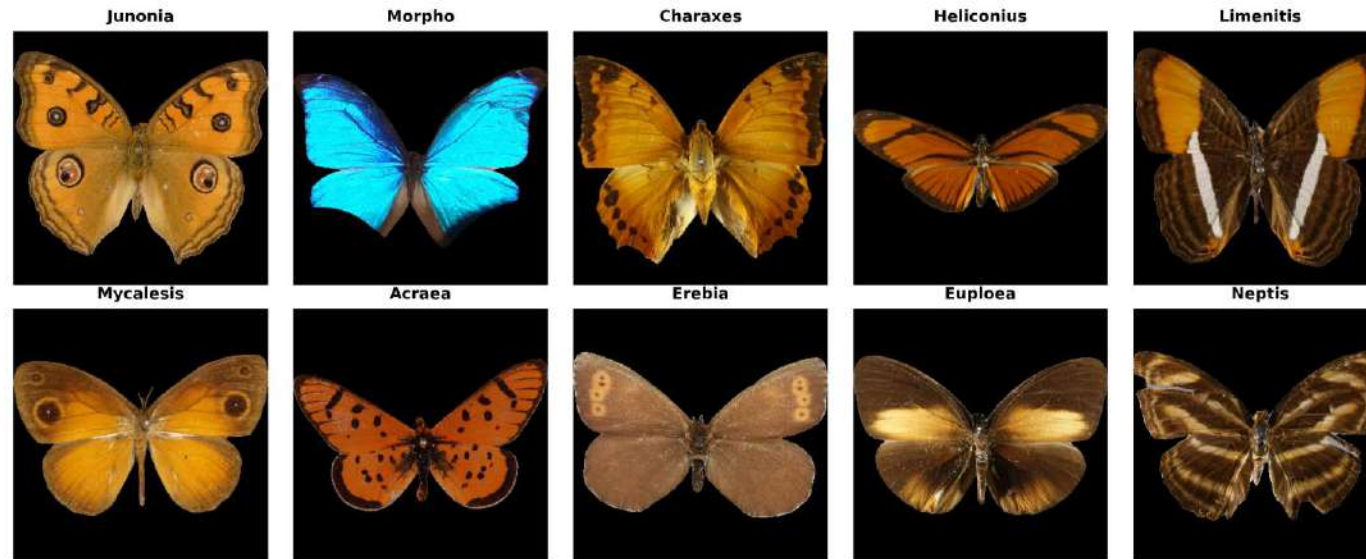


Self-Supervised Vision Embeddings Reveal Phenotype–Taxonomy Structure in Digitized Butterfly Collections

Arun Tipingiri 1 · Moritz Lürig 2 · Arthur Porto 2

1 Georgia Institute of Technology | 2 Florida Museum of Natural History, University of Florida



"Can self-supervised visual embeddings recover species structure—and more importantly reveal biologically meaningful departures?"

MOTIVATION & OPERATIONAL BOTTLENECK

Millions of Specimen Images, Limited Expert Review Capacity



Digitization Acceleration

Biological image archives are expanding far faster than manual expert curation capacity.



