Corresponding author: Heru Handika, handika.heru@gmail.com

Running Header: An ML-assisted biodiversity web application

BioCosmos: computer vision and large language models for large-scale image-based biodiversity data and trait exploration

Heru Handika[1] https://orcid.org/0000-0002-2834-7175, Moritz D. Lürig[2] (https://orcid.org/0000-0002-8175-6234), Kira L. Nichtawitz[1](https://orcid.org/0009-0003-5379-094X), Oliver M. Dobon[3] (https://orcid.org/0009-0003-3837-8789), Venkata S. Yarava[1] (https://orcid.org/0009-0009-3025-4478), Justin K. Sobayo[1] (https://orcid.org/0009-0002-5620-8348), Breanna D. Shi[3] (https://orcid.org/0009-0002-5332-4626), Michael J. Elliott[4] (https://orcid.org/0000-0001-9243-3978), Jose A. B. Fortes[4] (https://orcid.org/0000-0001-8870-5205) , Arthur Porto[1,5] (https://orcid.org/0000-0002-9210-8750)

[1]Florida Museum of Natural History, University of Florida, Gainesville, Florida, USA
[2]Leibniz Institute for the Analysis of Biodiversity Change, Center for Taxonomy & Morphology, Bonn, Germany
[3]College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA
[4]Advanced Computing and Information Systems Laboratory, Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida, USA
[5]Department of Biology, University of Florida, Gainesville, Florida, USA

Keywords: biodiversity database, computer vision, vector embedding, large language model, lepidoptera

Abstract

Millions of natural history specimen images are now online, but poor discoverability keeps their rich phenotypic and evolutionary data out of reach for most users. We developed BioCosmos, a web platform that combines conventional biodiversity databases with emerging computer vision and large language models for image retrieval, visual feature recognition, and natural language queries. Its first implementation, LepiVerse (https://lepiverse.com), is a butterfly web portal with more than 600,000 specimen images aggregated from natural history museums worldwide via GBIF, SCAN, and Ecdysis, and curated through taxonomic harmonization and georeference validation pipelines. BioCosmos features visual similarity, interactive morphospace plots, and text- and image-based searches, alongside species pages enriched with data from GenBank, LepTraits, Wikipedia, and Crossref. The platform demonstrates how an open-source, low-cost emerging AI can make vast image collections accessible to a broad audience, while highlighting the digitization gap between the Global North and South.

Introduction

In recent decades, the massive digitization effort of Natural History Collections (NHCs) has enabled worldwide access to collection-based biodiversity data for downstream research, disease surveillance, conservation management, citizen science, and other applications (Willis et al. 2017, Nelson and Ellis 2018, Hedrick et al. 2020, Popov et al. 2021, Blackburn et al. 2024). Through increasing investments in data mobilization and aggregation by global biodiversity data infrastructures, such as the Global Biodiversity Information Facility (GBIF) and similar platforms, scientists can now query the data in bulk from multiple participating institutions via direct downloads or programmatically using an Application Programming Interface (API) (Constable et al. 2010, Nelson et al. 2012), expanding its usage for large-scale species distribution modeling and utilization for deep-learning (DL), among several others (Hedrick et al. 2020, Steinke et al. 2025). However, NHC data in general remain largely fragmented, with derived specimen datasets, such as genetic sequences and traits, hosted across heterogeneous databases or as publicly accessible spreadsheets (Gallagher et al. 2020, Soria et al. 2021, Shirey et al. 2022, Oskyrko et al. 2024, Sayol et al. 2024). While some tools ease data aggregation, retrieval, and taxonomic updates (e.g., Mozzherin et al. 2017, Elliott et al. 2026), consolidating disparate resources still requires users to spend hours on data cleaning, wrangling, and analytics to uncover evolutionary patterns, identify sampling and research gaps, or simply to choose research specimens to loan from NHCs. Millions of specimen images containing rich phenotypic traits are generally discoverable only through associated metadata (which typically does not include trait information) in web applications, and their use and analysis begin after interested individuals analyze them. There is a need for software platforms that mobilize, validate, and summarize these underutilized specimen images, and consolidate disparate records in one place, to streamline biodiversity research and its application while enhancing the value of digitized NHCs by making them accessible to a broad audience.

The increasing application of DL methods in large-scale phenotypic datasets has shown exciting promise for revealing complex phenotypic variation, evolutionary convergence, and morphological diversification (Hoyal Cuthill et al. 2019, 2024, East et al. 2025). Among these methods, DL-based vision models have been increasingly adopted to extract biologically meaningful features from high-dimensional, unstructured data, such as specimen images (Zhao et al. 2023, Lürig et al. 2024). These models convert the unstructured data into fixed-dimension numerical representations, called vector embeddings. These embeddings contain information about organism phenotypes by encoding visual characteristics present in the images, such as color, shape, and texture, in a high-dimensional latent space. Embedding models that are trained using approaches such as metric learning (see Lürig et al. 2021, 2024) can organize the latent space such that proximity reflects phenotypic similarity (Figure 1). Compared to conventional methods, such as manual scoring of phenotypes (e.g., Oliver et al. 2009), targeted spectrometry (e.g., Dalrymple et al. 2015), or geometric morphometrics (e.g., Chazot et al. 2016), which typically capture a predefined subset of trait features, a vector embedding represents a broader range of visual phenotypes across a specimen image (He et al. 2024). When integrated with

associated spatial, temporal, taxonomic, and phylogenetic information, the encoder model approach through metric learning could potentially uncover subtle evolutionary patterns that would otherwise remain hidden in NHCs (Zhao et al. 2023, He et al. 2024, Hoyal Cuthill et al. 2024).

With the growing availability of digitized specimen images in publicly accessible biodiversity databases, butterflies represent a promising model system for testing the potential of DL encoder models to uncover evolutionary patterns (Hoyal Cuthill et al. 2024, Lürig et al. 2025, Lawrence et al. 2026). Their mostly flat wing structure and standardized specimen preparation are convenient for extracting and comparing visual features from specimen images (Oliver et al. 2009, Teng and Zhang 2024). Additionally, their remarkable variability in coloration, patterning, and shapes of the wings has made them popular model systems to study evolutionary mechanisms, such as natural selection, mimicry, and morphological convergence (Turner 1976, Mallet 2001, Reed et al. 2011, Suzuki et al. 2019). Recent studies have demonstrated that DL models can quantify phenotypic similarities from specimen images and reveal evolutionary convergence that is difficult to characterize using conventional trait-quantification approaches (Hoyal Cuthill et al. 2019, Puissant et al. 2023, 2026). Beyond their scientific value, butterflies are widely recognized as an important group for advancing conservation efforts and citizen-science initiatives (Bonebrake et al. 2010, Sanderson et al. 2021, Piccini et al. 2022, Rosa and Freitas 2024). Both the scientific value and public interest in butterflies make them an appealing use case to develop a machine-learning-enabled web platform that serves both academic and public audiences. Through an integration of computer vision DL models combined with emerging machine learning techniques, such as multimodal encoder models and Large Language Models (LLMs), we can design a biodiversity data platform that provides conventional features familiar to taxon experts and image-based trait exploration accessible to both experts and the general public. It will lower barriers to data access, broaden public engagement with biodiversity science, help identify sampling gaps, and streamline the generation of new hypotheses from previously underutilized specimen images.

We developed BioCosmos to support this workflow. Its first implementation as a butterfly web portal, LepiVerse (https://lepiverse.com), combines more than 600,000 museum images of Lepidoptera with taxonomic and geographic records and linked genetic, trait, and literature resources. The platform supports searches based on images, natural-language descriptions, and conventional database fields, together with interactive exploration of image-based phenotypic variation. Here, we describe the data preparation, search workflows, and implementation choices that connect these functions. Our aim is to make digitized specimen images more useful for finding comparative material, exploring phenotypic similarities, and identifying questions for further study.

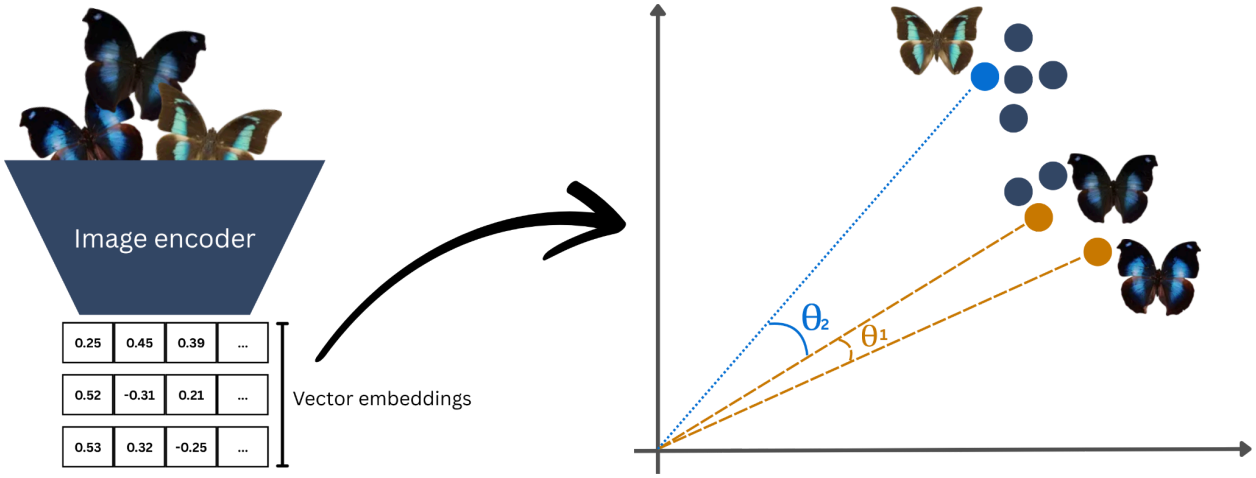


Figure 1. A simplified schematic illustrating encoding of specimen images into vector embeddings, the mapping of those embeddings into a latent vector space, and the cosine distance between them. The embedding numbers are for illustration only. They do not reflect the actual values from the images. The dots in the scatter chart represent clusters of reduced-dimensional embeddings. Images with similar coloration tend to have embeddings that are closer in vector space (θ_1), and vice versa (θ_2). Specimens shown from top to bottom of the chart: the Field Museum of Natural History (FMNH) INS125137, San Diego Natural History Museum (SDNHM) 150078, and Yale Peabody Museum (YPM) ENT412882.

Data Collection

We retrieved and compiled specimen images for seven Lepidoptera families (all butterflies) from biodiversity data aggregators: the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/), Ecdysis (https://ecdysis.org/), and SCAN (https://scan-all-bugs.org/) (https://doi.org/10.5281/zenodo.23840624). The download procedure and data collection followed (Lürig et al. 2025). Briefly, 1) we downloaded occurrence records of images from the data aggregators; 2) filtered for occurrences of preserved specimens with at least one associated image; 3) used the provided download links to retrieve the images from museum providers at their original resolution; 4) segmented the images with Grounded SAM (Ren et al. 2024); 5) removed junk with a custom trained YOLO classifier (Jocher and Qiu 2024); 6) computed MD5 hash values for cleaned images to remove duplicates; 7) renamed the images using the hash values; and 8) classified the images into dorsal or ventral sides with another custom trained YOLO classifier (Figure 2A; Lürig et al. 2025).

