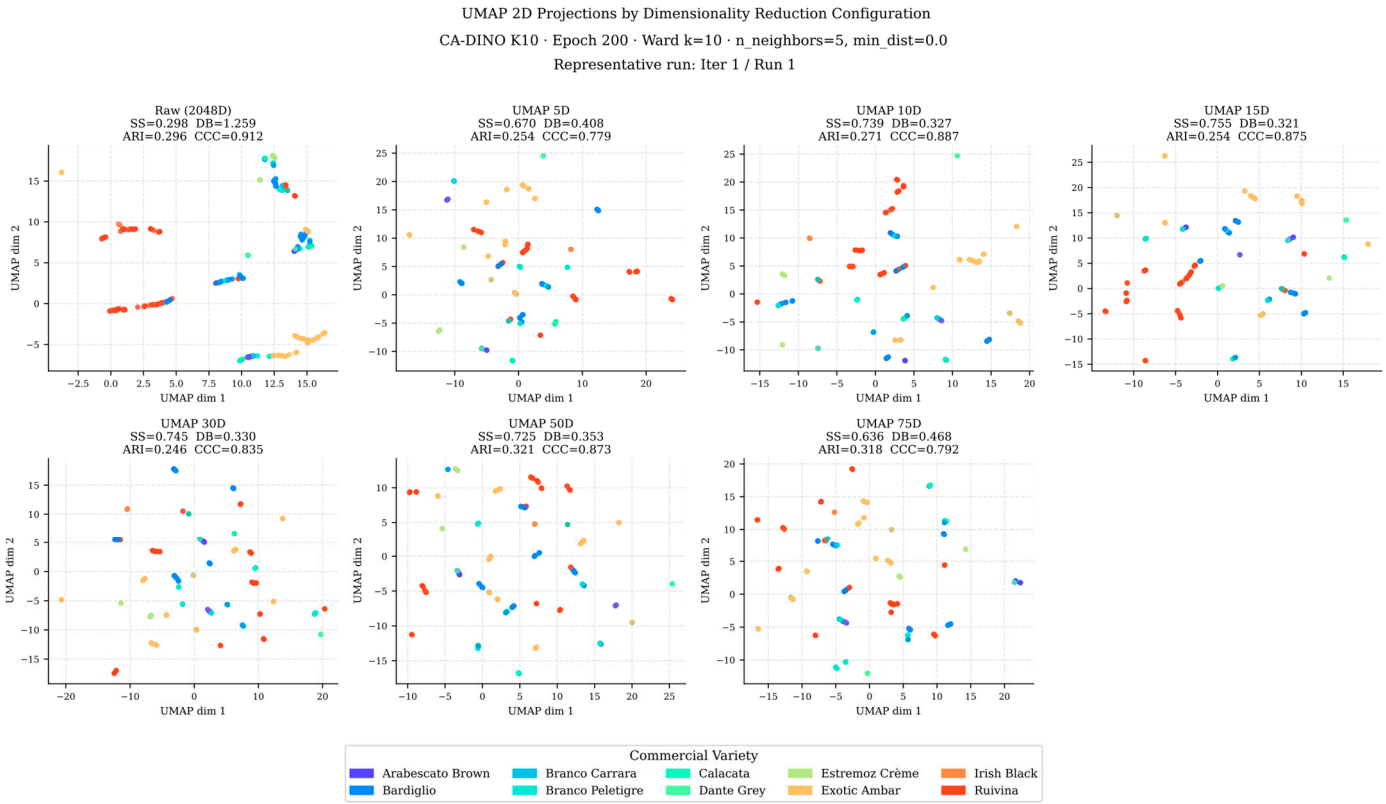

Figure 8. Two-dimensional UMAP projections of the CA-DINO K10 embedding space (epoch 200) for all seven dimensionality reduction configurations. Color encodes commercial variety (ground truth). Shown for Data split 1/Run 1 (representative run); quantitative results aggregated across all nine independent runs are reported in Table 4. UMAP 2D projection parameters: n_neighbors = 5, min_dist = 0.0.

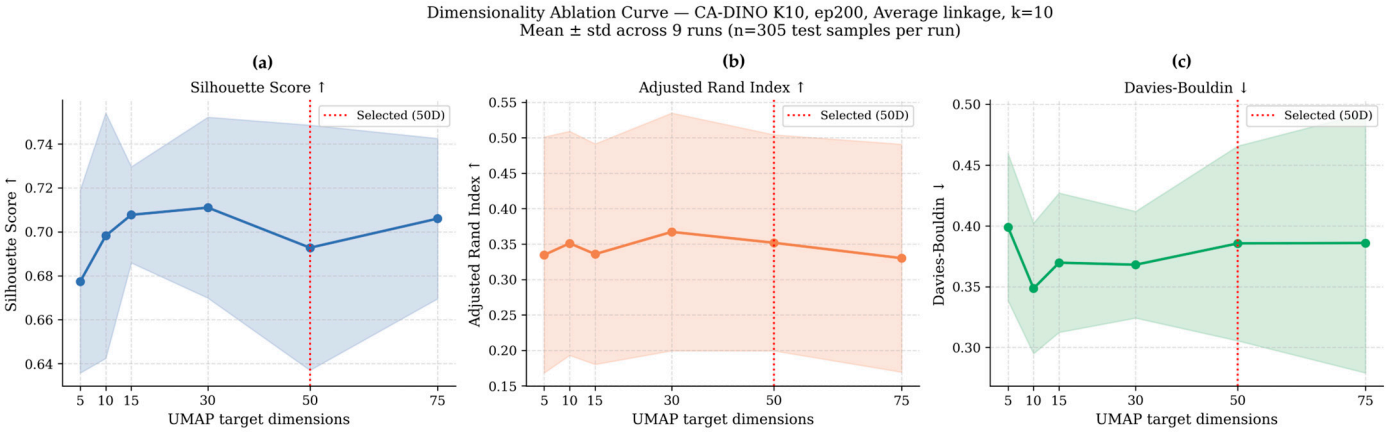

Figure 9. Mean ± standard deviation of (**a**) Silhouette Score, (**b**) Adjusted Rand Index, and (**c**) Davies-Bouldin Index, across nine runs as a function of UMAP target dimensionality (CA-DINO K10, epoch 200, Average linkage, k = 10, *n* = 305 test samples per run). The vertical dashed line marks the selected configuration (50D). Raw 2048D baseline excluded for legibility.

4.4. Linkage Method Ablation

Four agglomerative linkage criteria, namely Ward, Complete, Average, and Single, were compared on the UMAP 50D L2-normalized embeddings at epoch 200, evaluating internal clustering quality and hierarchical reproducibility at a dendrogram cut of k_c = 10 clusters. The UMAP configuration (50D, n_neighbors = 5, min_dist = 0.0) was fixed from the joint Phase 1 selection in Section 4.3 and was not re-optimized here. All metrics are

reported as mean ± standard deviation across nine independent measurements (3 data splits × 3 training runs, n = 305 test samples per split). Quantitative results are reported in Table 7. Dendrogram structures for a representative run are shown in Figure 10, and metric profiles across k_c = 2–20 are shown in Figure 11.

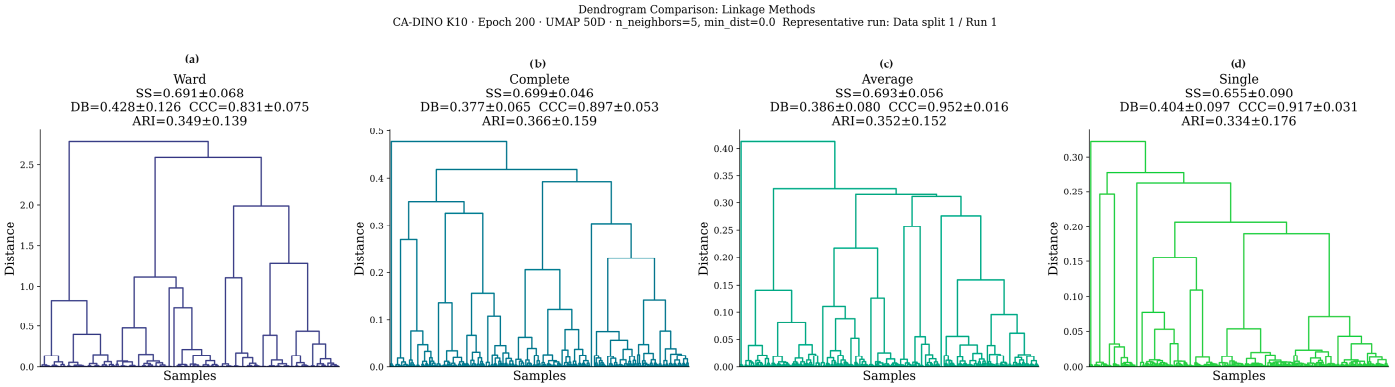

Figure 10. Dendrogram structures for (**a**) Ward, (**b**) Complete, (**c**) Average, and (**d**) Single linkage applied to UMAP 50D L2-normalized CA-DINO K10 embeddings at epoch 200. Each panel is annotated with mean ± standard deviation of SS, DB, CCC, and ARI across nine independent runs. Shown for run 1 of the first data split (representative run).

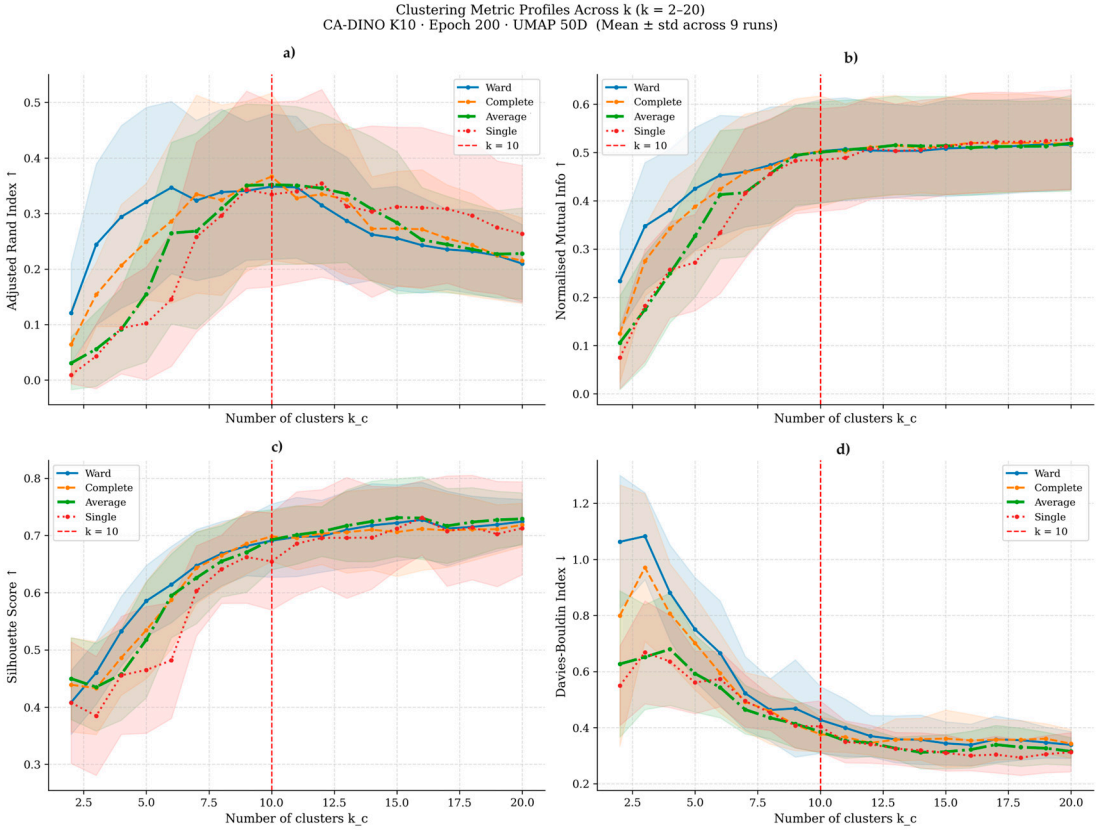

Figure 11. Clustering metric profiles across k_c = 2–20 for all four linkage methods (**a**) ARI, (**b**) NMI, (**c**) Silhouette Score and (**d**) Davies-Bouldin Index. Solid lines show the mean across 9 independent runs; shaded bands show ±1 standard deviation. The vertical dashed line marks k_c = 10. ARI and NMI values are included as descriptive references for alignment with commercial variety labels and do not constitute a selection criterion.