Repository noise

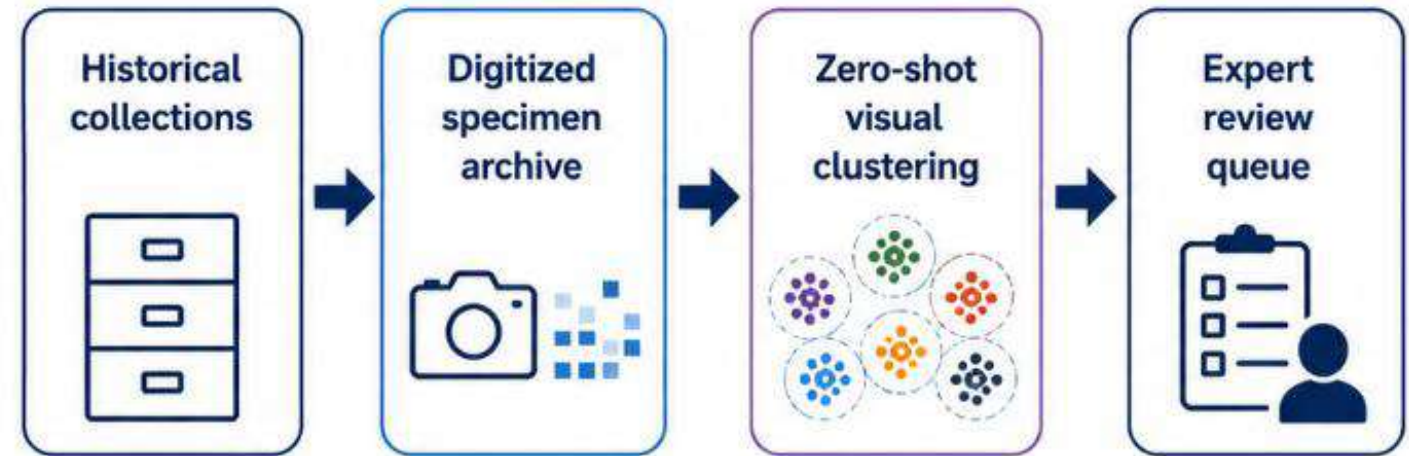
Historical collections contain labeling errors, taxonomic drift and unresolved population boundaries.



Supervised Trap

Supervised classifiers force images into predefined species labels, inheriting training label noise and failing on novel or unseen morphs.

Proposed Approach



Why this helps



Label validation



Collection audit



Retrieval



Review prioritization

Visual clustering offers a practical triage tool for large digitized natural history collections

CONCEPTUAL FRAMEWORK

Taxonomy-Free Visual Clustering: Emergence vs. Prediction

Phase 1: Pure Visual Structuring

Zero-Shot Representation:

- No taxonomic labels used during feature extraction.
- No labels used for dimensionality reduction (PCA/t-SNE/UMAP).
- No labels used for partition or density clustering.

Phase 2: Post-Hoc Taxonomic Audit

External Audit Signal:

- Ground-truth species names introduced *only afterward*.
- Intrinsic cluster coherence (Silhouette) Vs Extrinsic taxonomic alignment (ARI, V-Measure).
- Diagnoses biologically informative departures.

We do not train the model to predict taxonomy; we ask whether taxonomy naturally emerges from pure visual structure.

Why Butterflies?

A natural intersection of evolutionary biology and computer-vision difficulty

Natural history / Biological significance

- Wing patterns are dense morphological signals. it encodes mimicry, polymorphism, geography, seasonality, and divergence.
- Species labels and visual phenotypes are related but non-equivalent.
- Standardized specimen images retain provenance and taxonomic context.
- Mismatches can guide expert review and collection triage.



DIVERSITY • COMPLEXITY • AMBIGUITY

Computer vision significance

- Fine-grained complexity and Intra-species variation make recognition hard.
- Open-set structure exposes limits of fixed-label classification.
- Long-tailed sampling reflects real biodiversity archives.
- Enables clustering, retrieval, anomaly detection, and representation analysis.