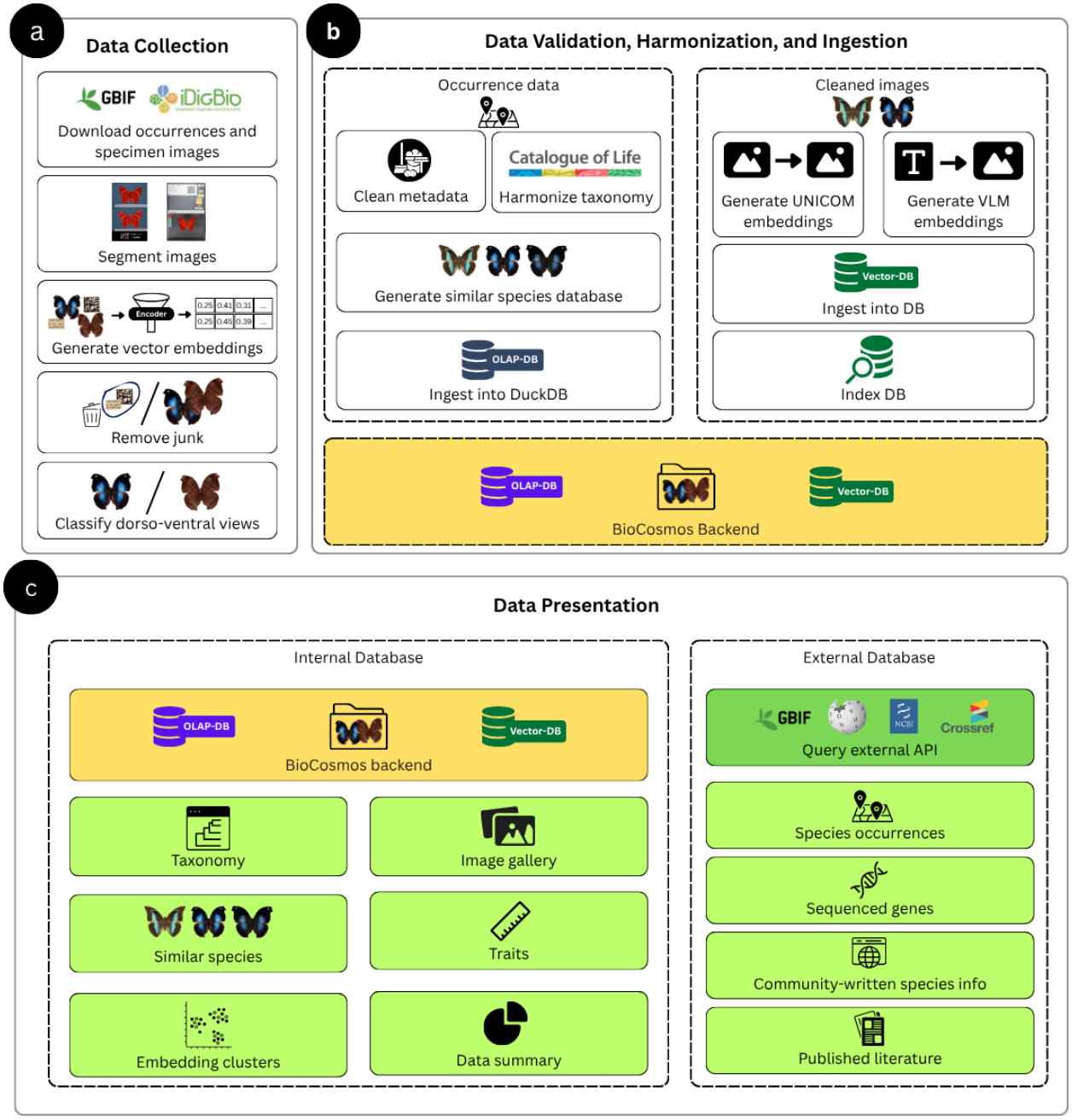


Figure 2. A schematic representation of a data preparation and presentation workflow from aggregation to visualization on LepiVerse: (a) data collection, (b) data harmonization and ingestion, (c) data presentation featuring key features provided by internal and external services. The workflows in panels A and B are ordered from top to bottom. Colors correspond to the module layers in Figure 5.

Data verification, validation, and harmonization

We developed an automated pipeline to validate and harmonize taxonomy, institution providers, and georeference records, implemented as packages within BioCosmos: *colharmonize,*

instharmonize, and *geoharmonize* packages, respectively. These packages were designed to update and validate hundreds of thousands of occurrence records on a personal computer and integrate directly with the application data management pipeline (Figure 2B). We tested all packages on a 14-inch MacBook Pro with an Apple M2 Pro processor and 16 GB of RAM. Our method for taxonomy validation and harmonization used the latest available taxonomy (version 2026-10-16 XR at the time of writing) from the Catalogue of Life (CoL; Bánki et al. 2026). We normalized the scientific names in both our records and CoL records. We used exact matching and string-distance approximations using Damerau-Levenshtein distance or Jaro-Winkler similarity. We matched the dataset names to CoL Name in three successive stages (species, subspecies, and genus), comparing canonical names with authorship removed. We retain up to five ranked candidates for each ambiguous match on the website for human review. To validate the georeferenced records, we downloaded the Database of Global Administrative Areas (GADM; https://gadm.org/) version 4.1 dataset (gadm_410-levels.gpkg) and validated the coordinates against the boundaries of GADM country and first-level administrative layer (ADM1; states and provinces). We ingested the validation categories into the database. For mismatches, we displayed the original coordinates and locality fields on the website for human review. To clean, update, and resolve specimen provider details, we used the Global Registry of Scientific Collections (GRSciColl; Schindel et al. 2016) and the GBIF dataset-publisher registry. We included each institution's website listed in GRSciColl or in the publisher record. Details on data verification and harmonization are available in Supplementary Method 1.

After validation and harmonization, our final dataset consisted of 619,739 images across seven Lepidoptera families: 38.4% had valid coordinates, 23.7% had missing coordinates, and the rest had invalid or mismatched coordinates at the state or country level (Figure 3a). Most records and taxa matched the CoL taxonomy exactly (Figure 3b and c). Across the dataset, 76.2% were dorsal images. Most images (92.8%) were retrieved from GBIF, with 53.2% of the total dataset contributed by both the Natural History Museum, London (NHMUK) and the Museum of Comparative Zoology, Harvard University, U.S. (MCZ). Nymphalidae was the most represented family in the dataset, accounting for 38.1% of all images, while *Lysandra coridon* was the most represented species, comprising 3.5% of the total. A large portion of the images were species distributed in the United States and Europe, while Brazil and Indonesia had the highest recorded species richness in the collection (Figure 4a–g).

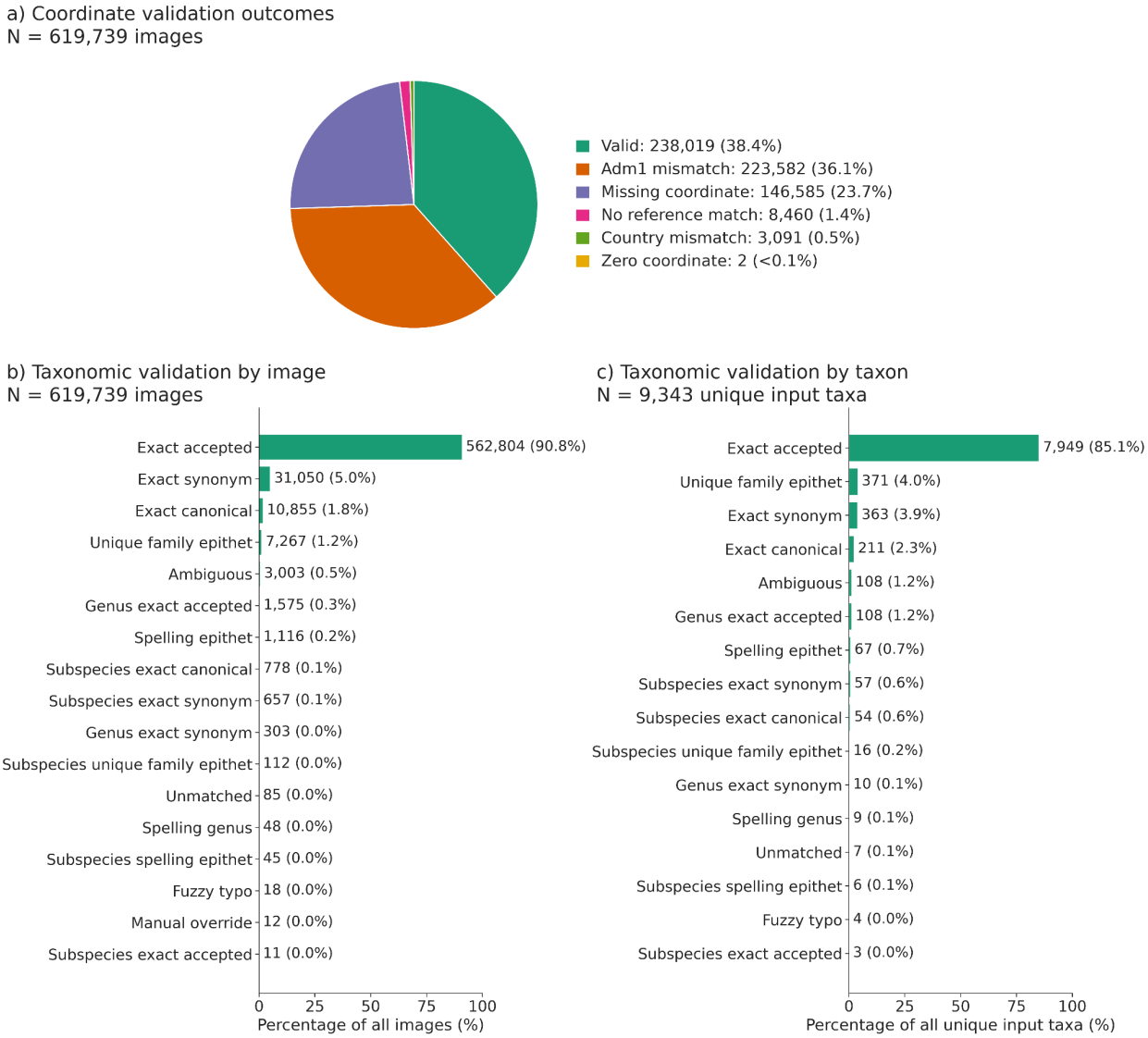


Figure 3. Outcomes of coordinate validation against the Database of Global Administrative Areas (GADM) and taxonomic validation against the Catalogue of Life Extended Release (CoL XR): (a) proportion of georeferenced records in each validation category; (b) proportion of records in each taxonomic validation category; and (c) proportion of taxa in each taxonomic validation category.