Table 7. Linkage criterion ablation: clustering metrics at dendrogram cut k_c = 10 (UMAP 50D, n_neighbors = 5, min_dist = 0.0, L2 normalization, CA-DINO K10, epoch 200). Values are mean ± standard deviation across 3 data splits × 3 training runs (N = 9, n = 305 test samples per split). The selected configuration (Average, marked with ‡) was chosen in the joint Phase 1 grid search of Section 4.3 using internal Silhouette Score on the validation set; commercial labels were not used during selection. Bold indicates best value per metric. ↑ higher is better; ↓ lower is better. Note: k_c denotes the number of clusters extracted from the dendrogram and is independent of the CA-DINO model parameter k.

Linkage	SS ↑	DB ↓	CH ↑	ARI ↑	NMI ↑	V-Measure ↑	CCC ↑
Ward	0.691 ± 0.064	0.428 ± 0.119	**632 ± 247**	0.349 ± 0.131	0.503 ± 0.107	0.503 ± 0.107	0.831 ± 0.071
Complete	**0.699 ± 0.044**	**0.377 ± 0.062**	609 ± 262	**0.366 ± 0.150**	**0.503 ± 0.115**	**0.503 ± 0.115**	0.897 ± 0.050
Average ‡	0.693 ± 0.053	0.386 ± 0.075	588 ± 280	0.352 ± 0.144	0.500 ± 0.105	0.500 ± 0.105	**0.952 ± 0.015**
Single	0.655 ± 0.085	0.404 ± 0.092	436 ± 254	0.334 ± 0.166	0.485 ± 0.107	0.485 ± 0.107	0.917 ± 0.030

Ward linkage achieved SS = 0.691 ± 0.064 and DB = 0.428 ± 0.119, with the highest inter-run instability of all methods on both metrics (CV = 9.3% for SS and 27.8% for DB). The cophenetic correlation coefficient was the lowest among the four methods (CCC = 0.831 ± 0.071, CV = 8.5%), indicating that Ward's variance-minimization objective produces a dendrogram geometry that is sensitive to variation in training initialization and data partition composition. Complete linkage achieved the best point estimates on geometric metrics (SS = 0.699 ± 0.044, DB = 0.377 ± 0.062) and the highest ARI across all four methods (0.366 ± 0.150), with reduced inter-run variance compared to Ward (SS CV = 6.3%). Its cophenetic correlation coefficient was intermediate (CCC = 0.897 ± 0.050, CV = 5.6%).

Average linkage, selected in Section 4.3 Phase 1 on the basis of internal Silhouette Score on the validation set, achieved SS = 0.693 ± 0.053 and DB = 0.386 ± 0.075 on the test set: values within one standard deviation of Complete linkage on both geometric metrics. The most distinctive property of Average linkage is its cophenetic correlation coefficient: CCC = 0.952 ± 0.015, a coefficient of variation of 1.6% across nine independent runs. This result indicates that the inter-sample distance relationships in the UMAP 50D embedding are preserved with near-perfect reproducibility in the Average linkage dendrogram across different data partitions and training initializations. The gap between Average and the next-best CCC is substantial: 0.055 points above Complete and 0.121 points above Ward, differences that exceed the standard deviation of either method. Single linkage was the weakest method on internal quality metrics overall (SS = 0.655 ± 0.085, DB = 0.404 ± 0.092, CH = 436 ± 254) and showed the highest inter-run variance on SS (CV = 13.0%), though it achieved a comparatively high CCC (0.917 ± 0.030), reflecting the characteristic behavior of single linkage in producing extended chain-like dendrogram structures that are geometrically faithful to the distance matrix even when cluster compactness is poor.

A notable feature of the results in the UMAP 50D space is the frequent coincidence of SS, DB, ARI, and NMI values across Ward, Complete, and Average linkage within individual runs (visible in the per-run output above). This pattern indicates that the three methods often produce identical or near-identical cluster assignments at the k_c = 10 cut level in this embedding geometry, with the primary differentiator being the structure of the dendrogram rather than the flat partition it yields at a specific cut depth. The CCC thus becomes the decisive discriminating metric among the three comparable methods, and Average linkage's CCC superiority is consistent and large across all nine runs. Single linkage diverges in cluster assignment quality, particularly in runs where the embedding contains borderline geometric structure, reflected in its highest SS variance.

The metric profiles across k_c = 2–20 (Figure 11) show that Silhouette scores peak at k_c = 2–3 for all methods and decline monotonically thereafter, reflecting the dominant large-

scale geometric structure of the 50D embedding space. Label alignment, reported as a descriptive reference only, given that commercial nomenclature does not constitute a visual categorization criterion for this pipeline, reaches a local maximum at k_c = 8–10 across all four methods, consistent with the ten commercial varieties present in the dataset. This pattern does not inform the linkage selection, which was made exclusively on internal geometric quality during Phase 1. Based on this analysis, Average linkage is confirmed as the pipeline default for all configurations reported in Sections 4.5 and 4.6.

4.5. Clustering Performance Across k Values

Table 8 reports clustering metrics for the Pure DINO baseline and all five CA-DINO configurations evaluated on the test set (*n* = 305 samples per split) using the pipeline locked from Sections 4.3–4.4 (UMAP 50D, n_neighbors = 5, min_dist = 0.0, Average linkage, dendrogram cut k_c = 10). This section is a post hoc sensitivity ablation around the already-adopted primary model, not a model selection contest. CA-DINO K10 was designated as the primary model prior to pipeline optimization on domain-alignment grounds, i.e., k = 10 matches the number of commercial varieties present in the dataset, making it the theoretically motivated prototype count for training a representation space intended to separate those categories. Sections 4.3 and 4.4 were conducted exclusively on CA-DINO K10 embeddings, and the pipeline configuration they produced is therefore specific to this model. The ablation variants (Pure DINO, K5, K8, K12, K15) are evaluated on first data split only (*N* = 3, 3 training runs) to establish directional sensitivity around the primary configuration; CA-DINO K10 is evaluated across the full three-iteration grid (*N* = 9, 3 data splits × 3 training runs) to characterize its cross-split and cross-initialization variability. This asymmetry in statistical coverage is fundamental to interpreting Table 8: ablation estimates from a single data partition cannot expose the cross-split variability visible in the *N* = 9 results, and any apparent metric advantage of a single-iteration ablation variant over K10 must be interpreted accordingly.

The most consistent finding across all configurations is that the cluster-aware objective substantially improves geometric embedding quality relative to the unconstrained self-supervised baseline. Pure DINO produced the weakest geometric cohesion of all tested models (SS = 0.660 ± 0.030, DB = 0.569 ± 0.012), with every CA-DINO variant yielding meaningfully higher Silhouette scores and substantially lower Davies-Bouldin indices. This confirms that the CA-DINO prototype-matching objective organizes the representation space into geometrically tighter and better-separated clusters regardless of the specific K value used during training, and that unconstrained self-supervised pretraining alone is insufficient to produce a partition-ready embedding at this level of geometric quality. Notably, Pure DINO achieved the highest cophenetic correlation coefficient of all models (CCC = 0.955 ± 0.002, CV = 0.2%), indicating that its Average-linkage dendrogram faithfully preserves pairwise distance structure in the UMAP 50D space even while producing geometrically looser clusters. This pattern reflects the known dissociation between hierarchical distance fidelity and cluster compactness: a dendrogram can be geometrically consistent yet fail to produce well-separated partitions if the underlying embedding lacks tight intra-class concentration.

Among the CA-DINO variants, no single K configuration dominates uniformly across all metrics, and the performance differences within the CA-DINO family are substantially smaller than the gap between any CA-DINO model and Pure DINO on internal geometric metrics. CA-DINO K8 achieves the highest mean Silhouette score (SS = 0.749 ± 0.051) and the highest NMI across all configurations (0.520 ± 0.071), while CA-DINO K15 achieves the lowest Davies-Bouldin index (DB = 0.321 ± 0.015) and the lowest SS inter-run variance (CV = 1.9%). However, both estimates derive from a single data partition. CA-DINO K10 evaluated on the first data split alone produces SS values of 0.725,

0.686, and 0.721 across its three training runs, fully competitive with K8's single-iteration estimate, and it is only the inclusion of the two extra data partitions (which introduce harder data partitions, as evidenced by lower geometric scores of the second data split) that depresses the K10 nine-run mean. This is precisely the information that single-iteration ablation cannot provide: the cross-split performance floor. CA-DINO K15 achieves the tightest geometry but at the cost of the weakest label alignment in the CA-DINO family (ARI = 0.229 ± 0.105, NMI = 0.401 ± 0.139), consistent with a training objective that over-partitions the representation space relative to the ten-variety structure. The anomalously high Calinski-Harabasz variance at K15 (CH = 1472 ± 876, CV = 59.5%) reflects the known sensitivity of this index to the ratio of between-cluster to within-cluster dispersion in compact spaces and is not a reliable quality signal at this configuration. CA-DINO K12 achieved the lowest CCC among CA-DINO variants (0.943 ± 0.020), indicating that increasing K beyond the target variety count introduces mild instability in hierarchical distance preservation.

CA-DINO K10 achieves SS = 0.693 ± 0.053, DB = 0.386 ± 0.075, ARI = 0.352 ± 0.144, NMI = 0.500 ± 0.105, and CCC = 0.952 ± 0.015 across the full $N = 9$ evaluation grid. These are not the single-best point estimates in Table 8, nor are they expected to be: the $N = 9$ design exposes cross-split variability that inflates standard deviation and moderates the mean relative to a cherry-picked single-iteration result. The appropriate comparison is not K10 ($N = 9$) against K8 ($N = 3$), but K10 against itself across all three data splits, which demonstrates consistent geometric performance and a CCC of variation of 1.6% across nine independent measurements. Alignment with commercial variety labels (ARI, NMI) is reported as a descriptive reference only: commercial nomenclature does not constitute a visual categorization target for this pipeline, and these metrics are not the basis for model selection. The primary selection criteria are geometric embedding quality, cross-validated hierarchical reproducibility, and domain alignment of the training objective. On all three grounds, i.e., theoretical alignment of K = 10 with the target variety count, competitive and stable geometric performance across the full three-split evaluation grid, and the comprehensive statistical coverage that no other configuration in this ablation provides, CA-DINO K10 is confirmed as the pipeline default and used exclusively in the analyses reported in Section 4.6.

Table 8. Clustering metrics: Pure DINO baseline vs. CA-DINO across K values (epoch 200, UMAP 50D, Average linkage, dendrogram cut $k_c = 10$, $n = 305$ test samples). CA-DINO K10: $N = 9$ (3 data splits × 3 training runs); all other models: $N = 3$ (Data split 1, 3 training runs). Values are mean ± standard deviation. Bold indicates best CA-DINO value per metric. ↑ higher is better; ↓ lower is better. Note: K denotes the CA-DINO training prototype parameter; k_c denotes the independent dendrogram cut used at inference.