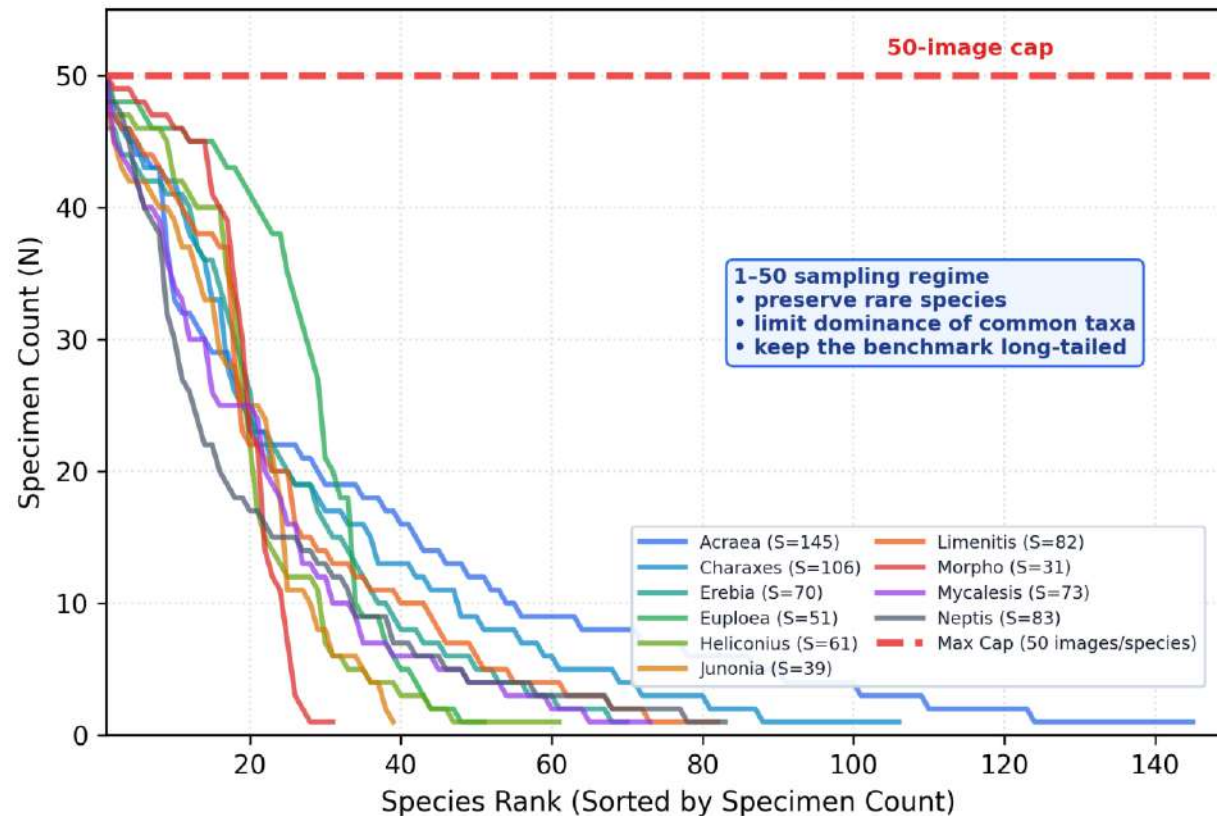
Butterflies are visually rich, biologically interpretable, and computationally challenging – an ideal testbed for taxonomy-free visual representations.