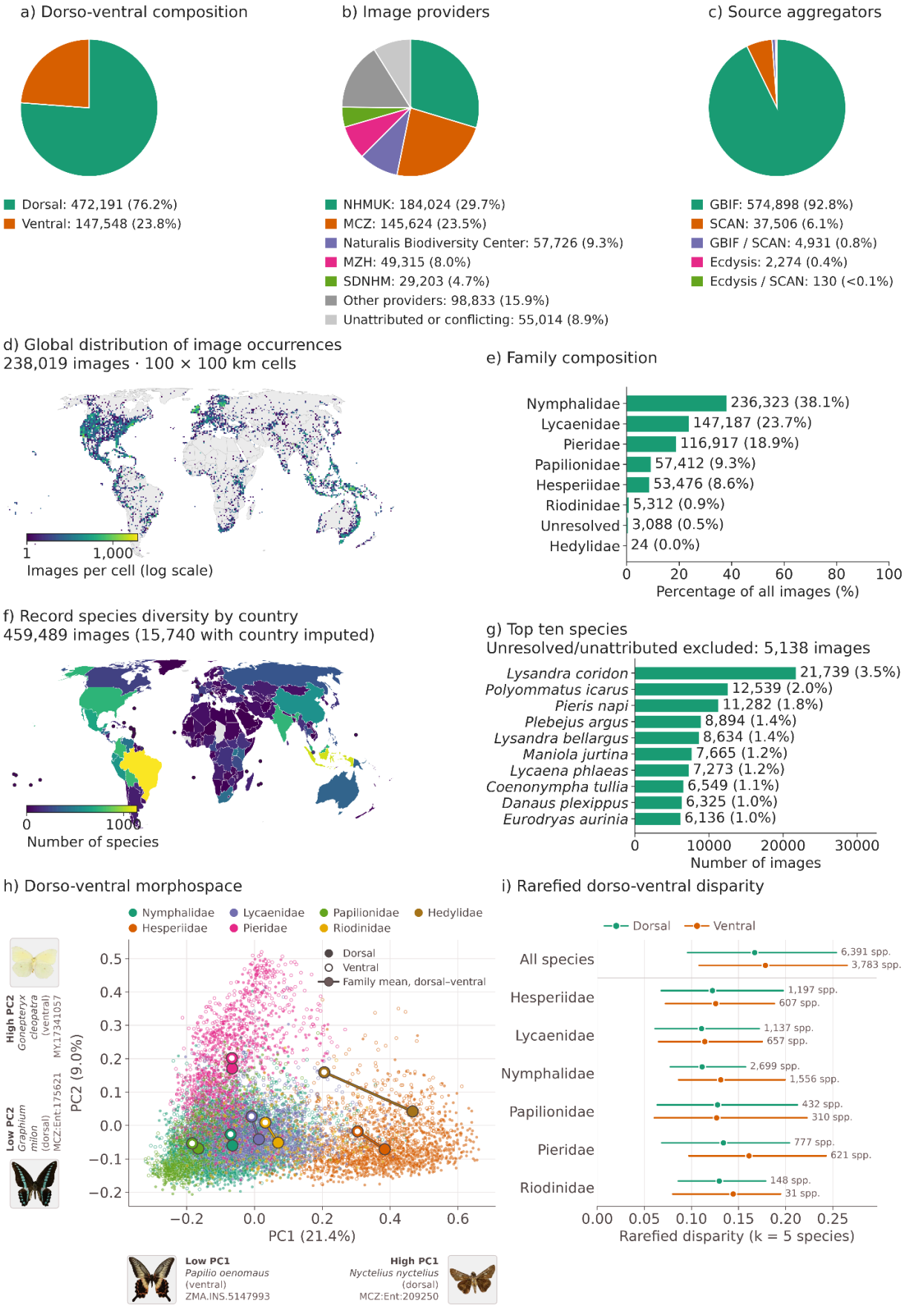

a) Dorso-ventral composition
Dorsal: 472,191 (76.2%)
Ventral: 147,548 (23.8%)
b) Image providers
NHMUK: 184,024 (29.7%)
MCZ: 145,624 (23.5%)
Naturalis Biodiversity Center: 57,726 (9.3%)
MZH: 49,315 (8.0%)
SDNHM: 29,203 (4.7%)
Other providers: 98,833 (15.9%)
Unattributed or conflicting: 55,014 (8.9%)
c) Source aggregators
GBIF: 574,898 (92.8%)
SCAN: 37,506 (6.1%)
GBIF / SCAN: 4,931 (0.8%)
Ecdysis: 2,274 (0.4%)
Ecdysis / SCAN: 130 (<0.1%)
d) Global distribution of image occurrences
238,019 images · 100 × 100 km cells
1
1,000
Images per cell (log scale)
e) Family composition
Nymphalidae 236,323 (38.1%)
Lycaenidae 147,187 (23.7%)
Pieridae 116,917 (18.9%)
Papilionidae 57,412 (9.3%)
Hesperiidae 53,476 (8.6%)
Riodinidae 5,312 (0.9%)
Unresolved 3,088 (0.5%)
Hedylidae 24 (0.0%)
0 20 40 60 80 100
Percentage of all images (%)
f) Record species diversity by country
459,489 images (15,740 with country imputed)
0
1000
Number of species
g) Top ten species
Unresolved/unattributed excluded: 5,138 images
Lysandra coridon 21,739 (3.5%)
Polyommatus icarus 12,539 (2.0%)
Pieris napi 11,282 (1.8%)
Plebejus argus 8,894 (1.4%)
Lysandra bellargus 8,634 (1.4%)
Maniola jurtina 7,665 (1.2%)
Lycaena phlaeas 7,273 (1.2%)
Coenonympha tullia 6,549 (1.1%)
Danaus plexippus 6,325 (1.0%)
Eurodryas aurinia 6,136 (1.0%)
0 10000 20000 30000
Number of images
h) Dorso-ventral morphospace
Nymphalidae Lycaenidae Papilionidae Hedylidae
Hesperiidae Pieridae Riodinidae
Dorsal
Ventral
Family mean, dorsal–ventral
High PC2
Low PC2
PC2 (9.0%)
PC1 (21.4%)
Low PC1
Papilio oenomaus
(ventral)
ZMA.INS.5147993
High PC1
Nyctelius nyctelius
(dorsal)
MCZ:Ent:209250
i) Rarefied dorso-ventral disparity
Dorsal
Ventral
All species 6,391 spp. 3,783 spp.
Hesperiidae 1,197 spp. 607 spp.
Lycaenidae 1,137 spp. 657 spp.
Nymphalidae 2,699 spp. 1,556 spp.
Papilionidae 432 spp. 310 spp.
Pieridae 777 spp. 621 spp.
Riodinidae 148 spp. 31 spp.
0.00 0.05 0.10 0.15 0.20 0.25
Rarefied disparity (k = 5 species)

Figure 4. Overview of the dataset showing: (a) Dorso-ventral composition; (b) Proportion of records by institution providers; (c) Composition of records by source aggregators; (d) Global distribution of the sampled images; (e) Record composition by family; (f) Record species richness by country; (g) Top ten species in the records; (h) PCA of dorsal and ventral species centroids in UNICOM embedding space (6,416 species, 3,758 with both sides); (i) rarefied disparity for each side, showing the mean and central 95% interval across 200 subsamples of five species. Unless specified, the total number of records is 619,739 images. Only records with valid family, species, coordinates, and countries were aggregated for family and species composition and global distribution. Missing countries were imputed using GADM country boundaries. The methods and abbreviations are described in the Data verification, validation, and harmonization section.

Data Presentation

Software design, architecture, and system integration

At a high level, the application consists of backend, frontend, and caching layers (Figure 5). The application also queries data via Application Programming Interfaces (APIs) from external services, such as GBIF, GenBank, Wikipedia, and Crossref, and hosts LepTraits (https://github.com/RiesLabGU/LepTraits; Shirey et al. 2022) in our database. For the backend, we used Python and the FastAPI framework (https://github.com/fastapi/fastapi) to build the application and *uv* (https://github.com/astral-sh/uv) to manage dependencies and the backend runtime. We used DuckDB (Raasveldt and Mühleisen 2019) for the analytical database and LanceDB (https://www.lancedb.com/) for vector embedding storage and operations. For the frontend, we used TypeScript and the Next.js framework (https://nextjs.org/), and Bun (https://bun.com/) to manage dependencies and the frontend runtime. Our frontend dependencies also support the Node.js (https://nodejs.org/) runtime and the Yarn package manager (https://yarnpkg.com/). For computationally intensive data processing that needed to run only once, we created separate packages in the "/packages" directory. These packages served as command-line tools or backend libraries, and Jupyter Notebook files in the "/analyses" directory performed data analyses for this manuscript. These packages also simplified updating records with new data or CoL taxonomy releases. A YAML configuration file in the backend provides settings for data ingestion, embedding model, and vector-index construction. The backend and analysis packages are members of one *uv* workspace, and the dependencies are resolved against a single lockfile (uv.lock) to ensure that every component uses identical versions. All backend, frontend, package, and analysis code described in this manuscript are maintained in a single repository codebase (https://github.com/BioVisionLab/BioCosmos).

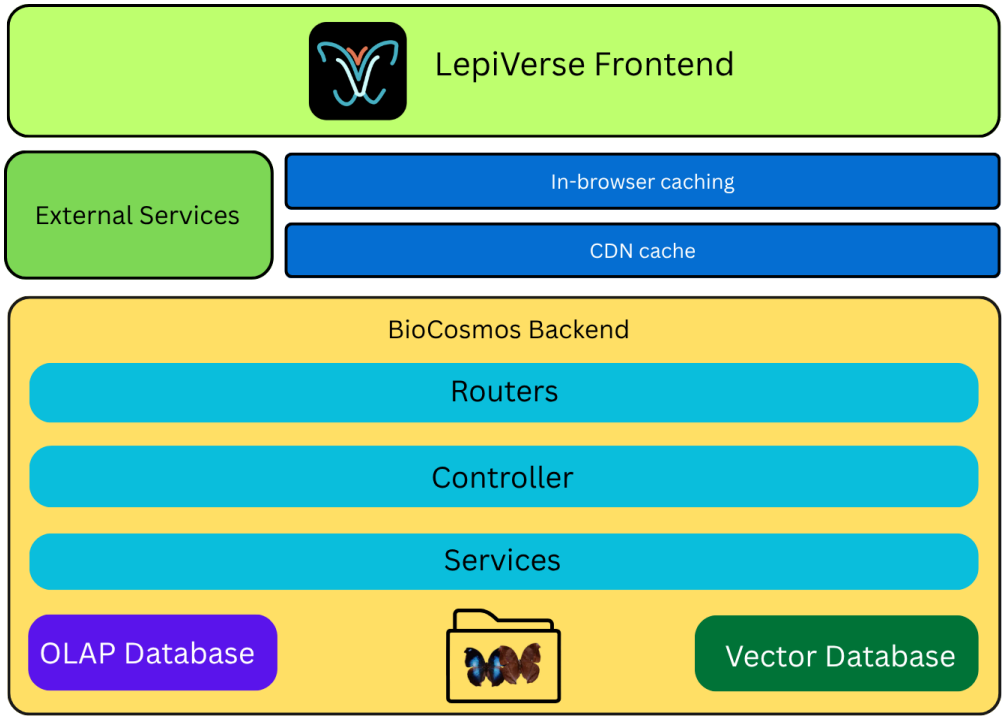


Figure 5. A high-level overview of the BioCosmos/LepiVerse modular architecture, illustrating each layer across the application stack. We used two layers of caching to minimize bandwidth to the backend. A specialized database for analytical workloads (OLAP; DuckDB) persists image occurrences, associated metadata, and taxonomy, and handles data shaping and analytics. An efficient vector database for vector storage and operations (LanceDB) persists image embeddings and handles vector operations. We stored high-resolution images and thumbnails as static files.

Machine learning integration, image standardization, and embedding generation

Our system integrates a Vision Model (VM) for image feature extraction, a Vision-Language Model (VLM) for image and text feature extraction, and a LLM for semantic text search orchestration and function calls. We used the Universal and Compact Representation model (UNICOM: ViT-L/14@336px; An et al. 2023) and the Contrastive Language-Image Pre-training model (CLIP: openai/clip-vit-base-patch32; Radford et al. 2021) as the VM and VLM, respectively. We first downscaled source images so their longest side was at most 512 px, preserving aspect ratio (smaller images were left unchanged). Each model's standard preprocessing then resized the shorter side and center-cropped the image to 336×336 px for UNICOM and 224×224 px for CLIP. Embeddings were L2-normalized. We ingested the embeddings, image identifiers (IDs), and processed image paths into LanceDB and stored associated specimen occurrence records in DuckDB, keyed by image IDs. To optimize storage, delivery, and visualization, we converted the images to the lossy WebP format at a maximum of 512 px on the longest side, with 128-px thumbnails. We then transferred these three datasets to our web backend for presentation on the website (Figure 2).

To provide optimum semantic search performance, we tested five open-weight LLMs (gemma-4-31b-it, mistral-small-3.1, gpt-oss-20b, meta-muse-glimmer-30b, and nemotron-3-nano-30b-a3b) using 18 queries across four categories: single intent, combination, ambiguous description, and restraint (Table 1). All the LLMs were hosted by the University of Florida NaviGator AI Toolkit and queried via the provided API using the OpenAI library (v3.16.2). We ran five replicates and measured accuracy, latency, and token consumption for 18 queries, totaling 450 runs across all tests, models, and replicates. We spaced each case run by 0.6 seconds to avoid hitting API rate limits. All tests were automated using the plannerbench v0.1.0 package in the BioCosmos codebase. Across all tests and replicates, gemma-4-31b-it outperformed the other models on all metrics, achieving 100% accuracy, the lowest median latency (0.44 s), and the lowest token consumption (about 1,890 tokens per successful call) in query planning (Figure 6a–c). We used it as the production model for the semantic search feature (see the search feature below).