Model	*N*	SS ↑	DB ↓	CH ↑	ARI ↑	NMI ↑	V-Measure ↑	CCC ↑
Pure DINO	3	0.660 ± 0.030	0.569 ± 0.012	167 ± 40	0.210 ± 0.042	0.423 ± 0.046	0.423 ± 0.046	0.955 ± 0.002
CA-DINO K5	3	0.712 ± 0.069	0.353 ± 0.033	**794 ± 126**	0.288 ± 0.060	0.467 ± 0.055	0.467 ± 0.055	0.952 ± 0.015
CA-DINO K8	3	**0.749 ± 0.051**	0.350 ± 0.061	679 ± 227	**0.355 ± 0.099**	**0.520 ± 0.071**	**0.520 ± 0.071**	0.953 ± 0.010
CA-DINO K10	9	0.693 ± 0.053	0.386 ± 0.075	588 ± 280	0.352 ± 0.144	0.500 ± 0.105	0.500 ± 0.105	0.952 ± 0.015
CA-DINO K12	3	0.693 ± 0.044	0.359 ± 0.030	682 ± 117	0.320 ± 0.115	0.450 ± 0.115	0.450 ± 0.115	0.943 ± 0.020
CA-DINO K15	3	0.733 ± 0.014	**0.321 ± 0.015**	1472 ± 876	0.229 ± 0.105	0.401 ± 0.139	0.401 ± 0.139	**0.955 ± 0.020**

4.6. Emergent Visual Grouping Structure

The qualitative analysis examines how the CA-DINO pipeline organizes the 305 test samples across the visual similarity space learned during training. The Average-linkage dendrogram over the UMAP 50D embeddings encodes a continuous multi-scale structure interrogated here across six levels of resolution, from the root binary split down to the

twenty-group fine-grained partition, tracing three recurring phenomena: cross-category merging of commercially distinct varieties that share visual properties; intra-category splitting of commercially unified varieties that harbor visual sub-populations; and coherent pure family formation where commercial and visual boundaries genuinely coincide. Across the five consecutive level transitions, 12 of the 39 tracked groups undergo a split while 27 pass through intact, confirming that structural change is concentrated at specific nodes rather than diffuse. The cluster composition heatmap at the k = 10 reference cut is shown in Figure 12, the level-by-level dendrogram split lineage in Figure 13, and the UMAP 2D projection with cluster assignments overlaid on commercial variety markers in Figure 14.

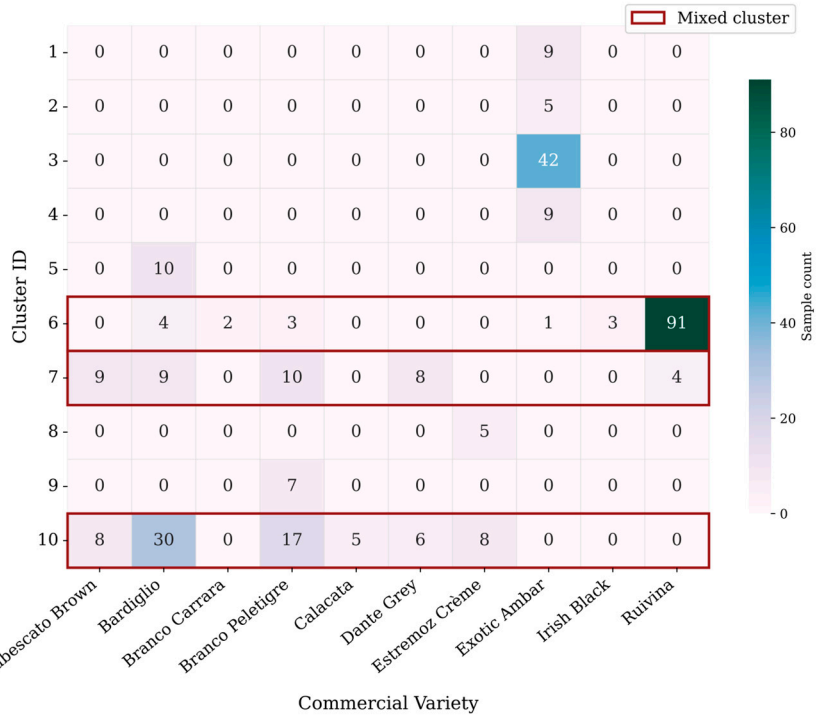

Figure 12. Cluster composition heatmap at the k = 10 reference cut (Average linkage, UMAP 50D, L2 normalization, epoch 200, n = 305; representative run: Data Split 1/Run 2). Rows correspond to cluster IDs; columns correspond to commercial varieties; cell values report sample counts. Red borders identify mixed clusters containing samples from more than one commercial variety. Seven of ten clusters are pure (C1–C5, C8–C9); three are mixed: C6 is dominated by Ruivina (91/104 samples) with minor contamination from Bardiglio, Irish Black, Branco Peletigre, Branco Carrara, and one Exotic Ambar specimen; C7 is the gray-veined cross-category cluster (Branco Peletigre 10, Bardiglio 9, Arabescato Brown 9, Dante Grey 8, Ruivina 4); and C10 is a heterogeneous dark complex dominated by Bardiglio (30/74) with Branco Peletigre, Arabescato Brown, Estremoz Crème, Dante Grey, and Calacata.

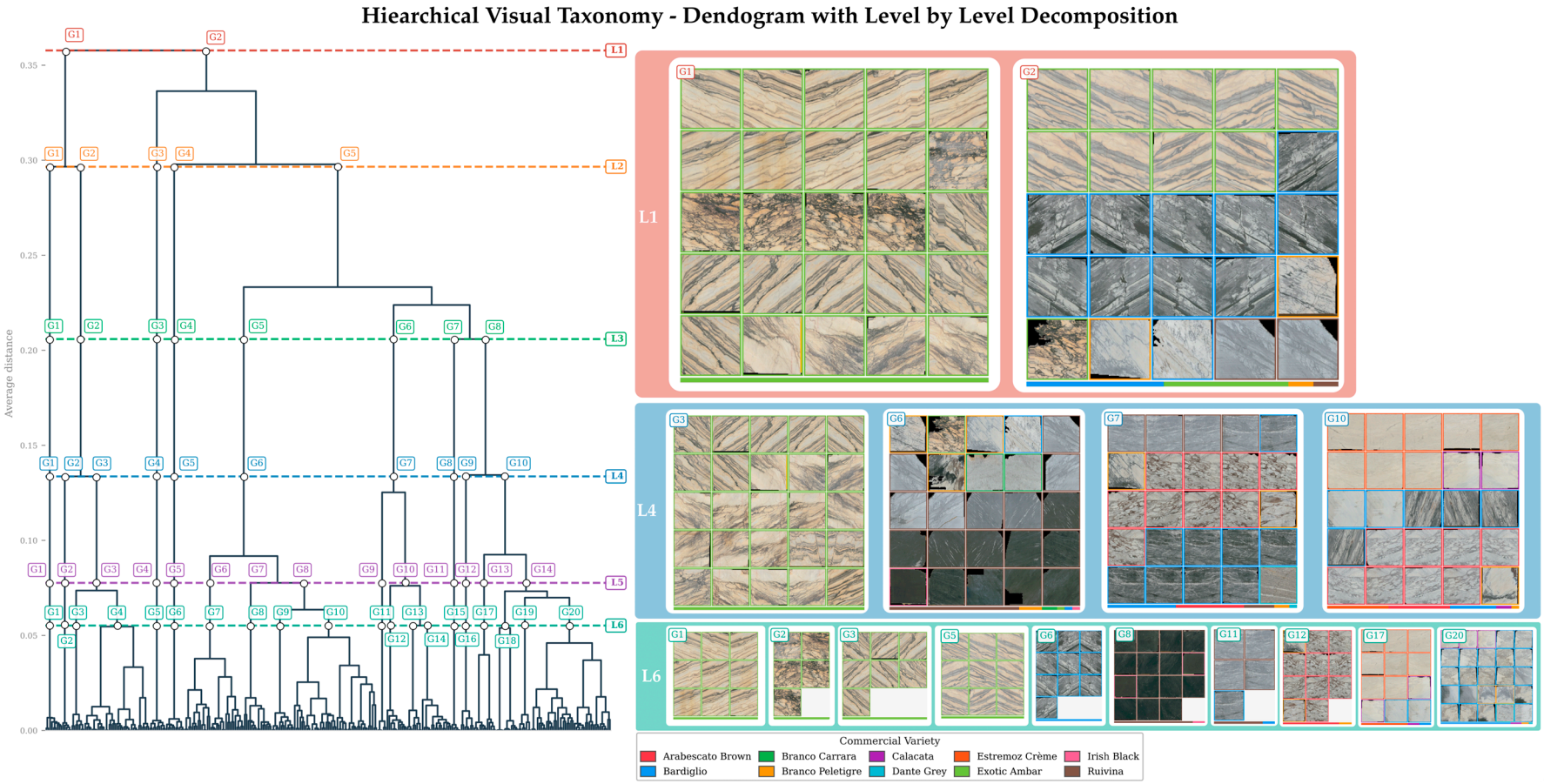

Figure 13. Level-by-level hierarchical decomposition of the CA-DINO K10 embedding space (Average linkage, UMAP 50D, epoch 200, test set, $n = 305$). Left: Full dendrogram annotated with six horizontal cut levels (L1–L6), yielding 2, 5, 8, 10, 14, and 20 groups respectively, with cluster labels color-coded by dominant commercial variety. Right: Representative image grids at three analytically significant levels: L1, where the root binary split isolates a pure Exotic Ambar branch ($n = 56$) from the cool-and-dark mixed majority ($n = 249$), reflecting the dominant chromatic axis without label supervision; L4 (the k = 10 reference cut), where Exotic Ambar has resolved into four pure sub-clusters documenting intra-variety heterogeneity, and a gray-veined cross-category family ($n = 40$; Branco Peletigre, Arabescato Brown, Bardiglio, Dante Grey) coheres across three commercial varieties; and L6, the finest cross-section, where pure single-variety isolates coexist with persistent cross-commercial groupings, exposing the full mismatch between visual and commercial taxonomies. Border colors in image grids correspond to commercial variety labels per legend.

https://doi.org/10.3390/app16094137

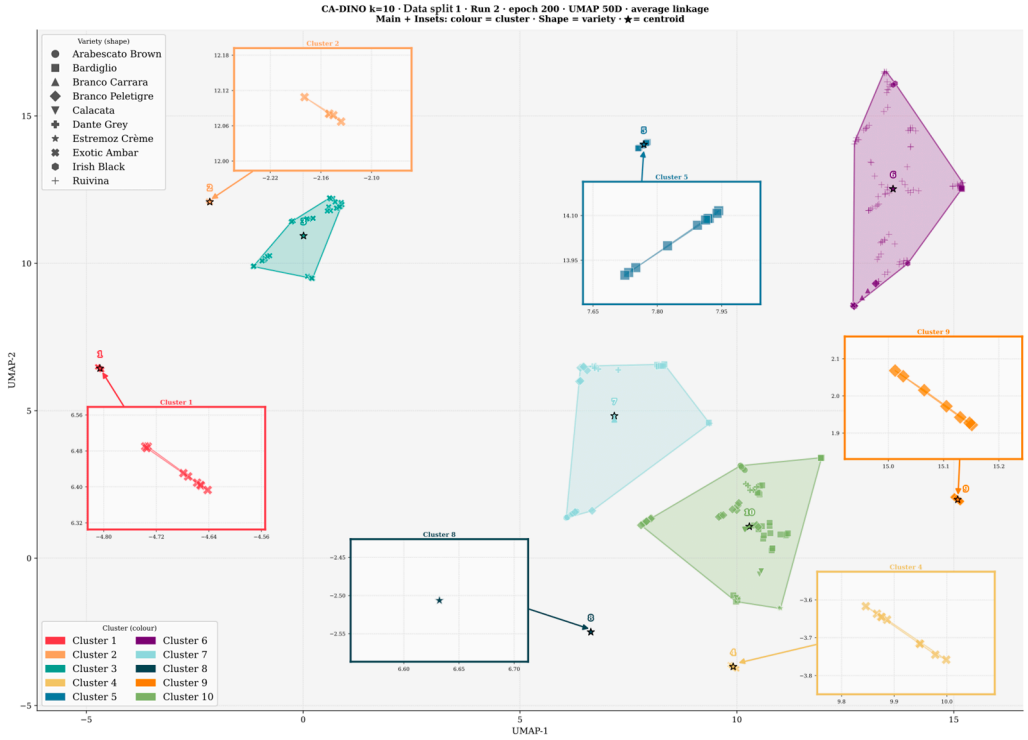

Figure 14. UMAP 2D scatter plot of CA-DINO test-set embeddings (Average linkage, UMAP 50D, epoch 200, n = 305; representative run: Data Split 1/Run 2). In the main panel, point color encodes cluster ID at the k = 10 cut and point shape encodes commercial variety; cluster centroids are marked with ★. Colored insets provide a per-cluster zoom with convex hull overlaid; arrow connectors link each inset to its centroid in the main projection. Regions where point shape and color do not align within a visually coherent zone identify cross-category groupings, i.e., specimens from distinct commercial varieties assigned to the same cluster on the basis of visual similarity alone.