DATASET & BENCHMARK CONSTRUCTION

Species-capped 1–50 Sampling Preserves Long-Tailed Collection Structure

230,000+
SOURCE IMAGES189,541
DORSAL ARCHIVE11,821
BENCHMARK IMAGES741
RECOGNIZED SPECIES

Long-Tailed Species Support



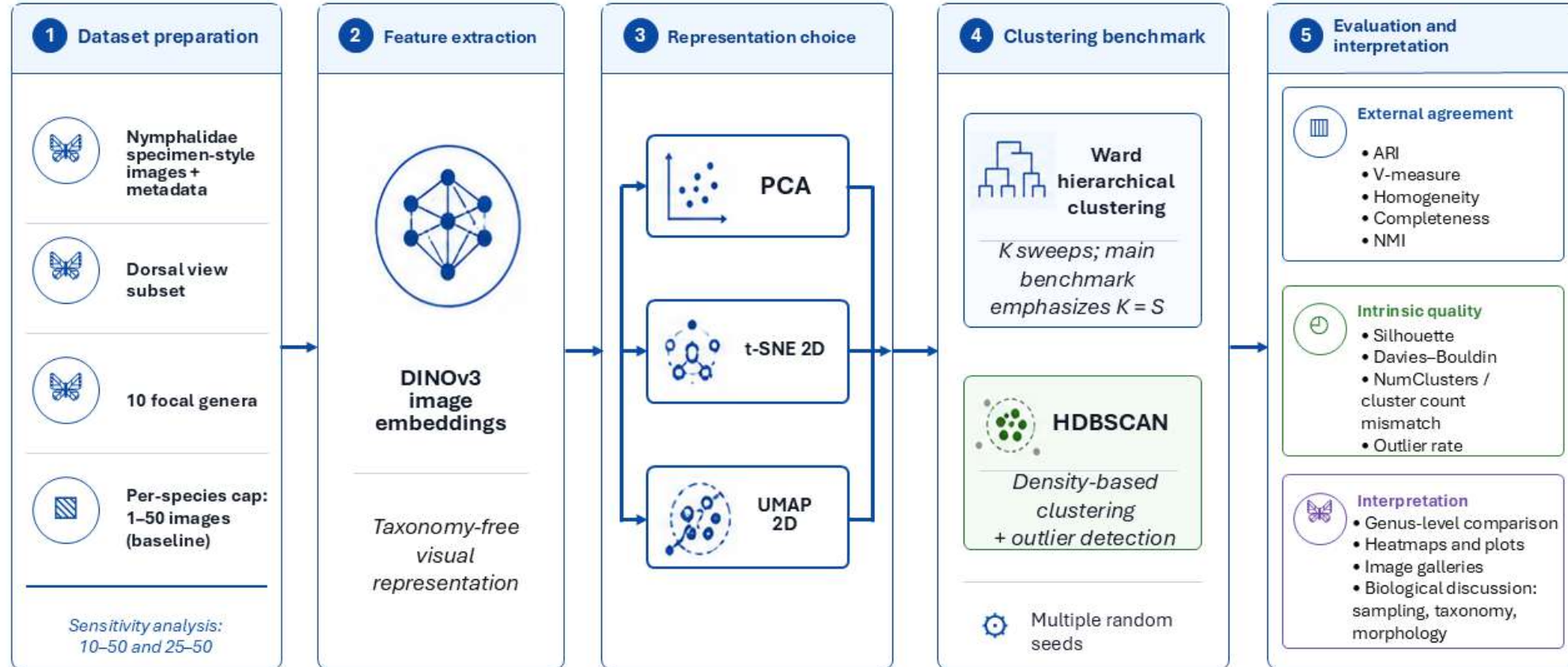
Genus	S	N	N/S
Junonia	39	919	23.56
Morpho	31	943	30.42
Charaxes	106	1,403	13.24
Heliconius	61	1,020	16.72
Limenitis	82	1,246	15.20
Mycalesis	73	1,030	14.11
Acraea	145	1,675	11.55
Erebia	70	1,206	17.23
Euploea	51	1,378	27.02
Neptis	83	1,001	12.06

1–50 sampling regime keeps rare species while preventing abundant species from dominating the evaluation.

METHODOLOGY & EVALUATION ARCHITECTURE

Zero-Shot Benchmarking Pipeline

Images → DINOv3 → PCA / t-SNE/UMAP → Ward / HDBSCAN → Evaluation



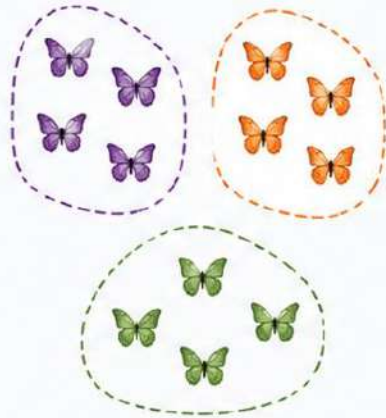
Goal: Assess whether DINOv3 visual embeddings recover species-consistent structure across genera, and diagnose when it does not.