To provide initial biological validation of embedding space beyond species classification, we tested whether it recovers published butterfly mimicry. We compiled 14 Müllerian and Batesian pairs from the literature, comprising 18 species, for which both had dorsal images in the collection (Supplementary Table 1). Three pairs of *Heliconius numata–Melinaea* spp. were not tested because the collection has no dorsal images for *H. numata*. For each tested species, we computed the centroid of its dorsal UNICOM embeddings, scored each of the other 8,350 species by the cosine distance from this centroid to that species' nearest dorsal image, and ranked the ten most similar species. We compared the mean partner percentile with three null models, drawing random species from the whole collection, from the query species' genus, or from the partner's genus (10,000 permutations each). The observed mean percentile (0.997) exceeded all three null means (0.50, 0.69, and 0.92; $p = 0.0001$). All pairs, except *Papilio polyxenes–B. philenor*, ranked among the top ten most similar species in at least one direction. Six pairs, including *Heliconius melpomene–H. erato* and *Limenitis archippus–Danaus plexippus*, appeared on the top-ten list for both directions (Figure 6d).

Morphospace analyses

We analyzed morphospace, morphological disparity, and correlations between dorso-ventral visual features for all records with valid taxonomy. We used principal component analysis of normalized UNICOM image embeddings to describe the morphospace of taxon visual features. We retained the first three principal components and ingested the coordinates and variance proportions into the DuckDB database for display on the website. We measured morphological disparity for each taxon set, separately for dorsal and ventral coloration, as the sum of variances of the species centroids. For groups with at least five species, we rarefied disparity to five species to account for differences in sample sizes. To measure the correlation between dorsal and ventral coloration, we ran a Mantel test for each family and genus with at least five species that had centroids for both sides. We computed pairwise cosine distances between species centroids separately for the dorsal and the ventral side and assessed significance with 999 permutations.

We ingested the results of each analysis into the application's DuckDB database (see Supplementary Method 2 for details).

The morphospace results are displayed in LepiVerse on species, genus, and family pages, whereas the morphological disparity and dorso-ventral correlation results are only displayed on the family and genus pages of the website. Figure 4h shows variation in morphospaces across all families in the collection, revealing substantial structure of visual features across sampled families. Hesperiidae, Pieridae, and Papilionidae occupied relatively distinct morphospace, whereas Nymphalidae, Lycaenidae, and Riodinidae showed substantial overlap. Figure 4i compares the rarefied morphological disparity between dorsal and ventral sides of six of the seven families in the collection. We excluded Hedylidae because it has fewer than five species with centroids. Ventral disparity exceeds dorsal disparity in every family, except Papilionidae, although the 95% intervals overlap overall in every family. The dorsal-ventral correlation is available on the website's family and genus pages with at least five species images on both sides.

Table 1. Semantic search categories, case identifiers, and queries used to test the performance of open-weight Large Language Models (LLMs).

No	Categories	Case IDs	Queries
1	Single intent	color-only	orange butterflies with black stripes
2		country-only	butterflies of Indonesia
3		adm1-ambiguous-name	butterflies in Georgia USA
4		similar-scientific	similar to Danaus plexippus
5		common-name	show me monarch
6		similar-common-name	monarch look-alikes
7		similar-common-name-morpho	species that resemble the blue morpho
8		trait-only	butterflies that live in the forest canopy
9	Combination	color-country	blue from brazil
10		common-name-country	monarch in Brazil
11		color-country-yellow	yellow butterflies from Indonesia
12		similar-country	butterflies like Morpho menelaus from Peru
13		color-trait	spotted butterflies in disturbed habitats
14		color-country-trait	orange canopy butterflies in Costa Rica
15	Ambiguous description	descriptive	owl-like butterfly
16		descriptive-country	owl-like butterfly from brazil

17		descriptive-glasswing	glasswing butterfly
18	Restraint	ignore-generic-terms	show me red butterfly species

Search Features

BioCosmos provides three search options: conventional text, image, and semantic text searches. The conventional text search accepts standard text with options to limit searches by database fields, such as taxonomy, traits, or coordinate points within a fixed 50km radius. It queries the database directly and returns species pages and image occurrences that match the query. The image search accepts JPEG, PNG, or WEBP input. It uses the UNICOM model to generate embeddings and find the closest cosine distance to the UNICOM specimen embeddings in the vector database. Semantic search allows natural language queries and uses gemma-4-31b-it to parse the user query into one or more predefined search functions and their parameters (Table 2; Figure 7). The model receives only the query and prompt instructions to match with provided functions. It does not see or interpret the search results. This design keeps a human in the loop for evaluating the biological relevance of the results.

The function uses CLIP for color-based and UNICOM for image-similarity search, respectively. For instance, a user searching for “butterflies like Morpho menelaus from Peru” will be parsed by the LLM into an image-similarity search for butterfly species "Morpho menelaus" and a geographic filter for "Peru". Other queries, such as “monarch look-alikes”, are parsed into an image-similarity search for the common name “monarch”, and the UNICOM embeddings of that species are queried for the closest cosine distance. Location, trait, and common-name functions first restrict the candidate species set, and the CLIP or UNICOM searches rank only within that set. Because CLIP text-to-image and UNICOM image-to-image distances lie on different scales, we rescale each search's cosine distances to 0.5–1 within its candidate pool and average them across searches that retrieve a species. If no embedding search is used, we list matching species alphabetically. Trait searches retrieve information from the existing LepTraits fields rather than infer ecological traits from specimen images.

To minimize retrieval latency, we indexed the image embeddings. We evaluated different vector indexing algorithms by search latency and recall@10, defined as:

$$Recall@10 = \frac{|R_{10} \cap G|}{|G|}$$

where R_{10} is a set of top 10 retrieved images, and G is the ground truth obtained by computing all pairwise cosine distances directly (without running optimized searches on an index). We tested three vector indexing algorithms available in LanceDB v0.39.0 against embeddings generated by CLIP and UNICOM. The IVF-HNSW-SQ index reached the best trade-off between recall and latency for both models. At the query settings used on BioCosmos (80 probes, refine factor 2), it reached recall@10 of 1.000 (UNICOM) and 0.999 (CLIP) at a mean of 4.1 ms and 3.9 ms per

query, compared with 633 ms and 378 ms for exact search (Figure 6e). We used the algorithm to index the final LanceDB dataset (see Supplementary Method 3 for details).

Table 2. Semantic text query functions and expected input parameters provided to the Large Language Model. The model selects the functions that match the query and supplies their parameters.

Functions	Input parameters
search_by_common_name	Common name (string, max 200 chars)
search_by_image_similarity	Reference species (string, max 128 chars; a scientific name, common name, or genus)
search_by_location	Country (ISO 3166-1 alpha-2 code, required); state/province (string, 1-100 chars, optional)
search_by_color	Color description (string, max 200 chars)
search_by_traits	At least one of LepTraits fields: canopy (closed/open/mixed/generalist/edge); diapause_stage (egg/larva/pupa/adult); disturbance (tolerant/avoidant); edge (associated/avoidant); flight_months (list of jan-dec); host_breadth (specialist/generalist); hostplant_family (plant family, letters only, up to 64 chars); moisture (wet/dry); oviposition (single/clustered); voltinism (univoltine/bivoltine/multivoltine); wing_size (small/medium/large)

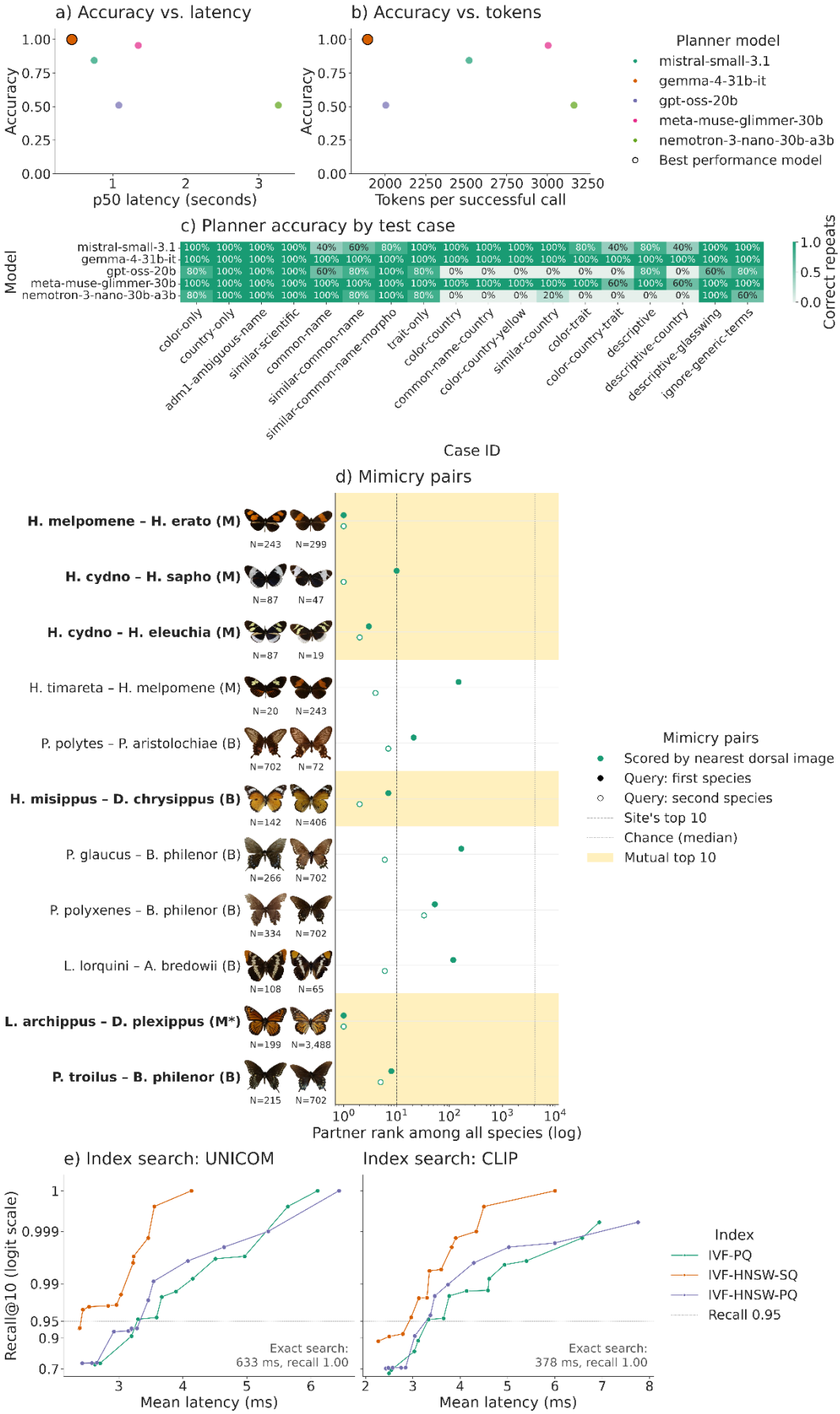

a) Accuracy vs. latency
b) Accuracy vs. tokens
Accuracy
p50 latency (seconds)
Tokens per successful call
Planner model
mistral-small-3.1
gemma-4-31b-it
gpt-oss-20b
meta-muse-glimmer-30b
nemotron-3-nano-30b-a3b
Best performance model
c) Planner accuracy by test case
Model
mistral-small-3.1 100% 100% 100% 100% 40% 60% 80% 100% 100% 100% 100% 100% 80% 40% 80% 40% 100% 100%
gemma-4-31b-it 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100%
gpt-oss-20b 80% 100% 100% 100% 60% 80% 100% 80% 0% 0% 0% 0% 0% 0% 80% 0% 60% 80%
meta-muse-glimmer-30b 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 100% 60% 100% 60% 100% 100%
nemotron-3-nano-30b-a3b 80% 100% 100% 100% 100% 80% 100% 80% 0% 0% 0% 20% 0% 0% 0% 0% 100% 60%
color-only
country-only
adm1-ambiguous-name
similar-scientific
common-name
similar-common-name
similar-common-name-morpho
trait-only
color-country
common-name-country
color-country-yellow
similar-country
color-trait
color-country-trait
descriptive
descriptive-country
descriptive-glasswing
ignore-generic-terms
Case ID
Correct repeats
d) Mimicry pairs
H. melpomene – H. erato (M)
N=243 N=299
H. cydno – H. sapho (M)
N=87 N=47
H. cydno – H. eleuchia (M)
N=87 N=19
H. timareta – H. melpomene (M)
N=20 N=243
P. polytes – P. aristolochiae (B)
N=702 N=72
H. misippus – D. chrysippus (B)
N=142 N=406
P. glaucus – B. philenor (B)
N=266 N=702
P. polyxenes – B. philenor (B)
N=334 N=702
L. lorquini – A. bredowii (B)
N=108 N=65
L. archippus – D. plexippus (M*)
N=199 N=3,488
P. troilus – B. philenor (B)
N=215 N=702
Partner rank among all species (log)
Mimicry pairs
Scored by nearest dorsal image
Query: first species
Query: second species
Site's top 10
Chance (median)
Mutual top 10
e) Index search: UNICOM
Index search: CLIP
Recall@10 (logit scale)
Exact search:
633 ms, recall 1.00
Exact search:
378 ms, recall 1.00
Mean latency (ms)
Index
IVF-PQ
IVF-HNSW-SQ
IVF-HNSW-PQ
Recall 0.95