4.6.1. Level-by-Level Structure

The root binary split at Level 1 partitions the 305 test samples into two markedly unequal branches (n = 56 and n = 249). The smaller branch is a pure Exotic Ambar group, isolating the dataset's warmest and most chromatically saturated variety. The larger branch consolidates all remaining varieties into a single mixed pool dominated by Ruivina (95), Bardiglio (53), and the gray-veined family, namely Branco Peletigre (37), Arabescato Brown (17), and Dante Grey (14), alongside Estremoz Crème (13), Calacata (5), Irish Black (3), and Branco Carrara (2). This primary split reproduces the most salient chromatic axis in the dataset, warm amber-ground stones versus cool and dark stones, confirming that the embedding's top-level organization is perceptually grounded without label supervision.

At Level 2 (five groups), the Exotic Ambar branch undergoes its first internal split into two pure sub-groups (n = 9 and n = 47), while within the cool branch a pure Bardiglio sub-population (n = 10) and a small pure Exotic Ambar group (n = 9) detach from the large mixed pool, the latter confirming that nine Exotic Ambar specimens, initially co-embedded with the cool-toned majority at Level 1, are recoverable as a distinct pure group at this finer resolution. The remaining mixed residual numbers 230 samples across nine varieties.

At Level 3 (eight groups), only the 230-sample mixed residual undergoes further splitting, producing four sub-groups. A 40-sample mixed group coheres from specimens sharing gray-on-gray veining, namely Branco Peletigre (10), Bardiglio (9), Arabescato Brown (9), Dante Grey (8), and four Ruivina samples whose visual texture is sufficiently

gray-toned to co-embed with this family, forming a cross-commercial visual cluster with no counterpart in the commercial taxonomy. Simultaneously, a Bardiglio-plurality dark complex (n = 81; Bardiglio 30, Branco Peletigre 24, Estremoz Crème 8, Arabescato Brown 8, Dante Grey 6, Calacata 5) separates from the Ruivina core (n = 104), and a compact pure Estremoz Crème group (n = 5) isolates.

At Level 4 (ten groups—the k = 10 reference cut), the Exotic Ambar branch has fully resolved into four pure clusters (C1: n = 9; C2: n = 5; C3: n = 42; C4: n = 9), capturing 65 of 66 Exotic Ambar test samples and documenting extensive intra-variety heterogeneity within this commercially unified category. The gray-veined cross-category cluster (C7, n = 40) persists intact, passing through from Level 3 without further splitting. Within the dark complex, a pure Branco Peletigre satellite (C9, n = 7) detaches, leaving the Bardiglio-dominated residual as C10 (n = 74; Bardiglio 30, Branco Peletigre 17, Estremoz Crème 8, Arabescato Brown 8, Dante Grey 6, Calacata 5).

At Level 5 (fourteen groups), three groups undergo splitting. The Ruivina core (C6, n = 104) fragments for the first time into three mixed sub-groups: a diffuse 32-sample boundary zone (Ruivina 24, plus Branco Peletigre, Branco Carrara, Exotic Ambar, Bardiglio, and Irish Black), a compact 15-sample cluster of Ruivina co-embedded with Irish Black (Ruivina 13, Irish Black 2), and a 57-sample sub-group that is overwhelmingly Ruivina (54/57). Concurrently, the gray-veined cluster (n = 40) undergoes its first internal split: a small 5-sample fringe group (Ruivina 4, Bardiglio 1) separates from the tighter 35-sample core (Branco Peletigre 10, Arabescato Brown 9, Bardiglio 8, Dante Grey 8), which retains the defining visual signature of the family. The Bardiglio dark complex (C10, n = 74) also splits, yielding a 12-sample light-toned satellite (Estremoz Crème 8, Calacata 2, Bardiglio 2) and a 62-sample dark residual (Bardiglio 28, Branco Peletigre 17, Arabescato Brown 8, Dante Grey 6, Calacata 3).

At Level 6 (twenty groups), the decomposition reaches its finest resolved cross-section. The large pure Exotic Ambar sub-group (n = 42) splits into two pure sub-groups (n = 7 and n = 35), yielding five pure Exotic Ambar groups in total (sizes: 9, 5, 7, 35, 9) that document the full intra-variety diversity of this category. Within the Ruivina lineage, the predominantly Ruivina 57-sample sub-group splits and produces a first fully pure Ruivina cluster (n = 45), confirming that a substantial Ruivina core exists at sufficient resolution. The 35-sample gray-veined core splits into three sub-groups: an Arabescato Brown-dominant mixed group (n = 11; Arabescato Brown 9, Branco Peletigre 2), a pure Bardiglio group (n = 8), and a Dante Grey–Branco Peletigre mixed boundary zone (n = 16; Dante Grey 8, Branco Peletigre 8). Similarly, the 62-sample dark residual splits into an Arabescato Brown-anchored mixed group (n = 12; Arabescato Brown 8, Bardiglio 4), a pure Branco Peletigre group (n = 7), and a Bardiglio-dominated residual (n = 43). This level provides the most analytically informative cross-section: the model simultaneously partitions specimens sharing a commercial label into distinct visual sub-types and clusters specimens from different commercial categories into a common visual family, exposing the full texture of the mismatch between commercial and visual taxonomies.

Variety-level tracking reveals differentiated behavior consistent with each variety's visual properties. Exotic Ambar is the most fragmented variety: 65 of its 66 test samples distribute across four pure clusters at the k = 10 cut (dominant core C3: n = 42; satellites C1, C4: n = 9 each; and C2: n = 5), with a single specimen in the mixed Ruivina-dominated C6, confirming that this commercially unified variety encompasses multiple visually distinguishable sub-populations. Ruivina presents the complementary pattern: 91 of 95 test samples (95.8%) concentrate in a single cluster (C6) at k = 10, with internal sub-structure, a pure 45-sample core, emerging only at Level 6. Bardiglio is the most dispersed variety, contributing to four clusters at the k = 10 cut (C5 pure n = 10; C6 n = 4; C7 n = 9; C10 n = 30), consistent with its broad visual range spanning dark monolithic to gray-

veined sub-types and making it the dominant contaminant in both the gray-veined cross-category family and the heterogeneous dark complex.

4.6.2. Three Illustrative Cases

The three cases shown in Figure 15 extract specific sub-trees from the full dendrogram to illustrate each of the three phenomena identified above.

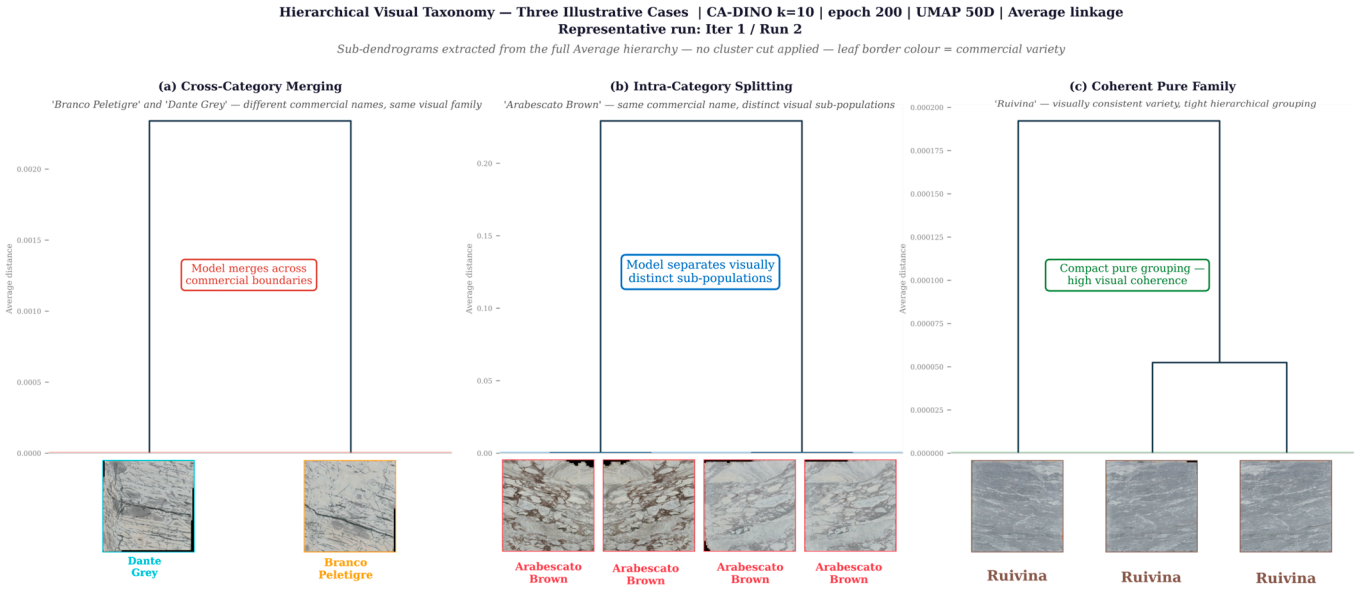

Figure 15. Three illustrative cases extracted from the Average-linkage dendrogram (UMAP 50D, epoch 200, n = 305, representative run: Data split 1/Run 2). (**a**): Cross-category merging—a Branco Peletigre specimen and a Dante Grey specimen joined at Average distance 0.0023, illustrating that the model places commercially distinct stones in immediate proximity when their visual signatures are indistinguishable at the texture level. (**b**): Intra-category splitting—two visually distinct Arabescato Brown sub-populations with intra-group distances of 0.0002 and 0.0003, separated by an inter-group distance of 0.2293 (inter-to-intra ratio of 891), demonstrating that a single commercial variety can harbor multiple recoverable visual sub-populations. (**c**): Coherent pure family—a 3-sample Ruivina subtree at Average distance 0.0002, corresponding to 0.001× the dataset mean pairwise distance of 0.2406, demonstrating that where commercial and visual boundaries align, the model produces exceptionally tight, stable, and interpretable groupings. Each panel shows the relevant sub-dendrogram above variety-labeled color-bordered thumbnails in dendrogram leaf order; no cluster cut line is imposed.