Zero-Shot Benchmarking : Evaluation - Extrinsic & Intrinsic

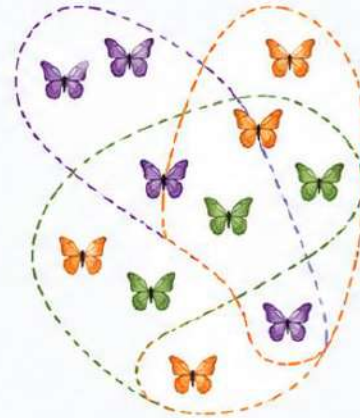


ARI: Taxonomic agreement

High agreement



Low agreement



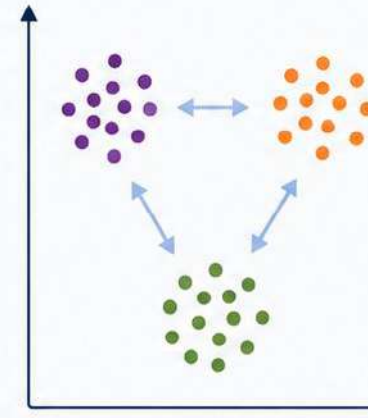
cluster-label agreement

higher is better

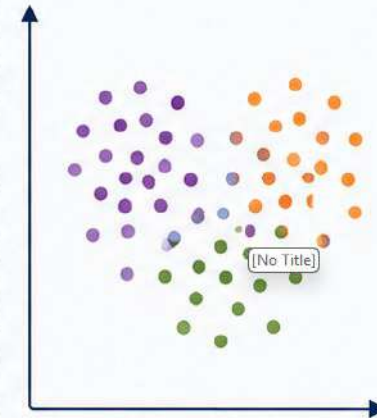


Silhouette: Geometric separation

Good (high silhouette)



Poor (low silhouette)



compact + separated

higher is better

Species labels (ground truth)