Figure 6. Performance tests and validation for machine learning features: (a) a scatter plot of open-weight model accuracy versus latency at the 50th percentile, and (b) versus token usage per successful call. (c) percentage of accuracy by test case across five runs. The query for each case ID is defined in Table 1. (d) Partner rank of known mimicry pairs from the literature among UNICOM visually similar species. The dashed vertical line marks the top ten and the dotted line the chance median; pairs with mutual top-ten results are shaded and in bold. Asterisks indicate context dependent mimicry type depending on local ecology. (e) Recall@10 against mean query latency (ms) for each index type, in separate panels for UNICOM and CLIP embeddings. Each line is an index's frontier across query-time settings, and exact-search latency is given as a note.

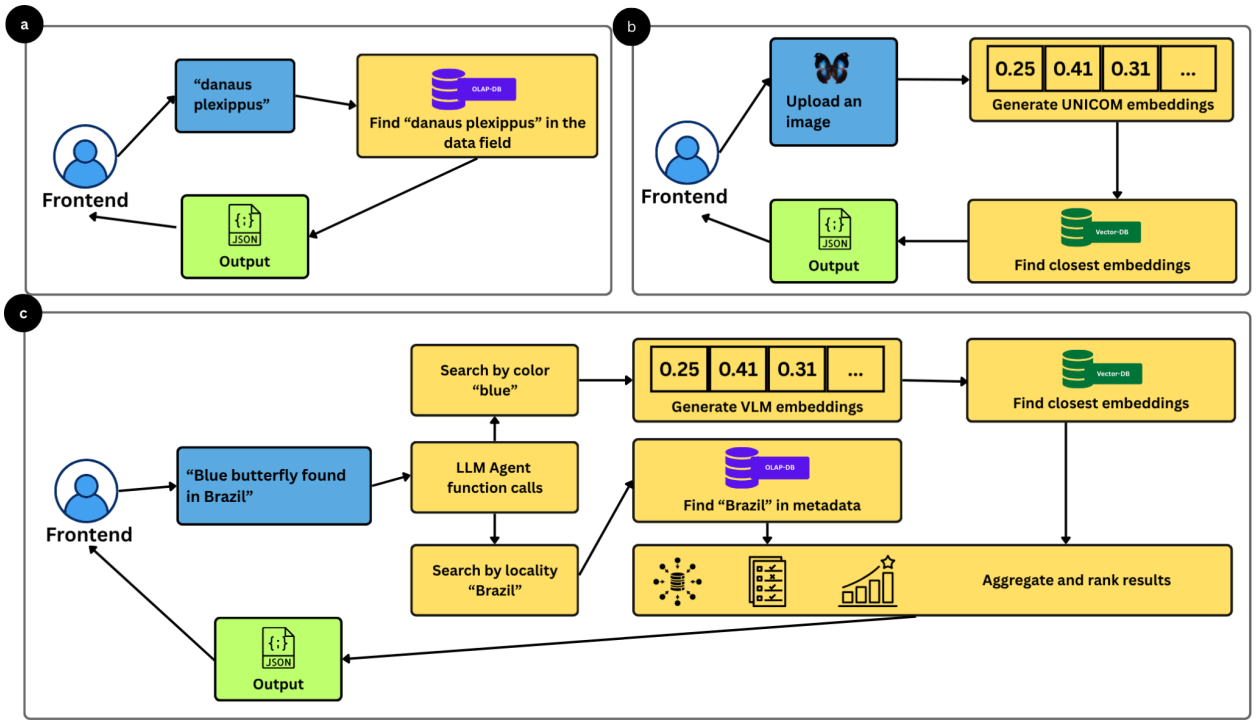


Figure 7. Simplified schemas of three search features available at BioCosmos highlighting data flows from frontend request (azure) to backend processing (yellow) and output production (lime): (a) conventional database search, (b) image search utilizing UNICOM embeddings, and (c) semantic search utilizing a large language model (LLM) for query parsing and function calls. Semantic searches using vector embeddings are ranked by cosine distance, rescaled within each search's candidate pool. The frontend parses the output JSON for visualization.

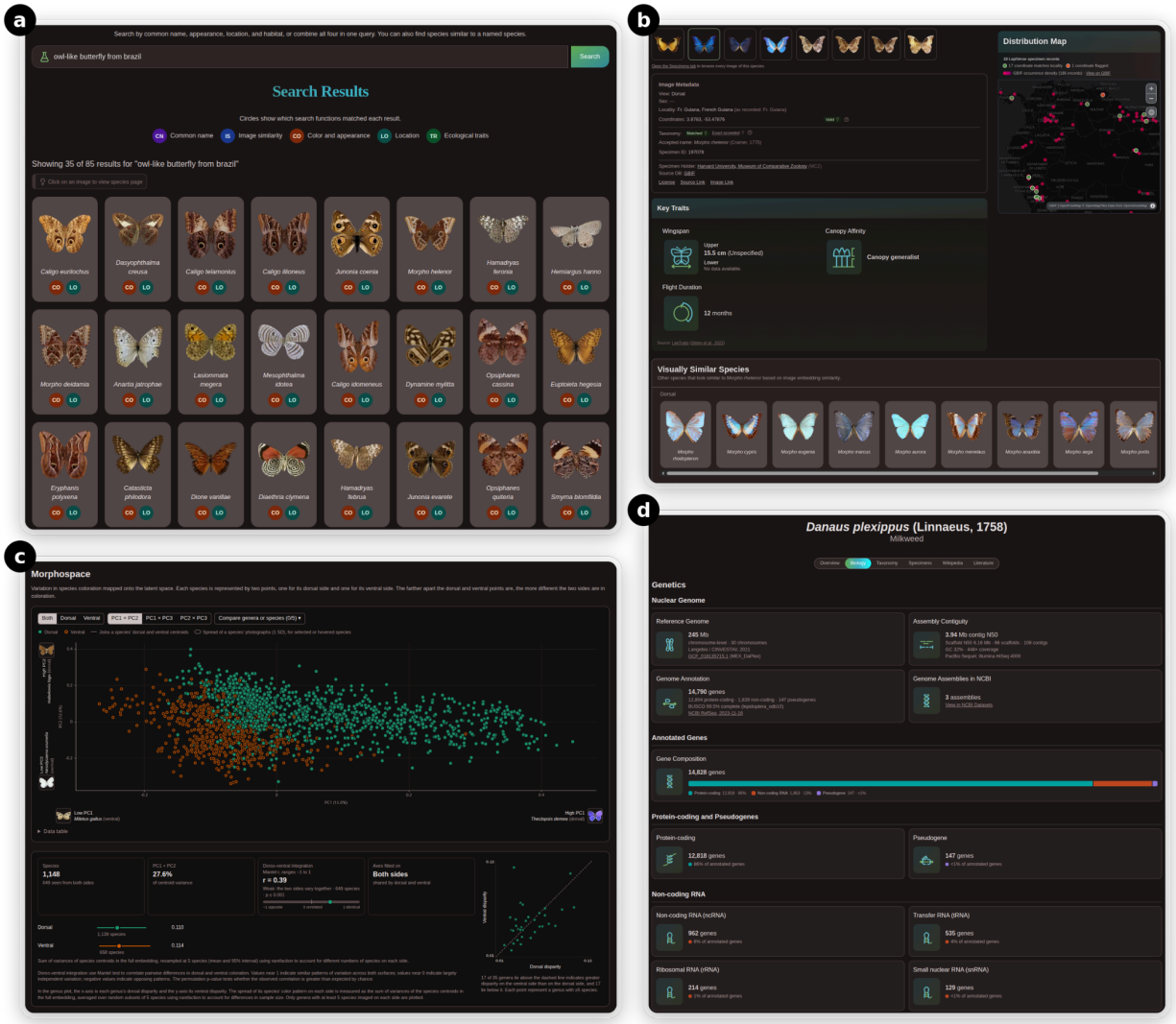


Figure 8. Feature screenshots of the LepiVerse butterfly portal: (a) semantic search results on "owl-like butterfly from Brazil"; (b) species page overview showing the *Morpho rhetenor* page; (c) family page showing morphospace of Lycaenidae; (d) the Biology section of the species page showing results for *Danaus plexippus*. The interface supports light and dark themes that users can toggle in the navigation menu. For consistency, all screenshots in this figure use the dark theme.

User interfaces

LepiVerse is the butterfly portal that integrates BioCosmos features for both internally and externally managed resources (Figure 2c). The key features are as follows: three search functions for standard query, semantic text search, and image search (see Figure 8a for search result interface); data-rich species pages, covering specimen galleries, visually similar species, distribution map, classification, genetic and trait data, Wikipedia pages, and literature (Figure 8b and d); higher taxon pages covering full taxonomy of all families and genera within Lepidoptera

available from the CoL accounts and the count and proportion of data available in the database; statistics of the data by taxonomy, country, specimen providers, and data aggregators.

On each species and higher-taxon page, we provide an interactive plot to explore the morphospace of species coloration based on its latent space in UNICOM. Users can filter and highlight to compare to other species within the same genus or family. The morphospace panels include principal component analyses of the vector embeddings up to PC3 for family and genus pages, and within-group variance, dorsum and ventrum disparity, and Mantel test results correlating the dorsum and ventrum embeddings (Figure 8c).

Discussion

We developed BioCosmos as a unified biodiversity data platform that enhances the use of specimen images from natural history collections (NHCs) while providing comprehensive features for studying taxonomy, distribution, and phenotypic diversity. The first implementation of the platform on the butterflies portal (LepiVerse) provides a proof of concept for BioCosmos's potential to improve access to and dissemination of image-based NHCs. Over recent decades, digitization efforts have increased the accessibility and use of natural history collections (Hedrick et al. 2020). Early digitization efforts largely focused on specimen inventory and collection management, such as tracking specimen loans at the Museum of Comparative Zoology (MCZ), Harvard University. Subsequently, digital specimen images have enabled new forms of access, with the Harvard University Herbaria experiencing substantial growth in specimen image distribution despite stable rates of physical loans (Hedrick et al. 2020). More recently, digitized specimens have expanded for large-scale data analytics, predicting species distribution, analyzing phenotypic variation, and reconstructing evolutionary and ecological changes. BioCosmos demonstrates how digital specimen records can be used with emerging machine learning techniques. These include using computer vision to extract visual features from specimen images to study complex coloration, and using vision-language and large-language models to enable natural-language queries of large digitized NHCs, an approach proposed as Digitization 2.0 in Hedrick et al. (2020).