Cross-category merging (Figure 15a). The most precise instance of cross-category merging in the hierarchy involves a Branco Peletigre specimen and a Dante Grey specimen joined at Average distance 0.0023—the closest inter-variety pair in the full embedding space of 305 samples. These two specimens, carrying distinct commercial designations, are nearest neighbors in the representation space, and their merge height of 0.0023 falls within the same order of magnitude as intra-variety merge heights throughout the hierarchy. This is not an isolated occurrence: at Level 2 of Figure 13, a 40-sample group consolidates Arabescato Brown, Bardiglio, Branco Peletigre, and Dante Grey (22%, 22%, 25%, 20% respectively) as a persistent and robust cross-category affinity among gray-veined stones: a visual family that is structurally invisible to any commercial classification system. The commercial significance of this finding is direct: it identifies specimens that are visually indistinguishable at the texture level despite carrying different market designations and origin labels, a form of discrimination that supervised classifiers trained on commercial categories cannot represent.

Intra-category splitting (Figure 15b). Within the commercial category of Arabescato Brown, the hierarchy identifies two sub-populations with intra-group distances of 0.0002 and 0.0003, respectively, separated by an inter-group distance of 0.2293, yielding an inter-to-intra ratio of 891. The two sub-populations only rejoin in the dendrogram at Average height 0.2293, three orders of magnitude above their individual merge heights, confirming that the separation is genuine and not an artifact of local neighborhood structure. The finding directly challenges the implicit assumption of commercial taxonomy that each named variety constitutes a visually homogeneous class: within a single commercial label, the model recovers visually distinct sub-populations that carry an inter-to-intra distance ratio of 891, a separation that would be entirely collapsed and concealed by any label-supervised approach. The same phenomenon is reproduced at coarser scale by Exotic Ambar (four pure clusters at k = 10) and Bardiglio (four cluster memberships at k = 10), confirming that intra-variety visual variability is structured and recoverable rather than random noise across multiple varieties in this dataset.

Coherent pure family (Figure 15c). A 3-sample Ruivina sub-tree forms a coherent pure family at an intra-family distance of 0.0001, corresponding to only 0.001× the dataset mean pairwise distance of 0.2406, the tightest pure family identified in the hierarchy. This compactness ratio places the Ruivina sub-tree at an extreme of the distribution of within-group distances, confirming that where commercial and visual boundaries align, the CA-DINO representation produces exceptionally concentrated groupings with no comparable parallel in the embedding space. At Level 6 of Figure 13, the same pattern is reproduced across multiple groups: the twenty-group fine-grained partition yields sub-populations of high visual coherence, with pure groups displaying near-identical texture and color profiles and mixed groups marking the genuine visual overlap zones that commercial boundaries fail to respect. Taken together, the three cases of Figure 15 demonstrate that the Average-linkage hierarchy encodes a genuine multi-scale visual taxonomy: it simultaneously reveals the commercial boundaries imposed unnecessarily on visually equivalent material (Case A), the internal heterogeneity that commercial classification conceals within a single label (Case B), and the visual coherence that commercial classification correctly captures where real discriminability exists (Case C). No single flat cut of the dendrogram captures all three simultaneously, a property intrinsic to the hierarchical structure and precisely what any flat k-partition, including the k = 10 reference cut, cannot represent.

4.7. Cross-Split Replication and Cluster Stability

Tier 2 evaluates whether the taxonomy produced by the frozen CA-DINO pipeline is a reproducible property of the data manifold or an artifact of a single training initialization and data partition. The frozen pipeline: UMAP 50 dimensions, n-neighbors = 5, minimum distance = 0.0, Average linkage, epoch 200 checkpoint, is applied identically to CA-DINO K10 embeddings extracted from all three independent data partitions, each trained with three independent weight initializations, yielding nine test-set measurements in total (3 partitions × 3 runs, n = 305 per evaluation). Three aspects of stability are examined: cross-partition metric consistency (Table 9), within-partition cluster label stability via pairwise Adjusted Rand Index (Table 10), and the persistence of the three key taxonomy phenomena identified in Section 4.6 across all nine runs (Table 11). No configuration decisions are made on the basis of these results; the pipeline is fully frozen before any test-set access at this tier.

Table 9. Cross-partition metric consistency for the frozen CA-DINO K10 pipeline was evaluated across three independent data partitions, each with three independent training initializations (N = 9 measurements total). Values report mean ± standard deviation over the three runs within each partition; the overall mean row aggregates all nine measurements. The frozen pipeline configuration is: UMAP 50D, n-neighbors = 5, minimum distance = 0.0, Average linkage, epoch 200 checkpoint, k = 10. SS: Silhouette Score (↑ better); DB: Davies-Bouldin Index (↓ better); CH: Calinski-Harabasz Index (↑ better); ARI: Adjusted Rand Index; NMI: Normalized Mutual Information; V: V-measure; CCC: Cophenetic Correlation Coefficient (↑ better). ARI, NMI, and V are computed against commercial variety labels used solely as a post hoc reference and are not optimization targets.

Partition	SS ↑	DB ↓	CH ↑	ARI	NMI	V	CCC ↑
Data Split 1 (3 runs)	0.692 ± 0.052	0.350 ± 0.034	581.6 ± 319.4	0.437 ± 0.089	0.551 ± 0.051	0.551 ± 0.051	0.954 ± 0.006
Data Split 2 (3 runs)	0.669 ± 0.023	0.411 ± 0.083	608.4 ± 381.2	0.322 ± 0.173	0.486 ± 0.138	0.486 ± 0.138	0.957 ± 0.018
Data Split 3 (3 runs)	0.665 ± 0.024	0.412 ± 0.066	760.5 ± 310.1	0.267 ± 0.094	0.461 ± 0.128	0.461 ± 0.128	0.952 ± 0.009
Overall mean (9 runs)	0.675 ± 0.034	0.391 ± 0.064	650.2 ± 304.7	0.342 ± 0.132	0.499 ± 0.105	0.499 ± 0.105	0.954 ± 0.011
CV ‡ (%)	5	16.2	46.9	38.4	21.1	21.1	1.1

‡ CV: coefficient of variation ($\sigma/|\mu| \times 100\%$). CH is highly sensitive to partition size and class-count ratios in imbalanced datasets; its high CV reflects this sensitivity rather than geometric instability.

Table 10. Within-partition cluster label stability measured by pairwise ARI between the three independently initialized training runs within each data partition. Each cell reports the ARI between the flat k = 10 partition assignments of the two runs. The final column reports the mean ± standard deviation of the three pairwise comparisons within each partition. A low pairwise ARI reflects label permutation sensitivity of the stochastic UMAP–Average linkage pipeline rather than structural disagreement; the Cophenetic Correlation Coefficient (Table 9) and phenomena persistence (Table 11) provide complementary evidence of structural robustness.

Partition	Run 1 vs. Run 2	Run 1 vs. Run 3	Run 2 vs. Run 3	Mean ARI ± Std
Data Split 1	0.317	0.293	0.583	0.398 ± 0.161
Data Split 2	0.161	0.153	0.349	0.221 ± 0.111
Data Split 3	0.197	0.284	0.183	0.221 ± 0.055
Overall mean ‡	—	—	—	0.280 ± 0.134

‡ The Overall mean row reports the mean ± standard deviation of the three partition-level Mean ARI values (0.398, 0.221, 0.221).

Table 11. Persistence of three key taxonomy phenomena across all nine independent measurements (3 data partitions × 3 training initializations). A tick (✓) indicates the phenomenon is present in the flat k = 10 partition of that run; a cross (✗) indicates absence. The frequency row reports the number of runs in which the phenomenon is detected out of nine. Phenomenon definitions are given in the table footnotes.

Run	Cross-Category Merging [a]	Intra-Category Splitting [b]	Coherent Pure Families [c]
Data Split 1/Run 1	✓	✓	✓
Data Split 1/Run 2	✓	✓	✓
Data Split 1/Run 3	✓	✓	✓
Data Split 2/Run 1	✓	✓	✓
Data Split 2/Run 2	✓	✓	✓
Data Split 2/Run 3	✓	✓	✓

Data Split 3/Run 1	✓	✓	✓
Data Split 3/Run 2	✓	✓	✓
Data Split 3/Run 3	✓	✓	✓
Frequency	9/9	9/9	9/9

[a] Cross-Category Merging: at least one cluster contains samples from ≥ 2 distinct gray-veined commercial varieties (Arabescato Brown, Dante Grey, Bardiglio, Branco Peletigre), indicating that visual affinity overrides commercial naming boundaries. [b] Intra-Category Splitting: Exotic Ambar samples are distributed across ≥ 2 distinct clusters, indicating that this commercially unified category encompasses visually distinct sub-populations. [c] Coherent Pure Families: at least one cluster achieves ≥ 80% compositional purity (dominant variety ≥ 80% of cluster membership), indicating that commercial and visual boundaries coincide for at least one variety in every run.

4.7.1. Cross-Partition Metric Consistency

Table 9 reports the mean and standard deviation of the full metric suite for each data partition and across all nine runs. The two metrics with the clearest interpretation as pipeline signatures, namely the Silhouette Score and the Cophenetic Correlation Coefficient, show low variation across partitions. The overall mean SS is 0.675 ± 0.034, with a coefficient of variation of 5.0%, placing it at the boundary of the low-variation threshold. The CCC is the most stable metric in the suite, with a overall mean of 0.954 ± 0.011 and a coefficient of variation of 1.1%, confirming that the hierarchical fidelity of the Average-linkage dendrogram is a structural property of the embedding manifold that is insensitive to both training initialization and data partition. Per-partition means range from 0.952 to 0.957, a spread of 0.005, which is negligible in practical terms.

The internal geometric metrics show moderate between-partition variation consistent with expected sampling effects. The Davies-Bouldin Index ranges from 0.350 ± 0.034 (Data split 1) to 0.412 ± 0.082 (Data splits 2 and 3), with a coefficient of variation of 16.2% across all nine runs. The Calinski-Harabasz Index shows the highest relative variation (CV = 46.9%), which is expected: CH is disproportionately sensitive to total sample count and cluster size ratios, both of which vary across the three partitions due to stratified sampling of an imbalanced dataset. This metric should therefore be interpreted within partitions rather than across them.

The external metrics, namely ARI, NMI, and V-measure, exhibit higher between-run variation than the internal and hierarchical metrics. The ARI overall mean of 0.342 ± 0.132 (CV = 38.5%) and the NMI overall mean of 0.499 ± 0.105 (CV = 21.1%) reflect the dual role of these indices: they measure not only clustering consistency but also the degree to which the discovered visual structure aligns with the commercial variety labels used as the external reference. As established in Section 3.5.3, this alignment is neither guaranteed nor expected for an appearance-based taxonomy discovering visual structure that commercial labels do not reliably encode. The per-partition means are systematically higher for Data split 1 (ARI = 0.437 ± 0.090, NMI = 0.551 ± 0.050) than for data splits 2 and 3 (ARI = 0.322 ± 0.172 and 0.267 ± 0.094, respectively), a pattern addressed in Section 4.7.2.

4.7.2. Within-Partition Cluster Label Stability

Table 10 reports the pairwise Adjusted Rand Index between the three training runs within each data partition. The overall mean pairwise ARI is 0.280 ± 0.134, a value that falls below the threshold conventionally associated with stable cluster partitions. This result warrants careful interpretation, as the pairwise ARI conflates two distinct sources of variability: genuine structural disagreement about which samples belong together, and label permutation, in which the same visual groupings are recovered but receive different cluster indices due to the stochastic nature of UMAP initialization and the hierarchical merge order.

Within data split 1, the pairwise stability is notably higher for the Run 2 vs. Run 3 comparison (ARI = 0.583) than for comparisons involving Run 1 (ARI = 0.317 and 0.293). This asymmetry suggests that two of the three runs converge to nearly identical partition structures, while one run finds a locally distinct but geometrically equivalent solution. The same pattern, in which one run within a partition diverges from the other two, is visible in data split 2 (where Run 3 has a higher ARI = 0.349 against Run 2 than the Run 1–Run 2 pair at ARI = 0.161) and partially in Data split 3. This is the expected behavior of average linkage applied to UMAP-reduced embeddings: the hierarchical structure is reproducible at the level of merge order and cophenetic distances (CCC = 0.954 ± 0.011), but small perturbations in the embedding manifold from different random initializations shift the precise assignment of borderline samples near cluster boundaries, producing partition label differences that disproportionately inflate the ARI penalty. The phenomenon's persistence analysis in Section 4.7.3 confirms that these partition-level differences do not alter the substantive taxonomy findings.