Species A



Species B



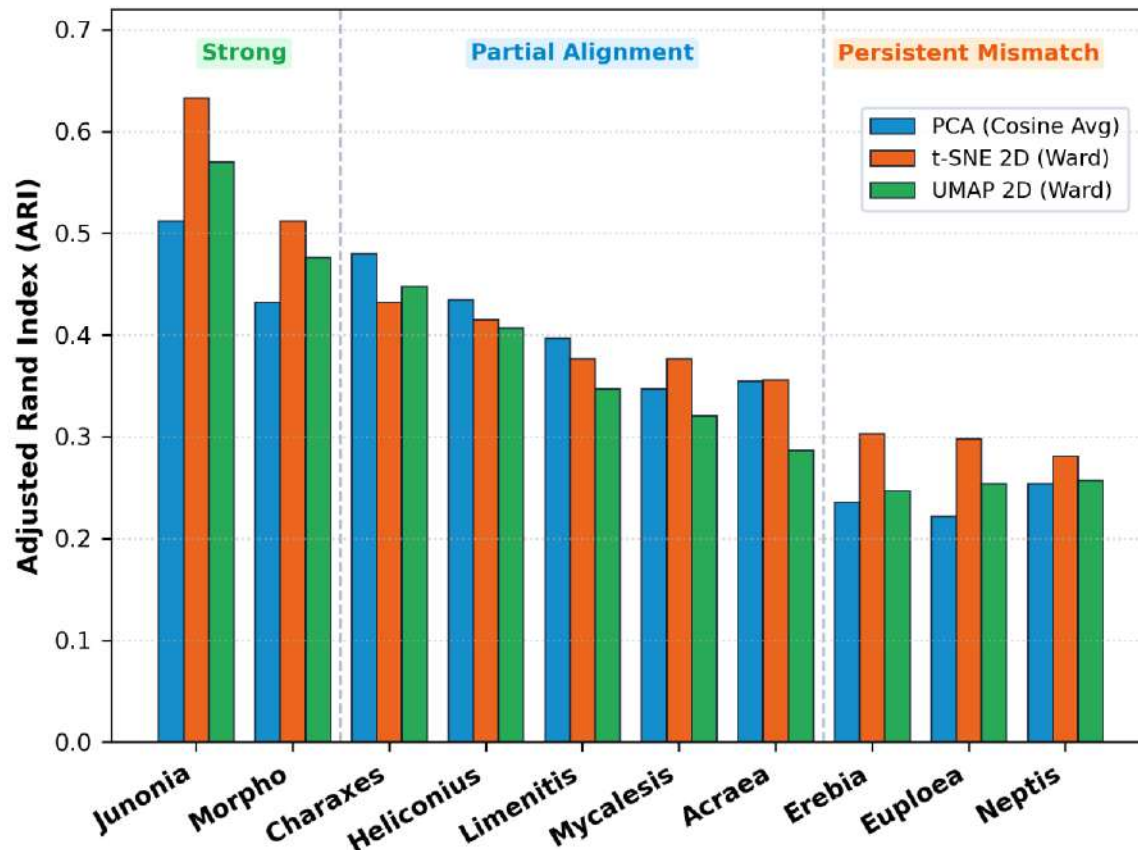
Species C

Predicted clusters



Genus Identity Matters More Than Representation Method

Best ARI by Genus Across Representation Spaces



Strong Alignment

Junonia & Morpho: Discrete, visually striking wing patterns align tightly with species boundaries.

Partial Alignment

Charaxes, Heliconius, Limenitis: Coherent visual clusters split or merge nominal species boundaries due to sexual dimorphism or mimicry.