The development of accessible biodiversity informatics tools is particularly urgent given growing concerns about declining taxonomic expertise. The shortage, further exacerbated by limited taxonomic education globally, has been recognized as a challenge for biodiversity discovery and conservation efforts (Engel et al. 2021, Groom et al. 2026). BioCosmos helps to address these challenges by making existing biodiversity knowledge more accessible to both specialists and non-specialists. BioCosmos features, particularly image-based retrieval and interactive phenotypic data exploration, could help users find relevant specimens and recognize phenotypic variation across major taxonomic ranks up to the family level. To support taxonomic work and education, we integrate the whole Catalogue of Life taxonomy for Lepidoptera (Cachuela-Palacio 2006, Hobern et al. 2021) and design an intuitive interface to track taxonomic changes, type information, name usages, and higher taxon taxonomy, listing all available records in each rank regardless of whether we have representative images in the database. We enhanced

the species page with the Crossref library to help track taxon-specific scientific literature. Nevertheless, our computational approaches complement, rather than replace, traditional taxonomic expertise, which requires hands-on experience with physical specimens (Engel et al. 2021).

Developing a sustainable, non-profit web application that integrates deep learning and large-language models presents substantial challenges in maintaining computational resources needed to operate the system. We mitigated this issue by designing BioCosmos to minimize unnecessary computation and limit the LLM to the generation of function calls that initiate database queries. We also precomputed embeddings and static results, such as visually similar species, and implemented multiple caching layers to limit runtime computation for each user request. Model selection is another key consideration in designing AI-powered applications, as model size alone does not reliably indicate performance or computational efficiency (Hoffmann et al. 2022). Our LLM test results showed that the 31-billion-parameter Gemma model outperformed all other models (20B-30B parameters) in accuracy and had the lowest latency and token consumption, despite being the largest model we tested. Furthermore, all the models we use, both LLMs and other DL models, are open-weight and hosted internally or institutionally, so other institutions can replicate or adapt our approach. Together, these strategies show that careful model selection, precomputation, caching, and lightweight system design are important for balancing operational costs with the functionality, accessibility, and replicability of AI-assisted biodiversity informatics platforms.

In addition to its application to studying butterfly diversity, BioCosmos demonstrates the feasibility of developing resource-efficient AI-assisted platforms that extend the scientific utility of existing digitized natural history collections (NHCs). Our evaluation of the UNICOM vision model shows its potential for identifying mimetic pairs using global butterfly image datasets. However, in some tested samples, the top ten results only return mimetic matches in one direction. The asymmetry is technically and biologically plausible. Retrieval performance depends on consistent lighting, specimen positioning, and color calibration across images (East et al. 2025). In some species, mimetic pairs vary by distribution and sex (Puissant et al. 2023). Misidentification of specimens is widespread in natural history collections and could distort the results, and taxonomic harmonization alone would not fix it (Goodwin et al. 2015). Because deep learning is rapidly evolving, BioCosmos's modular architecture offers an opportunity to adapt new models, and fine-tuning existing models could improve retrieval performance. Expanding deep learning features to detect potential specimen misidentification will increase the value of the platform. Nonetheless, BioCosmos is designed to aid discovery and hypothesis formulation. We harmonized and resolved institutional records to simplify locating and further examining the findings using physical collections.

Despite considerable progress in specimen digitization, substantial portions of global natural history collections remain digitally inaccessible (Johnson et al. 2023). Our dataset showed that most specimen images came from digitization efforts in the United States and England. This digitization geography was concentrated far from the most diverse regions in our

dataset and mirrors the spatial, taxonomic, colonial-legacy, and collector biases of the underlying natural history collections (Troudet et al. 2017, Daru et al. 2018, Moura et al. 2026). These biases propagate downstream and distort subsequent uses of the collections (Beck et al. 2014, Lipiński 2026), and BioCosmos is no exception. Coordinated efforts to develop affordable digitization tools and invest in supporting museum digitization are needed to cover these gaps, while ensuring data-sharing practices always credit the natural history museums that host and maintain the collections (Rouhan et al. 2017, Hedrick et al. 2020).

Acknowledgements

We thank Ethan Wolfe, Elan Grossman, Julian Perez, Romouald Dombrovski, Thomas Deatherage, and Zahara Shuemaker for their code contributions to the earlier version of the application. Stephanie Killingsworth, Katherine Wolcott, Caleb Gordon, and Lucas T. Bezerra provided feedback on the application features and its usage for K-12 teachers and a broader audience. We thank all the natural history museums listed in this publication for digitizing and sharing their specimen records. The University of Florida AI Seed Award, the University of Florida Biodiversity Institute, the University of Florida Research Computing, and the Florida Museum Undergraduate Internship Program provided funding and computational resources to maintain the application. The Human Augmented Analytics Group (HAAG) provided logistical support in developing earlier versions of the application. We benefited greatly from the open-source libraries and deep learning models provided by the Python and Node.js developers, as well as the UNICOM, CLIP, and Gemma developers. Other biodiversity database platforms, such as the Mammal Diversity Database, inspired some of the LepiVerse features.

Generative Large Language Model Usages

We used multiple frontier models, including GPT 5.5, 5.6, the 6 series models, and Claude 4.5, 5, and 5.5, during development to assist with writing code, prototyping user interfaces, and editing language. For coding, we used all models through the harnesses provided by the model developers. The authors decided the architecture, user interfaces and experience (UI/UX), dependency choices, code management, data flow, and all the features and their integration. We wrote the initial code by hand, and the authors reviewed all LLM-generated code. When needed, LLMs iteratively rewrote the code, or the authors rewrote it by hand. Pydantic, Pyright, and ESLint provided additional measures for data type checking and compliance with the language conventions.

Author contribution

Heru Handika (conceptualization, data curation, methodology, software, manuscript writing), Moritz D. Lürig (data curation, methodology, manuscript writing), Oliver M. Dobon (methodology), Kira L. Nichtawitz (software), Venkata S. Yarava (software), Justin K. Sobayo (software), Breanna D. Shi (conceptualization, methodology), Michael J. Elliott (conceptualization, methodology), Jose A. B. Fortes (conceptualization, methodology, funding

acquisition), Arthur Porto (conceptualization, methodology, software, manuscript writing, funding acquisition). All authors approved and contributed to the manuscript revisions.

Code Availability

The manuscript described BioCosmos/LepiVerse up to commit number 6afdf35. The code for the platform, the LepiVerse butterfly portal, and supported packages for data preparation and analyses are available at https://github.com/BioVisionLab/BioCosmos. The portal is served at https://lepiverse.com/. The image occurrence metadata is available at Zenodo (https://doi.org/10.5281/zenodo.23840624). The GBIF occurrence records used in this study are available at https://doi.org/10.15468/dl.a8vkwd.

References

An X, Deng J, Yang K, Li J, Feng Z, Guo J, Yang J, Liu T. 2023. Unicom: Universal and compact representation learning for image retrieval. ArXiv.

Bánki O, Roskov Y, Döring M, Ower G, Hernández Robles DR, Plata Corredor CA, Stjernegaard Jeppesen T, Örn A, Pape T, Hobern D, Garnett S, Little H, DeWalt RE, Miller J, Orrell T, Aalbu R, Abbott J, Abreu C, Acero P A, Acevedo-Charry O, Aedo C, Aescht E, Agboola B, Agosti D, Agudelo-Zamora HD, Ahlmer W, Ahyong S, Akhoundi M, Aldea C, Alexander S, Alonso-Zarazaga MA, Alvarez B, Ananjeva N, Anderson R, Andrade AJ, Andrella GC, Anions M, Anonymus, Antić DZ, Antonietto LS, Arango C, Arias-Buriticá JA, Arriaga-Varela E, Artois T, Ascuntar-Osnas O, Atahuachi Burgos M, Atkinson S, Atwood JJ, Avendaño JE, Avendaño C. JE, Backlund M, Bagnatori Sartori ÂL, Bailly N, Baixeras J, Baker E, Balan A, Balletto E, Bamber R, Bandyopadhyay S, Barber A, Barber-James H, Barbosa Pinto R, Barrett R, Barros-Barrios M, Bartolozzi L, Bartsch I, Beccaloni G, Bellamy CL, Bellan-Santini D, Bellinger PF, Beltrán León BS, Ben-Dov Y, Benichou L, Berlanga H, Bermúdez-Higinio MF, Bernot J, Beta J, Biffi G, Blasco-Costa I, Boatwright JS, Bock P, Bogota-Gregory JD, Bohn H, Bolton B, Borges LM, Borges P, Bortoluzzi R, Bossard RL, Bota-Sierra C, Botero JP, Bouchard P, Boury-Esnault N, Bousquet Y, Bouzan R, Boxshall G, Boyko C, Brandão S, Braun H, Bray R, Brazil Flora, Group, Brehm G, Brinda JC, Brock PD, Broich SL, Brosens D, Brouillet L, Brown J, Brown S, Brullo S, Bruneau A, Bush L, Büscher T, Błażewicz-Paszkowycz M, Cabras A, Cacabelos E, Cairns S, Calabuig I, Calderón-Parra R, Calonje M, Campos-Filho IS, Cardinal-McTeague W, Cardona-Duque J, Cardoso D, Cardoso L, Carvalho ABG de L, Carvalho P, Castaño Rivera A, Castilho RC, Castro EB, Castro Silva IC, Cervantes A, Chamorro-Fuertes WR, Chernyshev A, Chevillotte H, Chittaro Y, Choo LM, Choong H, Christiansen KA, Cianferoni F, Cifuentes-Ruíz P, Cigliano MM, Cisteil Xinum CX, Clarke R, Clavijo-Bustos J, Cobra e Monteiro T, Collins A, Collins S, Compton J, Copilaș-Ciocianu D, Corbari L, Cordeiro R, CoreoideaSF Team, Cortés-Hernández K, Costello MJ, Coursol F, Crameri S, Creuwels J, Cruz-López JA, Cuervo AM, Cárdenas P, Dagley B, Daly M, Daneliya M, Dangi R, Dauvin J-C, Davie P, Davies A, De Broyer C, De Grave S, De Lima HC, De Prins J, De Prins W, De Sousa F, De la Estrella M, DeSalle R, Decker P, Decock W, Delgado-Salinas A, Deliry C, Dellapé PM, Den Heyer J, Denno BD, Desmet P, Diewald W, Dijkstra K-D, Dincă V, Dmitriev DA, DoNascimiento C, Dohrmann M, Donsker D, Dorado Ó, Dorkeld F, Downey R, Duan L, Duno de Stefano R, Dämmrich F, Dépaquit J, Díaz M-C, Eades DC, Ebert M, Echeverri-Galvis MÁ, Egan AN, Eitel M, El Nagar A, Eleaume M, Emig CC, Enghoff H, Esquete Garrote P, Estela FA, Evans GA, Evenhuis NL, Falcão M, Faltynek Fric Z, Farruggia F, Fauchald K, Fautin D, Favreau M, Favret C, Fenner D, Ferm J, Figuerola B, FinBIF, Fisher B, Fišer C, Fonseca-Cortés A, Fontinha S, Forró L,