4.7.3. Phenomena PersistenceT

Table 11 reports the presence or absence of the three key taxonomy phenomena, namely cross-category merging, intra-category splitting, and coherent pure family formation, in each of the nine independent runs. All three phenomena are present in every run, yielding a frequency of 9/9 for each. This result is the central stability finding of the Tier 2 evaluation: the structural properties of the visual taxonomy documented qualitatively in Section 4.6 are not artifacts of a particular training initialization or data partition but are universal properties of the CA-DINO embedding space for this dataset.

Specifically, in all nine runs: (i) at least two gray-veined varieties (Arabescato Brown, Dante Grey, Bardiglio, Branco Peletigre) co-occur within a single cluster, confirming the persistent cross-category visual affinity identified at Level 2 of the dendrogram; (ii) Exotic Ambar spans at least two distinct clusters, confirming the structural fragmentation of this commercially unified but visually heterogeneous category; and (iii) at least one cluster achieves ≥ 80% compositional purity, confirming that coherent pure families, i.e., where commercial and visual boundaries coincide, are a robust and recoverable property of the representation space. The universality of these three phenomena across training runs with different random seeds, and across data partitions with different train-test stratifications establishes them as genuine geometric properties of the learned embedding rather than partition-specific coincidences.

Taken together, the three tiers of Tier 2 evidence, namely stable CCC (0.954 ± 0.011, CV = 1.1%), adequate SS stability (0.675 ± 0.034, CV = 5.0%), and 9/9 phenomena persistence, confirm that the hierarchical visual taxonomy is reproducible across the experimental conditions of this study. The moderate pairwise ARI (0.280 ± 0.134) reflects label permutation sensitivity inherent to stochastic dimensionality reduction rather than structural disagreement, a distinction addressed in the Discussion.

5. Discussion

The results presented in Section 4 demonstrate that the proposed CA-DINO pipeline discovers a hierarchical visual organization of marble varieties that is both internally coherent and structurally distinct from the commercial taxonomy. This section interprets these findings in the context of existing literature, examines the implications for evaluation methodology, discusses the feature priorities learned by the model, identifies limitations of the current study, and outlines directions for future work.

5.1. Learned Visual Structure Versus Commercial Classification

The three phenomena documented in Section 4.6, cross-category merging, intra-category splitting, and coherent pure family formation, collectively demonstrate that the relationship between commercial naming conventions and intrinsic visual structure is neither arbitrary nor fully aligned, but rather partially overlapping. The pipeline correctly recovers commercial boundaries where they correspond to genuine visual discontinuities, as evidenced by the coherent Bardiglio subtrees and the stable Arabescato Brown grouping across all six hierarchical levels. Simultaneously, it identifies commercially invisible relationships, such as the gray-veined family grouping Branco Peletigre, Dante Grey, Arabescato Brown, and Bardiglio at Level 2, and commercially concealed heterogeneity, such as the multiple Ruivina sub-populations separated by an inter-to-intra distance ratio of 257. These findings validate the central premise of this work: that supervised methods trained on commercial labels are structurally incapable of uncovering this partial misalignment, as they would be forced to either collapse the gray-veined family into separate categories or merge the distinct Ruivina sub-populations.

The novelty of this pipeline is therefore empirical rather than architectural: it demonstrates, through a rigorous three-tier validation design, that a specific synthesis of existing components reveals a hierarchical visual structure that is inaccessible to any individual component alone and irreproducible by supervised alternatives trained on the inconsistent commercial labels this work seeks to transcend.

This result extends the observations of Brondolo and Beaussant [17] and Scabini et al. [18], who demonstrated the effectiveness of DINO-based SSL in geological and material science domains, by showing that self-supervised representations are not only competitive with supervised approaches in classification accuracy but are qualitatively superior when the objective is to discover rather than replicate categorical structure.

5.2. The Role and Limitations of Evaluation Metrics

A central methodological insight of this study concerns the misalignment in this domain between standard extrinsic clustering metrics and the objective of unsupervised taxonomy discovery. As demonstrated in Sections 4.3 and 4.5, configurations achieving the highest ARI and NMI scores, most notably the raw 2048-dimensional embeddings (ARI = 0.302 ± 0.212, mean ± std across nine runs), did so as an artifact of high-dimensional noise inflation rather than genuine clustering quality, as confirmed by their poor and highly unstable internal geometry (SS = 0.372 ± 0.238, DB = 1.157 ± 0.632; Table 6). Conversely, CA-DINO at k = 10, the configuration producing the most perceptually coherent hierarchy, achieved a moderate mean ARI of 0.352 ± 0.144 across the same nine independent measurements (Table 8), precisely because it correctly grouped visually similar stones from different commercial categories rather than replicating the commercial label structure.

This observation has broader implications for unsupervised learning research in domains with unreliable labels. When the ground truth itself encodes the inconsistencies that the method aims to resolve, extrinsic metrics measure conformity to a flawed reference rather than discovery quality. We therefore advocate for a multi-criterion evaluation strategy combining internal geometric measures with systematic qualitative dendrogram analysis, an approach aligned with the broader recognition that hierarchical structure requires evaluation tools sensitive to multi-scale organization rather than flat partition agreement [36,37]. We note that this observation is derived from a single domain with a single label set; whether the same misalignment holds in other domains with convention-driven labels remains an empirical question, though the underlying logic that extrinsic metrics penalize correct visual groupings that deviate from flawed references is domain-general in principle.

5.3. Feature Priorities in the Learned Embedding Space

The level-by-level dendrogram analysis reveals a consistent hierarchy of visual feature priorities encoded by the CA-DINO representations. The root split at Level 1 separates warm-toned from cool-toned stones, indicating that base chromaticity constitutes the dominant axis of variation in the embedding space. Subsequent splits progressively resolve finer distinctions: veining density and pattern topology at Levels 2–3, and textural granularity and background saturation at Levels 4–6. This general-to-specific organization mirrors the perceptual strategy reported by domain experts who select marble primarily based on color family before attending to veining and textural details, suggesting that the self-supervised objective, combined with the cluster-aware loss, learns a feature hierarchy that is perceptually grounded without explicit supervision. Critically, this hierarchical feature organization was observed consistently across all nine independent measurements (3 data partitions × 3 training initializations), as confirmed by the taxonomy phenomena persistence analysis in Section 4.7 (Table 11), which recorded all three structural findings, namely cross-category merging, intra-category splitting, and coherent pure family formation, in 9/9 runs.

The superiority of UMAP-reduced embeddings over raw 2048-dimensional representations for downstream clustering (Section 4.3; mean SS = 0.372 ± 0.238 for raw vs. 0.693 ± 0.053 for the selected UMAP configuration, Table 6) further indicates that the discriminative features learned by the ViT backbone are encoded non-linearly, consistent with previous findings that Vision Transformers capture complex spatial relationships through multi-head self-attention. The failure of linear dimensionality reduction to preserve this structure underscores the importance of manifold-aware post-processing for ViT-derived embeddings in fine-grained visual domains.

5.4. Sensitivity to the Cluster Count Hyperparameter

The ablation across k values (Section 4.5) revealed that pipeline performance is sensitive to the alignment between the training cluster count and the dataset's intrinsic visual complexity. CA-DINO at k = 10 demonstrated competitive internal clustering geometry with the most comprehensive cross-split statistical evidence in this study ($N = 9$ measurements, 3 partitions × 3 initializations), while the sensitivity analysis revealed that performance differences among CA-DINO variants are substantially smaller than the gap between any CA-DINO configuration and the Pure DINO baseline; notably, k = 8, despite matching the number of commercial varieties represented in the test set, did not produce superior internal geometric cohesion relative to k = 10 under the full nine-run evaluation, suggesting that alignment between the clustering target k and the number of commercial labels is not a reliable proxy for geometric quality in this domain. This counterintuitive result indicates that the optimal training cluster count does not correspond to the number of commercial categories but rather to the number of visually distinguishable sub-populations in the data, a quantity that may exceed the commercial category count due to intra-variety heterogeneity. Future deployments of this pipeline should therefore treat k as a hyperparameter to be tuned against internal clustering quality on a held-out validation set, rather than setting it to the expected number of output categories. We acknowledge that k = 10 was selected on domain-alignment grounds, matching the number of commercial varieties, rather than through data-driven cluster-count estimation methods such as gap statistics, the elbow criterion, or stability-based approaches. While the sensitivity analysis across $k \in \{5, 8, 10, 12, 15\}$ demonstrates that the pipeline is not brittle to this choice and that all CA-DINO variants outperform Pure DINO, applying formal cluster-count estimation to the CA-DINO embedding space remains a direction for future work that could identify the intrinsic visual complexity of the dataset independently of the commercial category count.

5.5. Cross-Split Replication: Structural Stability Versus Label Stability

The Tier 2 results reveal a dissociation that deserves explicit attention: the pipeline produces highly stable hierarchical geometry, i.e., CCC = 0.954 across nine measurements with a coefficient of variation of 1.1%, and SS stable to within 5%, yet the pairwise Adjusted Rand Index between runs within the same data partition is low (overall mean 0.280). These two observations are not contradictory; they describe different levels of the clustering pipeline and must be interpreted separately.

The pairwise ARI measures agreement between flat partition label assignments, specifically, which of the ten cluster indices each sample receives. This assignment is determined by two stochastic components: the UMAP projection, which, despite a fixed random seed, introduces manifold perturbations when the input embedding distribution shifts between training runs, and the hierarchical merge order of average linkage, which is sensitive to small inter-point distance variations near cluster boundaries. When the embedding manifold shifts slightly between runs, a sample near the boundary between two visual clusters may cross the boundary, triggering a cascade of reattachments through average linkage's chain-sensitive merging behavior. The resulting partition differs substantially in label assignment even when the underlying visual groupings, which samples are nearest neighbors in the manifold, are preserved. The CCC of 0.954, which measures the fidelity of the dendrogram's pairwise distance structure independently of any flat cut, confirms precisely this: the hierarchical organization is preserved; what changes between runs is where the flat cut at k = 10 intersects the dendrogram, not the dendrogram itself.

This distinction is not merely a technical footnote. The phenomena persistence results provide the most direct evidence of structural stability: if the three key taxonomy findings (cross-category merging of gray-veined varieties, intra-category splitting of Exotic Ambar, and coherent pure family formation) are present in 9/9 independent measurements with different seeds and data partitions, then the underlying visual structure is real and recoverable, even if the precise label indices shift between runs. Cluster label indices are arbitrary identifiers; the visual groupings they represent are the substantive finding. A stability analysis that evaluates the former will systematically understate the reproducibility of the latter when the pipeline involves stochastic components.

This observation extends the metric misalignment argument of Section 5.2 from the between-method comparison to the cross-run stability context. Just as ARI penalizes the pipeline for correctly grouping Branco Peletigre and Dante Grey on the basis of their shared gray-veined appearance, pairwise ARI penalizes cross-run stability for correct but differently labeled recoveries of the same visual family. In both cases, the metric measures conformity to commercial labels in the between-method comparison, to a reference run's label assignment in the stability analysis, rather than the quality of the visual grouping itself. Internal metrics and phenomena persistence provide a more appropriate stability criterion for an unsupervised visual taxonomy whose output is a dendrogram and whose substantive claims are structural.