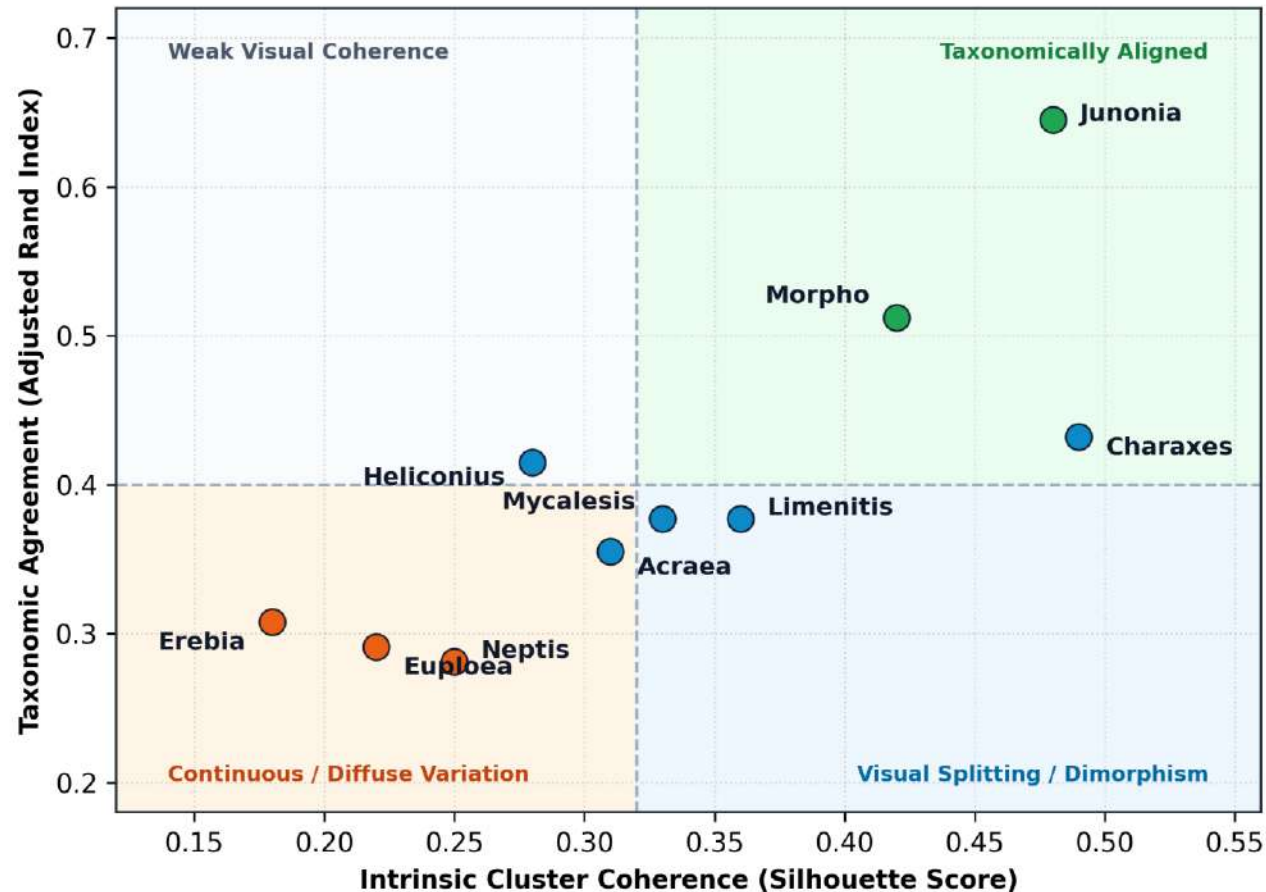
Persistent Mismatch

Erebia & Euploea: Continuous or convergent brown and black phenotypes cause persistent visual mixing across taxa.

Visual clustering performance is shaped by evolutionary biology and phenotypic structure.

A Visually Good Clustering Is Not Necessarily Taxonomically Correct

Intrinsic Quality vs. Taxonomic Agreement



Taxonomically Aligned

Junonia & Morpho - High ARI, High Silhouette: Visually discrete species boundaries. Visual geometry matches taxonomy.

Visual Splitting / Dimorphism

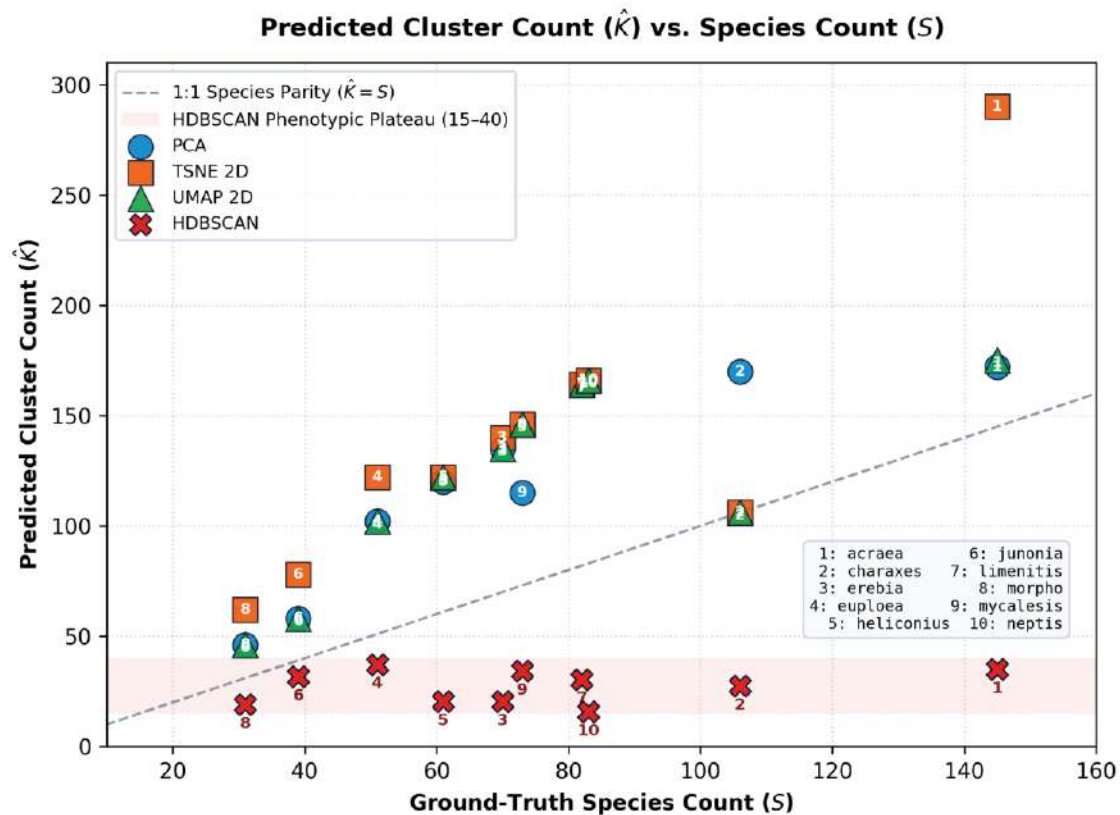
Charaxes, Heliconius, Limenitis - Moderate ARI, High Silhouette: Coherent visual groups that cut across species. Form dense visual clusters, but split nominal taxa due to sexual dimorphism.