Fortuna-Perez AP, Fortune-Hopkins H, Franquinho Aguiar A, Fritsch P, Fuchs A, Fujimoto S, Furuya H, Gagnon E, Galati E, Galea H, García K, García Morales M, García-Prieto L, Gargominy O, Garic R, Gasca R, Gasca Álvarez HJ, Gattolliat J-L, Geiger M, Geiser M, Georgiev G, Georgiev T, Gerken S, Gibau de Lima A, Gibson D, Gielis C, Gill F, Gilligan T, Giribet G, Girón Duque JC, Gittenberger A, Giusso del Galdo G, Gofas S, Golikov A, Gomez C, Goncharov M, Gondim AI, Gonseth Y, González-Alvarado A, González-Córdoba M, Goodwin C, Govaerts R, Grabowski M, Granado A de A, Gregório B de S, Grehan JR, Grether R, Gross O, Guerra-García JM, Guglielmone A, Guilbert E, Guillou L, Guldberg Frøslev T, Gurumse W, Gusenleitner J, Guzmán G, Gyeltshen C, Gómez C, Gómez de Silva H, Göker M, Haas F, Haas M, Hadfield KA, Hajdu E, Hassler M, Hastriter MW, Hausmann A, Hayward BW, Hendrich L, Hendrycks E, Hennen D, Henry TJ, Hernandes FA, Hernandez D, Hernandez F, Hernández D, Hernández Macías H, Hernández-Crespo JC, Herrera-Collazos EE, Herrera-R G, Hincapié-Montoya D, Hincapié-Montoya DM, Hine A, Ho B-C, Hodson A, Hoeksema B, Hoenemann M, Holstein J, Hooge M, Hooper J, Hopkins H, Hopkins M, Horak I, Horton T, Hošek J, Hughes C, Hughes L, Hughes T, Huys R, Háva J, Häuser C, Höfer H, Hüne M, Iganci J, Iniesta LFM, Institute of Zoology Chinese Academy of Sciences, International Union for Conservation of Nature, Ishikawa T, Ishong J, Jackson C, Janssens F, Jardim R, Jaume D, Javadi F, Jazdzewski K, Jenkins Shaw J, Jersabek CD, Johnson KP, Johnston MA, Jordão L, Jóźwiak P, Kajihara H, Kakui K, Kallies A, Kamiński MJ, Kanda K, Kathirithamby J, Kelly M, Khalikov R, Kilian N, Kim Y-H, King R, Kirk P, Kitching I, Klautau M, Klazenga N, Kleukers R, Klitgaard BB, Kmentova N, Koenemann S, Koenen E, Kostova R, Kotov A, Kramina T, Krapp-Schickel T, Kremenetskaia A, Kroh A, Kroupa AS, Kury AB, Kury MS, Kvaček J, Lachenaud O, Ladino-Peñuela A, Lado C, Lamas G, Lambert G, Lana C. Atunes L, Lavin M, Lazarus D, Le Coze F, Le Roux M, LeCroy S, Ledis Linares J, Lee H, Lee S, Leitner MF, Leuro Robles NG, Lewis GP, Li S-J, Li Y-H, Li-Qiang J, Lichtwardt RW, Lim S-C, Lindsay D, Liu H, Lohrmann V, Longhorn SJ, Lorenz W, Lowry J, Loyer M, Lozano F, Lukhtanov V, Lumen R, Lyal CHC, López-Bedoya P, Lörz A-N, Maes D, Magnien P, Mah C, Makhov I, Mal N, Malécot V, Mamos T, Manconi R, Mansano V, Markello K, Martens K, Martin JH, Martin P, Martínez Muñoz CA, Martínez-Revelo DE, Mashego KS, Maslakova S, Maslin B, Mattapha S, Mańko M, McFadden C, McKamey S, McMurtry JA, Meades S, Means J, Medina-Uribe CA, Medrano MA, Mees J, Meier-Kolthoff JP, Meland K, Melo I, Mendes AC, Mendoza-Garfias B, Menezes de Sequeira M, Merrin K, Mesa NC, Messing C, Mielke CGC, Migeon A, Migliore LJ, Miller DR, Miller GL, Mills C, Milto K, Minelli A, Miranda LS, Mironov V, Mita T, Mitchell D, eds. M, Molodtsova T, Monje JC, Monnerat C, Montenegro Valls JF, Montgomery A, Mooi R, Morales-Martínez DM, Morandini A, Moreira da Rocha R, Moreu I, Morrow C, Moteetee A, Munguira ML, Murillo-Ramos L, Murphy B, Narita JPZ, Nascimento EA, Nebel M, Neill K, Neita Moreno JC, Nelson W, eds. N, Neto AI, Neu-Becker U, Neubauer TA, Neuhaus B, Newton A, Ng Kee Lin P, Nguyen A, Nicolson D, Nielsen JE, Nijhof A, Nishikawa T, Niño-Suárez MA, Norenburg J, Novoa L, O'Donnell SA, O'Grady P, O'Hara T, Obondo D, Ochieng J, Ochoa R, Ohashi H, Ohashi K, Okposio E, Oliveira A, Oliveira P, Ollerenshaw J, Oosterbroek P, Opresko D, Ortega-Lara A, Ortega-Álvarez R, Osborne R, Osigus H-J, Oswald JD, Ota Y, Otte D, Ouvrard D, Paganucci de Queiroz L, Palacios Rodríguez JS, Pandey A, Parente MI, Parte AC, Paulay G, Paulson D, Penev L, Pennington RT, Pereira J da S, Pereira SGG, Perez-Gelabert D, Perveen F, Phillipson P, Pica D, Pinheiro U, Pires Morim M, Pisera A, Pitkin B, Plata C, Plotkin D, Poatskievick Pierezan B, Polanco A, Polanco F. A, Poore G, Povydysh M, Powell E, Prasad K, Praxedes RA, Prudhomme J, Pulawski WJ, Pyle R, Pérez-Hernández CX, Pühringer F, Rajaei H, Rakotonirina N, Ramos G, Ramírez-Chaves HE, Rando J, Randrianambinintsoa FJ, Ranzato Filardi F, Raper C, Rasmussen P, Ratcliffe S, Raz L, Read G, Reeder DM, Rees T, Reich M, Reimer JD, Reimer LC, Rein JO, Rendón Ramírez A, Reynolds J, Rincón J, Rius M, Robertson T, Robinson G, Robinson GS(†), Rodríguez E, Rodríguez WD, Rodríguez-Contreras V, Rosenfeld S, Ruggiero M, Ríos P, Sabater LM, Sabroux R, Salmela J, Samimi-Namin K, Sanborn A, Sanchez A, Sandoval-Espinel MK, Sanjappa M, Santos-Guerra A, Sardà Carbasse J, Sartori M, Sattler K, Sauter G, Schierwater B, Schilling S, Schley R, Schmid-Egger C, Schmidt S, Schmidt-Rhaesa A,

Schoolmeesters P, Schorr M, Schrire B, Schuchert P, Schuh RT, Schönberg C, Schütz Rodrigues R, Scoble M, Seijo G, Seleme EP, Senna A, Serejo C, Serrano A, Sforzi A, Shao K-T, Shenkar N, Shimabukuro P, Shimizu G, Siegel V, Sierwald P, Sihvonen P, Sikes D, Silva Flores A, Silva de Carvalho C, Sim-Sim M, Simmons E, Simon MF, Simonsen T, Simpson CE, Sinev S, Sinniger F, Sirichamorn Y, Sissa-Dueñas YP, Skipper L, Skvarla M, Smirnov I, Smirnov R, Smith A, Smith AD, Smith VS, Soares Gissi D, Sokoloff D, Sotuyo S, Souma J, Sousa da Silva G, Souza-Filho JF, Spearman L, Spelda J, Steiner A, Steiner W, Stemme T, Sterrer W, Stevenson D, Stiewe MBD, Stiles FG, Stirton CH, Straub S, Stueber G, Stöhr S, Subramaniam S, Swalla B, Swedo J, Sá HAB, Sáfián S, Sánchez Santos N, Sánchez-González LA, Sánchez-Ruiz M, Sérgio C, Sørensen MV, Taboada-Verona CA, Takiya DM, Tandberg AH, Tavakilian G, Tavera Vargas JJ, Taylor K, The Danish Agency for Green Land Use and Aquatic Environment, Thessen A, Thomas D, Thomas JD, Thomas P, Thomson S, ThripsWiki, Thuesen E, Thulin M, Thurston M, Thuy B, Todaro A, Todorov M, Toonen R, Torke BM, Tsai S-Y, Turiault M, Turner JRG, Turner T, Turon X, Tyler S, UCD Community, UNITE Community, Uchima Taborda DA, Uetz P, Ulmer JM, Usma Oviedo JS, Utria Suarez J, Vacelet J, Vachard D, Vader W, Valls Domedel G, Van der Burgt X, Vandepitte L, Vanhoorne B, Vargas-Canales V, Vatanparast M, Vaulot D, Verhoeff T, Verovnik R, Vieira P, Vila R, Villa-Navarro FA, Vliegenthart A, Vongphayloth K, Vonk R, Väinölä R, Wahlberg N, Walker-Smith G, Walter TC, Wambiji N, Wanke D, Warren M, Watling L, Weaver H, Webb J, Welbourn WC, Wesener T, Whipps C, White K, Wiemers M, Wilding N, Williams G, Wilson A, Wilson AJG, Wilson DE, Wing P, Winitsky S, Wirth CC, Wojciechowski M, Woodman S, Xavier J, Yi T, Yoder M, Yu DSK, Yunakov N, Zahniser J, Zapata F, Zapata LA, Zapata Padilla LA, Zaragoza-Caballero S, Zeidler W, Zhang R, Zinetti F, d'Hondt J-L, de Moraes GJ, de Oliveira ABR, de Voogd N, del Campo J, del Río MG, iNaturalist contributors, van Haaren T, van Nieukerken EJ, van Ofwegen L, van Soest R, van Swaay CAM, van der Meij S, Şentürk O, American Society of Mammalogists, Barcode of Life Data Systems (BOLD), Cornell Lab of Ornithology, Cornell University, ITIS, International Barcode of Life project (iBOL), International Committee on Taxonomy of Viruses (ICTV), Korovchinsky N M & Petrusek, Legume Phylogeny Working Group (LPWG), Natural History Museum Bern, Paleobiology Database contributors, Smithsonian Institution National Museum of Natural History, USDA, Agricultural Research Service, National Germplasm Resources Laboratory, WoRMS Editorial Board, World Flora Online. 2026. Catalogue of Life.

Beck J, Böller M, Erhardt A, Schwanghart W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics 19: 10–15.

Blackburn DC, Boyer DM, Gray JA, Winchester J, Bates JM, Baumgart SL, Braker E, Coldren D, Conway KW, Rabosky AD, de la Sancha N, Dillman CB, Dunnum JL, Early CM, Frable BW, Gage MW, Hanken J, Maisano JA, Marks BD, Maslenikov KP, McCormack JE, Nagesan RS, Pandelis GG, Prestridge HL, Rabosky DL, Randall ZS, Robbins MB, Scheinberg LA, Spencer CL, Summers AP, Tapanila L, Thompson CW, Tornabene L, Watkins-Colwell GJ, Welton LJ, oVert Project Team, Stanley EL. 2024. Increasing the impact of vertebrate scientific collections through 3D imaging: The openVertebrate (oVert) Thematic Collections Network. Bioscience 74: 169–186.

Bonebrake TC, Ponisio LC, Boggs CL, Ehrlich PR. 2010. More than just indicators: A review of tropical butterfly ecology and conservation. Biological Conservation 143: 1831–1841.

Cachuela-Palacio M. 2006. Towards an index of all known species: the Catalogue of Life, its rationale, design and use. Integrative Zoology 1: 18–21.

Chazot N, Panara S, Zilbermann N, Blandin P, Le Poul Y, Cornette R, Elias M, Debat V. 2016. Morpho morphometrics: Shared ancestry and selection drive the evolution of wing size and shape in Morpho butterflies: *MORPHO*MORPHOMETRICS. Evolution; International Journal of Organic Evolution

70: 181–194.

Constable H, Guralnick R, Wieczorek J, Spencer C, Peterson AT, VertNet Steering Committee. 2010. VertNet: a new model for biodiversity data sharing. PLoS Biology 8: e1000309.

Dalrymple RL, Kemp DJ, Flores-Moreno H, Laffan SW, White TE, Hemmings FA, Tindall ML, Moles AT. 2015. Birds, butterflies and flowers in the tropics are not more colourful than those at higher latitudes: Birds, butterflies and flowers are not more colourful in the tropics than at higher latitudes. Global Ecology and Biogeography: A Journal of Macroecology 24: 1424–1432.

Daru BH, Park DS, Primack RB, Willis CG, Barrington DS, Whitfeld TJS, Seidler TG, Sweeney PW, Foster DR, Ellison AM, Davis CC. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. The New Phytologist 217: 939–955.

East A, Campolongo EG, Meyers L, Rayeed SM, Stevens S, Zarubiieva I, Fluck IE, Girón JC, Jousse M, Lowe S, Perry KI, Betancourt I, Charney N, Donoso E, Fox N, Landsbergen KJ, Nepovinnykh E, Ramirez M, Singh P, Thapa-Magar K, Thompson M, Waite E, Berger-Wolf T, Lapp H, Mabee P, Stewart C, Taylor G, Record S. 2025. Optimizing image capture for computer vision‐powered taxonomic identification and trait recognition of biodiversity specimens. Methods in Ecology and Evolution 16: 2260–2275.