The moderate between-partition variation in external metrics, ARI is highest for Data split 1 (0.437 ± 0.090) and substantially lower for Data split 3 (0.267 ± 0.094), warrants acknowledgment. This between-partition spread is partly a consequence of the dataset's class imbalance: the three partitions stratify differently for rare varieties (Branco Carrara, n = 7; Irish Black, n = 10; Calacata, n = 22), and the precise composition of the test set determines how many boundary-region samples the flat k = 10 cut must classify, directly affecting the ARI. The SS and CCC are less sensitive to this effect because they assess the geometry of the full embedding rather than the agreement between partition labels and external references. Future work applying this pipeline to larger and more balanced

datasets would be expected to reduce between-partition ARI variation, as the stochastic sampling effects that drive it diminish with sample size.

In summary, the Tier 2 evidence establishes that the CA-DINO visual taxonomy is reproducible at the level that matters for the claims of this paper: the hierarchical dendrogram structure is stable (CCC = 0.954 ± 0.011), the geometric cluster quality is consistent (SS CV = 5.0%), and the three structural phenomena documented in Section 4.6 are universal across all nine independent measurements. Practitioners deploying this pipeline should be aware that flat partition label assignments are subject to permutation across runs and should evaluate taxonomic claims at the level of proximity structure and phenomena persistence rather than cluster index agreement.

5.6. Impact of Class Imbalance

The dataset exhibits substantial class imbalance, ranging from seven images for Branco Carrara to 472 for Ruivina (32% of the total). In a fully supervised pipeline, such disproportion would introduce systematic label bias, as over-represented classes would dominate the loss landscape. In the current unsupervised pipeline, no commercial labels are provided during training, thereby eliminating direct label bias. However, over-represented varieties may exert disproportionate influence on the learned feature manifold and on the k-means centroid positions generated during CA-DINO Stage 2, potentially biasing the pseudo-label distribution toward visual characteristics prevalent in high-frequency varieties.

The Dynamic Loss Gate mechanism provides partial mitigation: by fitting a two-component Gaussian Mixture Model to the per-sample loss distribution and down-weighting high-loss samples as unreliable, the gate implicitly reduces the influence of pseudo-labels that are likely to arise from under-represented varieties with poorly initialized centroids.

The qualitative analysis in Section 4.6 offers indirect evidence of robustness to imbalance. Rare varieties, namely Calacata (22 images), Irish Black (10), and Branco Carrara (7), form identifiable groups or join visually coherent clusters despite their low frequency, suggesting that the DINO backbone learned discriminative features for these varieties even without proportional representation in training. Nevertheless, systematic evaluation on class-balanced datasets is required to isolate imbalance effects from other sources of variability and is identified as a direction for future work.

5.7. Limitations

Several limitations of the current study should be acknowledged. First, the dataset comprises 1480 images in total, with 944 used for model training and 305 test images per partition across three independent data partitions, a scale sufficient to demonstrate the methodology but not yet representative of the full diversity of commercially available marble. Second, the evaluation relies on a single geological material; generalizability to other natural stones (granite, travertine, slate) or to broader material science domains remains to be empirically validated. Third, the pipeline currently operates on individual slab images and does not model spatial continuity across adjacent slabs from the same block, a property relevant to industrial matching applications. Fourth, while the qualitative dendrogram analysis provides strong evidence of hierarchical coherence, it is inherently subjective; developing quantitative metrics specifically designed for unsupervised taxonomy evaluation remains an open challenge. Fifth, the current evaluation is entirely computational; no industrial deployment, domain-expert user study, or business-oriented metrics (e.g., stock-matching efficiency, substitution accuracy) have been assessed. The three phenomena identified, namely the cross-category merging, intra-category splitting, and coherent pure family formation, are directly actionable for

specific industrial workflows: cross-category merging can inform automated stock matching by identifying visual substitutes across supplier catalogs; intra-category splitting supports visual quality stratification by distinguishing sub-grades within a single commercial label; and coherent pure family formation enables systematic catalog organization. Validating these applications in a real-world deployment setting, ideally with domain-expert participation, is a priority for future work. Sixth, the selection of epoch 200 as the training checkpoint was guided by three convergence criteria (Section 4.2), but principled early-stopping strategies for self-supervised ViT training on small, imbalanced datasets remain an open methodological question; future work should investigate automated convergence detection criteria suited to this regime.

The pipeline was evaluated exclusively with a ViT-S/8 backbone and UMAP for dimensionality reduction; comparison against alternative backbones (e.g., ResNet-50, DINOv2 ViT-B/14) and dimensionality reduction methods (e.g., PCA, t-SNE) would further isolate the contribution of each component. A simple baseline of ImageNet-pretrained features followed by UMAP and hierarchical clustering, without any domain-specific fine-tuning, would further clarify the value added by domain adaptation.

Regarding the scope of comparative evaluation, the experimental design of this study deliberately prioritizes depth of ablation over breadth of model comparison. Within the DINO family, we conducted a six-model comparison (Pure DINO plus CA-DINO at $k \in \{5, 8, 10, 12, 15\}$) with multi-run evaluation, and the primary configuration (CA-DINO $k = 10$) was further validated across nine independent measurements spanning three data partitions and three training initializations. This design isolates the specific contribution of the cluster-aware objective under controlled conditions, with identical backbone, identical augmentation regime, and identical post-processing pipeline, and it is the source of the central empirical claim of the paper: that cluster-aware training measurably improves embedding geometry relative to the unconstrained self-supervised baseline on this dataset.

A broader comparison against alternative self-supervised paradigms (SwAV, SimCLR with k-means, MoCo, DINOv2), alternative backbones (ResNet-50, ViT-B/14), and alternative dimensionality reduction methods (PCA, t-SNE, PaCMAP) would further characterize the relative performance envelope of the proposed pipeline. Such a comparison is outside the scope of the present study for two reasons. First, each alternative requires its own full training schedule, its own joint post-processing optimization on the validation set to ensure fair comparison, and its own multi-run statistical characterization—a combinatorial expansion of experimental cost that would multiply the present nine-measurement evaluation by the number of compared methods. Second, the contribution framed in Section 1 is not a claim of superiority over all available self-supervised methods but rather the demonstration that a specific, validated integration of established components, namely cluster-aware DINO, non-linear dimensionality reduction, and agglomerative hierarchical clustering, produces a reproducible hierarchical taxonomy in a domain where existing labels are structurally unsuitable as learning targets. The methodological contribution is the validated integration and the evaluation framework that exposes the misalignment between standard extrinsic metrics and unsupervised taxonomy objectives; cross-paradigm benchmarking is a natural follow-up but does not bear on the internal validity of the present claims.

The multi-split evaluation design, namely the nine independent measurements (3 data partitions × 3 training initializations) with mean and standard deviation reported in Table 8 provides a principled basis for assessing the reliability of the reported comparisons. The cross-split consistency analysis (Section 4.7, Table 9) confirms that the CA-DINO $k = 10$ advantage over Pure DINO is directionally consistent across all nine

measurements (SS CV = 5.0%, CCC CV = 1.1%), and the DB improvement ($\Delta = -0.183$) exceeds the standard deviation of both estimates, providing converging evidence that the observed differences reflect genuine representational gains rather than sampling variability.

5.8. Prerequisites for Cross-Domain Transferability

Regarding the staged validation and the role of single-domain evidence, the evaluation presented in this work is deliberately scoped to a single material (marble) acquired under controlled industrial conditions from a single equipment family. This scoping is not a concession but a methodological choice: before a pipeline can be meaningfully tested across materials, acquisition setups, and geographic contexts, its internal validity within a single well-characterized setting must first be established. The Tier 1 and Tier 2 evaluations demonstrate precisely this internal validity: the pipeline produces geometrically stable embeddings (SS CV = 5.0% across nine independent measurements), a reproducible hierarchical structure (CCC CV = 1.1%), and three structural phenomena that persist in nine of nine runs spanning three data partitions and three training initializations. These results establish that the taxonomy is a reproducible property of the data manifold under the tested conditions, which is the prerequisite for any subsequent claim of external generalization.

Claims of broader robustness (to alternative scanner types, uncontrolled lighting, variable viewing angles, different geological materials, and unseen commercial varieties) require dedicated datasets acquired under those varied conditions and are outside the evidentiary scope of the present dataset. We do not claim such generalization in this paper; we claim that the methodology, having been shown to work under rigorously characterized conditions on 1480 marble images, provides a validated foundation that future work can extend to those broader settings. The prerequisites listed below (conditions a–e) identify the conditions under which such extension is plausible and define a concrete research agenda.

The pipeline presented here was developed and evaluated on a single industrial dataset of marble slab images. While the methodology is not inherently marble-specific, its transferability to other domains is conditional on a set of prerequisites that should be verified before application.

The following conditions are expected to be necessary: (a) the target domain must involve imageable materials or objects with visually heterogeneous categories distinguishable at standard imaging resolution; (b) existing category labels must be assigned by convention, market tradition, or subjective consensus rather than by systematic physical or chemical measurement, i.e., domains where labels are physics-derived are less likely to benefit from purely visual unsupervised grouping; (c) sufficient within-class visual variability must exist to motivate unsupervised discovery of sub-structure beyond the existing label set; (d) a dataset of sufficient scale for self-supervised pretraining should be available, with a minimum on the order of 1000 images to provide adequate embedding diversity; and (e) the visual distinctions relevant to domain stakeholders must be capturable at the spatial scale and resolution of standard imaging.

Candidate domains where these conditions likely hold include other dimension stones (granite, travertine, slate, quartzite), ceramic and porcelain tiles, decorative wood grain, and textile fabrics. Domains where transferability is less certain include materials classified primarily by spectroscopic, mechanical, or crystallographic properties, where visual appearance is weakly correlated with the labeling criterion. Empirical cross-domain validation remains a direction for future work.

6. Conclusions

This work presented and validated an unsupervised pipeline for the hierarchical visual taxonomy of marble natural stone, combining cluster-aware self-supervised learning (CA-DINO) with UMAP dimensionality reduction and agglomerative hierarchical clustering using Average linkage. Through systematic ablation studies on a dataset of 1480 marble images spanning 10 commercial varieties, we demonstrated that each pipeline component contributes measurably to the quality of the final taxonomy: the cluster-aware training objective produces geometrically improved embeddings compared to pure self-supervised learning, as measured by internal clustering metrics across nine independent evaluations (3 data partitions × 3 training initializations; Table 8); UMAP 50D compression resolves the high-dimensional noise pathologies of raw 2048-dimensional embeddings while preserving the non-linear manifold structure essential for fine-grained visual discrimination; and Average linkage yields the most reproducible hierarchical geometry across independent runs, as evidenced by a Cophenetic Correlation Coefficient of 0.952 ± 0.015 (CV = 1.6%) in the nine-run evaluation (Table 7).

The resulting visual taxonomy exhibits three properties that distinguish it from commercial classification: it groups visually similar stones across commercial boundaries (cross-category merging), it separates visually distinct sub-populations within single commercial categories (intra-category splitting), and it preserves commercially meaningful groupings where visual and commercial boundaries genuinely coincide (coherent pure family formation). Crucially, these three phenomena were detected in all nine independent measurements of the Tier 2 cross-split replication study (Section 4.7, Table 11), establishing them as genuine geometric properties of the learned embedding space rather than artifacts of a particular data partition or training initialization.

From a methodological standpoint, we demonstrated that standard extrinsic clustering metrics (ARI, NMI) are misaligned with the objectives of unsupervised taxonomy in this domain when the reference labels encode the inconsistencies the method aims to resolve. This finding extends to the cross-run stability context: pairwise ARI between runs underestimates reproducibility by penalizing differently labeled recoveries of the same visual structure, while the CCC and phenomena persistence provide more appropriate stability criteria for a pipeline whose output is a dendrogram rather than a fixed partition. These insights apply beyond the marble domain to any unsupervised learning task in which existing categorical labels reflect convention rather than systematic visual criteria.