Continuous / Diffuse Variation

Erebia, Euploea & Neptis - Low ARI, Low Silhouette: weak visual coherence, reflecting continuous phenotypic clines

High cluster coherence does not guarantee taxonomic agreement—unsupervised geometry isolates biological morphospace (e.g., dimorphism, mimicry) rather than nominal labels.

Dense Visual Archetypes Are Often Coarser Than Recognized Species



15 – 40 mode ceiling

HDBSCAN tends to recover a smaller envelope of dense visual modes, roughly 15 to 40 per genus

Phenotype Compression

In uniform or mimetic clades (*Neptis*, *Acraea*), multiple nominal species converge onto single shared dense archetypes.

Outlier Noise Isolation

HDBSCAN isolates 3.5%–16.7% of specimens as unclustered noise (rare morphs, damaged specimens, atypical lighting) rather than forcing them into artificial groups.

Density clustering recovers a smaller set of high-density visual modes, suggesting that dominant visual structure is often coarser than species-level taxonomy.

QUALITATIVE EMBEDDING ANALYSIS

Visual Embedding Landscapes Across Phenotypic Regimes

Strong Alignment**Junonia**

Discrete, well-separated visual neighborhoods track species boundaries clearly.

Partial Alignment**Charaxes**

Coherent neighborhoods form, but sexual dimorphism splits nominal species.

Weak / Diffuse**Erebia**

Diffuse visual cloud where subtle, continuous brown wing patterns resist clean partitioning.

Disagreement is not automatically a model failure—It reflects biological realities like polymorphism, Müllerian mimicry, sexual dimorphism, and historical curation artifacts that single-label taxonomies oversimplify.

Toward Collaborative Natural-History AI

What this benchmark enables

- A **taxonomy-free** indexing strategy to organize large archives by visual similarity.
- A **triage workflow** that isolates outlier noise, flags splits or merges, and detects labeling errors
- A **reproducible benchmark** for evaluating foundation-model representations in biological collections.

Future work / collaboration opportunities

- **Multimodal grounding:** Integrate specimen metadata, taxonomic literature, and phylogenetic context (spatial, temporal, and evolutionary signals).
- **High-dimensional density discovery:** Identify phenotypic modes directly in foundation-model space.
- **Cross-collection validation:** Test robustness across museums, taxa, imaging standards, and model families.

Open Reproducibility Artifacts Released: Metadata, dorsal images, run-level result tables, precomputed DINOv3 embeddings, and notebooks.

Looking to connect biodiversity collections • multimodal foundation models • graph learning • open-set evaluation