Future work will pursue three directions: (1) scaling the pipeline to larger and more diverse natural stone datasets, including granite and travertine, to assess cross-material generalizability; (2) integrating slab-level spatial continuity to support industrial block-matching applications; and (3) developing quantitative evaluation metrics specifically tailored to hierarchical taxonomy coherence in the absence of reliable ground truth. The pipeline's modular architecture provides a potentially transferable methodology for visual taxonomy problems in materials science, manufacturing, and other domains characterized by unreliable or convention-driven labeling systems, though cross-domain applicability remains to be empirically validated under the conditions outlined in Section 5.7.

Author Contributions: Conceptualization, M.F. and A.A.d.C.; methodology, M.F. and A.A.d.C.; software, M.F. and C.M.A.D.; validation, M.F. and A.A.d.C.; formal analysis, M.F.; investigation, M.F.; resources, G.P. and P.A.; data curation, C.M.A.D. and M.F.; writing—original draft preparation, M.F.; writing—review and editing, M.F., A.A.d.C., M.F., C.M.A.D., G.P. and P.A.; visualization, M.F. and A.A.d.C.; supervision, A.A.d.C.; project administration, P.A.; funding acquisition, G.P. and P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Sustainable Stone by Portugal project, proposal number C644943391-00000051, co-financed by the PRR—Recovery and Resilience Plan of the European Union (Next Generation EU).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author. The proprietary source images are not part of the original contributions and therefore are not shared publicly.

Acknowledgments: The authors, A.A.d.C., C.M.A.D., and G.P. gratefully acknowledge the support of the CERENA through FCT Project UID/04028/2025 (https://doi.org/10.54499/UID/PRR2/04028/2025, accessed on 1 February 2026). The author P.M. acknowledges Fundação para a Ciência e a Tecnologia (FCT) for its financial support via LAETA (project https://doi.org/10.54499/UID/50022/2025, accessed on 1 February 2026).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CA-DINO	Cluster-Aware Distillation with No Labels
DINO	Distillation with No Labels
ViT	Vision Transformer
SSL	Self-Supervised Learning
DEC	Deep Embedded Clustering
UMAP	Uniform Manifold Approximation and Projection
MLP	Multilayer Perceptron
GELU	Gaussian Error Linear Unit
EMA	Exponential Moving Average
DLG	Dynamic Loss Gate
GMM	Gaussian Mixture Model
KL	Kullback–Leibler
PCA	Principal Component Analysis
GLCM	Gray-Level Co-occurrence Matrix
LBP	Local Binary Patterns
CNN	Convolutional Neural Network
SS	Silhouette Score
DB	Davies–Bouldin Index
CH	Calinski–Harabasz Index
ARI	Adjusted Rand Index
NMI	Normalized Mutual Information
CCC	Cophenetic Correlation Coefficient
SSIM	Structural Similarity Index Measure
CS	Cosine Similarity

References

1. Pereira, D.; Marker, B. The Value of Original Natural Stone in the Context of Architectural Heritage. *Geosciences* **2016**, *6*, 13. https://doi.org/10.3390/geosciences6010013.
2. Navarro, R.; Pereira, D.; Gimeno, A.; del Barrio, S. Verde Macael: A Serpentinite Wrongly Referred to as a Marble. *Geosciences* **2013**, *3*, 102–113. https://doi.org/10.3390/geosciences3010102.
3. Muñoz-Cervera, M.C.; Rodriguez-Garcia, M.Á.; Cañaveras, J.C. Aesthetic Quality Properties of Carbonate Breccias Associated with Textural and Compositional Factors: Marrón Emperador Ornamental Stone (Upper Cretaceous, Southeast Spain). *Appl. Sci.* **2022**, *12*, 2566. https://doi.org/10.3390/app12052566.

4. Strzałkowski, P.; Köken, E.; Sousa, L. Guidelines for Natural Stone Products in Connection with European Standards. *Materials* **2023**, *16*, 6885. https://doi.org/10.3390/ma16216885.
5. Badouna, I.; Koutsovitis, P.; Karkalis, C.; Laskaridis, K.; Koukouzas, N.; Tyrologou, P.; Patronis, M.; Papatrechas, C.; Petrounias, P. Petrological and Geochemical Properties of Greek Carbonate Stones, Associated with Their Physico-Mechanical and Aesthetic Characteristics. *Minerals* **2020**, *10*, 507. https://doi.org/10.3390/min10060507.
6. Alper Selver, M.; Akay, O.; Alim, F.; Bardak, S.; Ölmez, M. An Automated Industrial Conveyor Belt System Using Image Processing and Hierarchical Clustering for Classifying Marble Slabs. *Robot. Comput. Integr. Manuf.* **2011**, *27*, 164–176. https://doi.org/10.1016/j.rcim.2010.07.004.
7. Elbehiery, H.; Hefnawy, A.; Elewa, M. Surface Defects Detection for Ceramic Tiles Using Image Processing and Morphological Techniques. *Int. J. Inf. Control Comput. Sci.* **2007**, *1*, 1488–1492.
8. Haileslassie, F.; Leta, A.; Desalegn, G.; Kalayu, M. Classification of Marble Using Image Processing. *Int. J. Data Sci. Technol.* **2019**, *5*, 57. https://doi.org/10.11648/j.ijdst.20190503.11.
9. Turan, E.; Ucar, F.; Dandil, B. A Novel Marble Recognition System Using Extreme Learning Machine with LBP and Histogram Features. *Concurr. Comput.* **2021**, *33*, e6438. https://doi.org/10.1002/cpe.6428.
10. Ouzounis, A.G.; Sidiropoulos, G.K.; Papakostas, G.A.; Sarafis, I.T.; Stamkos, A.; Solakis, G. Interpretable Deep Learning for Marble Tiles Sorting. In *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications, DeLTA 2021*; SciTePress: Setúbal, Portugal, 2021; pp. 101–108.
11. Canayaz, M.; Uludağ, F. Marble Classification Using Deep Neural Networks. *Eur. J. Tech.* **2020**, *10*, 52–63. https://doi.org/10.36222/ejt.671527.
12. Ouzounis, A.G.; Taxopoulos, G.; Papakostas, G.A.; Sarafis, I.T.; Stamkos, A.; Solakis, G. Marble Quality Assessment with Deep Learning Regression. In *Proceedings of the 5th International Conference on Intelligent Computing in Data Sciences, ICDS 2021*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2021.
13. Selver, M.A.; Akay, O.; Ardali, E.; Yavuz, B.A.; Önal, O.; Özden, G. Cascaded and Hierarchical Neural Networks for Classifying Surface Images of Marble Slabs. *IEEE Trans. Syst. Man. Cybern. Part C Appl. Rev.* **2009**, *39*, 426–439. https://doi.org/10.1109/TSMCC.2009.2013816.
14. Al-Zoubi1, H.R.; Al-Khassaweneh, M.A.; Altawil, I.A. An Image Processing Approach for Marble Classification. *Jordan J. Electr. Eng.* **2015**, *1*, 73–81.
15. Sipko, E.; Kravchenko, O.; Karapetyan, A.; Plakasova, Z.; Gladka, M. The System Recognizes Surface Defects Of Marble Slabs Based On Segmentation Methods. *Sci. J. Astana IT Univ.* **2020**, *1*, 50–59. https://doi.org/10.37943/aitu.2020.1.63643.
16. Sidiropoulos, G.K.; Ouzounis, A.G.; Papakostas, G.A.; Lampoglou, A.; Sarafis, I.T.; Stamkos, A.; Solakis, G. Hand-Crafted and Learned Feature Aggregation for Visual Marble Tiles Screening. *J. Imaging* **2022**, *8*, 191. https://doi.org/10.3390/jimaging8070191.
17. Brondolo, F.; Beaussant, S. DINOv2 Rocks Geological Image Analysis: Classification, Segmentation, and Interpretability. *arXiv* **2024**, arXiv:2407.18100.
18. Scabini, L.; Sacilotti, A.; Zielinski, K.M.; Ribas, L.C.; De Baets, B.; Bruno, O.M. A Comparative Survey of Vision Transformers for Feature Extraction in Texture Analysis. *J. Imaging* **2025**, *11*, 304. https://doi.org/10.3390/jimaging11090304.
19. Zhu, T.; Braytee, A.; Thiyagarajan, K.; Zi, X.; Mustapha, S.; Tao, X.; Prasad, M. Autonomous Detection of Concrete Cracks Using Self-Supervised DinoV2. *Mach. Intell. Res.* **2026**, *23*, 168–184. https://doi.org/10.1007/s11633-025-1553-5.
20. Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; Tao, D. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9052–9071.
21. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A Survey on Contrastive Self-Supervised Learning. *Technologies* **2021**, *9*, 2.
22. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 132–149.
23. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv* **2020**, arXiv:2002.05709.
24. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Piscataway, NJ, USA, 2020; pp. 9729–9738.
25. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. *arXiv* **2021**, arXiv:2104.14294.

26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the ICLR 2021 The Ninth International Conference on Learning Representations, Virtual, 3–7 May 2021.
27. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. https://doi.org/10.1109/TPAMI.2022.3152247.
28. Leiber, C.; Miklautz, L.; Plant, C.; Böhm, C. An Introductory Survey to Autoencoder-Based Deep Clustering—Sandboxes for Combining Clustering with Deep Learning. *arXiv* **2025**, arXiv:2504.02087.
29. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. *arXiv* **2016**, arXiv:1511.06335.
30. Wang, R.; Li, L.; Wang, P.; Tao, X.; Liu, P. Feature-Aware Unsupervised Learning with Joint Variational Attention and Automatic Clustering. In *Proceedings of the International Conference on Pattern Recognition*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; pp. 923–930.
31. Zhou, S.; Xu, H.; Zheng, Z.; Chen, J.; li, Z.; Bu, J.; Wu, J.; Wang, X.; Zhu, W.; Ester, M. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions. *arXiv* **2022**, arXiv:2206.07579.
32. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
33. Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J.T.; Peng, X. Contrastive Clustering. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 8547–8555.
34. Li, J.; Zhou, P.; Xiong, C.; Hoi, S.C.H. Prototypical Contrastive Learning Of Unsupervised Representations. In Proceedings of the ICLR 2021 The Ninth International Conference on Learning Representations, Virtual, 3–7 May 2021.
35. Han, B.; Chen, Z.; Qian, Y. Self-Supervised Learning with Cluster-Aware-DINO for High-Performance Robust Speaker Verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *32*, 529–541.
36. Naumov, S.; Yaroslavtsev, G.; Avdiukhin, D. Objective-Based Hierarchical Clustering of Deep Embedding Vectors. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 9055–9063.
37. Yang, J.; Parikh, D.; Batra, D. Joint Unsupervised Learning of Deep Representations and Image Clusters. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 5147–5156.
38. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data An Introduction to Cluster Analysis*; John Wiley& Sons, Inc.: Hoboken, NJ, USA, 1990.
39. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
40. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
41. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
42. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.
43. Caliński, T.; Harabasz, J. A Dendrite Method For Cluster Analysis. *Commun. Stat.* **1974**, *3*, 1–27. https://doi.org/10.1080/03610927408827101.
44. Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218.
45. Strehl, A.; Ghosh, J. Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
46. Rosenberg, A.; Hirschberg, J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*; Eisner, J., Ed.; Association for Computational Linguistics: Prague, Czech Republic, 2007; pp. 410–420.
47. Sokal, R.R.; Rohlf, F.J.; James, F.; Lawrence, R. The Comparison of Dendrograms by Objective Methods. *Taxon* **1962**, *11*, 33–40. https://doi.org/10.2307/1217208.
48. Sainburg, T.; McInnes, L.; Gentner, T.Q. Parametric Umap Embeddings for Representation and Semisupervised Learning. *Neural Comput.* **2021**, *33*, 2881–2907. https://doi.org/10.1162/neco_a_01434.

49. Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn.* **2011**, *12*, 2825–2830.
50. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.