Elliott MJ, Luciano M, Fortes JAB. 2026. iChatBio: an Extensible Agentic System for Scientific Exploration and Integration of Open Biodiversity Data. 2026.

Engel MS, Ceríaco LMP, Daniel GM, Dellapé PM, Löbl I, Marinov M, Reis RE, Young MT, Dubois A, Agarwal I, Lehmann A. P, Alvarado M, Alvarez N, Andreone F, Araujo-Vieira K, Ascher JS, Baêta D, Baldo D, Bandeira SA, Barden P, Barrasso DA, Bendifallah L, Bockmann FA, Böhme W, Borkent A, Brandão CRF, Busack SD, Bybee SM, Channing A, Chatzimanolis S, Christenhusz MJM, Crisci JV, D'elía G, Da Costa LM, Davis SR, De Lucena CAS, Deuve T, Fernandes Elizalde S, Faivovich J, Farooq H, Ferguson AW, Gippoliti S, Gonçalves FMP, Gonzalez VH, Greenbaum E, Hinojosa-Díaz IA, Ineich I, Jiang J, Kahono S, Kury AB, Lucinda PHF, Lynch JD, Malécot V, Marques MP, Marris JWM, Mckellar RC, Mendes LF, Nihei SS, Nishikawa K, Ohler A, Orrico VGD, Ota H, Paiva J, Parrinha D, Pauwels OSG, Pereyra MO, Pestana LB, Pinheiro PDP, Prendini L, Prokop J, Rasmussen C, Rödel M-O, Rodrigues MT, Rodríguez SM, Salatnaya H, Sampaio Í, Sánchez-García A, Shebl MA, Santos BS, Solórzano-Kraemer MM, Sousa ACA, Stoev P, Teta P, Trape J-F, Dos Santos CV-D, Vasudevan K, Vink CJ, Vogel G, Wagner P, Wappler T, Ware JL, Wedmann S, Zacharie CK. 2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. Zoological Journal of the Linnean Society 193: 381–387.

Gallagher RV, Falster DS, Maitner BS, Salguero-Gómez R, Vandvik V, Pearse WD, Schneider FD, Kattge J, Poelen JH, Madin JS, Ankenbrand MJ, Penone C, Feng X, Adams VM, Alroy J, Andrew SC, Balk MA, Bland LM, Boyle BL, Bravo-Avila CH, Brennan I, Carthey AJR, Catullo R, Cavazos BR, Conde DA, Chown SL, Fadrique B, Gibb H, Halbritter AH, Hammock J, Hogan JA, Holewa H, Hope M, Iversen CM, Jochum M, Kearney M, Keller A, Mabee P, Manning P, McCormack L, Michaletz ST, Park DS, Perez TM, Pineda-Munoz S, Ray CA, Rossetto M, Sauquet H, Sparrow B, Spasojevic MJ, Telford RJ, Tobias JA, Violle C, Walls R, Weiss KCB, Westoby M, Wright IJ, Enquist BJ. 2020. Open Science principles for accelerating trait-based science across the Tree of Life. Nature Ecology & Evolution 4: 294–303.

Goodwin ZA, Harris DJ, Filer D, Wood JRI, Scotland RW. 2015. Widespread mistaken identity in tropical plant collections. Current Biology 25: R1066-7.

Groom QJ, De Nolf M, Estupinan-Suarez LM, Meeus S. 2026. Balancing the supply and demand for taxonomy: An analysis of European taxonomic capacity and policy needs. PloS One 21: e0347332.

He Y, Mulqueeney JM, Watt EC, Salili-James A, Barber NS, Camaiti M, Hunt ESE, Kippax-Chui O, Knapp A, Lanzetti A, Rangel-de Lázaro G, McMinn JK, Minus J, Mohan AV, Roberts LE, Adhami D, Grisan E, Gu Q, Herridge V, Poon STS, West T, Goswami A. 2024. Opportunities and challenges in applying AI to evolutionary morphology. Integrative organismal biology (Oxford, England) 6: obae036.

Hedrick BP, Heberling JM, Meineke EK, Turner KG, Grassa CJ, Park DS, Kennedy J, Clarke JA, Cook JA, Blackburn DC, Edwards SV, Davis CC. 2020. Digitization and the Future of Natural History Collections. Bioscience 70: 243–251.

Hobern D, Barik SK, Christidis L, T.Garnett S, Kirk P, Orrell TM, Pape T, Pyle RL, Thiele KR, Zachos FE, Bánki O. 2021. Towards a global list of accepted species VI: The Catalogue of Life checklist. Organisms, Diversity & Evolution 21: 677–690.

Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, Casas DDL, Hendricks LA, Welbl J, Clark A, Hennigan T, Noland E, Millican K, Van Den Driessche G, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Vinyals O, Rae J, Sifre L. 2022. Training Compute-Optimal Large Language Models. Paper presented at Advances in Neural Information Processing Systems 35. 2022, San Diego, California, USA.

Hoyal Cuthill JF, Guttenberg N, Huertas B. 2024. Male and female contributions to diversity among birdwing butterfly images. Communications Biology 7: 774.

Hoyal Cuthill JF, Guttenberg N, Ledger S, Crowther R, Huertas B. 2019. Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. Science advances 5: eaaw4967.

Jocher G, Qiu J. 2024. ultralytics: Ultralytics YOLO27, YOLO26, YOLO11, YOLOv8 — object detection, instance segmentation, semantic segmentation, image classification, pose estimation, object tracking. Github.

Johnson KR, Owens IFP, Global Collection Group. 2023. A global approach for natural history museum collections. Science (New York, N.Y.) 379: 1192–1194.

Lawrence CG, Ramirez M, Berger-Wolf T, McMilan O, Rubenstein D. 2026. The perceptual and spatial architecture of Mullerian mimicry in Heliconius Butterflies. bioRxiv 2026.03.02.709095.

Lipiński S. 2026. Spatial bias in open biodiversity data: How GBIF record quality shapes conservation analyses. Conservation 6: 40.

Lürig MD, Di Martino E, Porto A. 2024. BioEncoder: A metric learning toolkit for comparative organismal biology. Ecology letters 27: e14495.

Lürig MD, Donoughe S, Svensson EI, Porto A, Tsuboi M. 2021. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. Frontiers in ecology and evolution 9: 642774.

Lürig MD, Shirai LT, Mota L, Willmott KR, Freitas AVL, Porto A. 2025. Aposematic color patterns are the dominant axis of phenotypic diversification in Nymphalid butterflies. bioRxiv 2025.09.30.678834.

Mallet J. 2001. Mimicry: an interface between psychology and evolution. Proceedings of the National Academy of Sciences of the United States of America 98: 8928–8930.

Moura MR, Carvalho RL, Ceron K, Guedes JJM, Moroti M de T, Nakamura G. 2026. Local expertise anchors biodiversity documentation, but geopolitical power drives parachute discovery. Npj Biodiversity 5.

Mozzherin DY, Myltsev AA, Patterson DJ. 2017. 'gnparser': a powerful parser for scientific names based on Parsing Expression Grammar. BMC Bioinformatics 18: 279.

Nelson G, Ellis S. 2018. The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 374: 20170391.

Nelson G, Paul D, Riccardi G, Mast AR. 2012. Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 209: 19–45.

Oliver JC, Robertson KA, Monteiro A. 2009. Accommodating natural and sexual selection in butterfly wing pattern evolution. Proceedings. Biological Sciences 276: 2369–2375.

Oskyrko O, Mi C, Meiri S, Du W. 2024. ReptTraits: a comprehensive dataset of ecological traits in reptiles. Scientific Data 11: 243.

Piccini I, Pittarello M, Di Pietro V, Lonati M, Bonelli S. 2022. New approach for butterfly conservation through local field‐based vegetational and entomological data. Ecosphere (Washington, D.C) 13: e4026.

Popov D, Roychoudhury P, Hardy H, Livermore L, Norris K. 2021. The value of digitising natural history collections. Research Ideas and Outcomes 7.

Puissant A, Chotard A, Condamine FL, Debat V, Llaurens V. 2026. The coevolution of colour patterns and hindwing shapes on a large phylogenetic scale reveals predation-driven adaptive syndromes in swallowtail butterflies. Ecology Letters 29: e70303.

Puissant A, Chotard A, Condamine FL, Llaurens V. 2023. Convergence in sympatric swallowtail butterflies reveals ecological interactions as a key driver of worldwide trait diversification. Proceedings of the National Academy of Sciences of the United States of America 120.

Raasveldt M, Mühleisen H. 2019. DuckDB: an Embeddable Analytical Database. Paper presented at SIGMOD/PODS '19: International Conference on Management of Data. 25 June 2019, New York, NY, USA.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. 2021. Learning transferable visual models from natural language supervision. ArXiv.

Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, Jiggins CD, Chamberlain NL, Kronforst MR, Chen R, Halder G, Nijhout HF, McMillan WO. 2011. Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. Science (New York, N.Y.) 333: 1137–1141.

Ren T, Liu S, Zeng A, Lin J, Li K, Cao H, Chen J, Huang X, Chen Y, Yan F, Zeng Z, Zhang H, Li F, Yang J, Li H, Jiang Q, Zhang L. 2024. Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv [cs.CV].

Rosa AHB, Freitas AVL. 2024. The role of citizens in conservation science: a case study with threatened Brazilian butterflies. Journal of Insect Conservation 28: 1149–1160.

Rouhan G, Dorr LJ, Gautier L, Clerc P, Muller S, Gaudeul M. 2017. The time has come for Natural History Collections to claim co‐authorship of research articles. Taxon 66: 1014–1016.

Sanderson C, Braby MF, Bond S. 2021. Butterflies Australia: a national citizen science database for monitoring changes in the distribution and abundance of Australian butterflies. Austral Entomology 60: 111–127.

Sayol F, Wayman JP, Dufour P, Martin TE, Hume JP, Jørgensen MW, Martínez-Rubio N, Sanglas A, Soares FC, Cooke R, Mendenhall CD, Margolis JR, Illera JC, Lemoine R, Benavides E, Lapiedra O, Triantis KA, Pigot AL, Tobias JA, Faurby S, Matthews TJ. 2024. AVOTREX : A global dataset of extinct birds and their traits. Global Ecology and Biogeography: A Journal of Macroecology 33: e13927.

Schindel DE, Miller SE, Trizna MG, Graham E, Crane AE. 2016. The Global Registry of Biodiversity Repositories: A call for community curation. Biodiversity Data Journal 4: e10293.

Shirey V, Larsen E, Doherty A, Kim CA, Al-Sulaiman FT, Hinolan JD, Itliong MGA, Naive MAK, Ku M, Belitz M, Jeschke G, Barve V, Lamas G, Kawahara AY, Guralnick R, Pierce NE, Lohman DJ, Ries L. 2022. LepTraits 1.0 A globally comprehensive dataset of butterfly traits. Scientific data 9: 382.

Soria CD, Pacifici M, Di Marco M, Stephen SM, Rondinini C. 2021. COMBINE: a coalesced mammal database of intrinsic and extrinsic traits. Ecology 102: e03344.

Steinke D, Gemeinholzer B, Martínez-Meyer E, Noesgaard D, Young A, Schigel D. 2025. Globally aggregated biodiversity data impact predictive and descriptive research. Proceedings of the National Academy of Sciences of the United States of America 122: e2519119122.

Suzuki TK, Tomita S, Sezutsu H. 2019. Multicomponent structures in camouflage and mimicry in butterfly wing patterns. Journal of morphology 280: 149–166.

Teng D, Zhang W. 2024. The diversification of butterfly wing patterns: progress and prospects. Current Opinion in Insect Science 61: 101137.

Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific Reports 7: 9132.

Turner JRG. 1976. Adaptive radiation and convergence in subdivisions of the butterfly genus Heliconius (Lepidoptera: Nymphalidae). Zoological Journal of the Linnean Society 58: 297–308.

Willis CG, Law E, Williams AC, Franzone BF, Bernardos R, Bruno L, Hopkins C, Schorn C, Weber E, Park DS, Davis CC. 2017. CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. The New Phytologist 215: 479–488.

Zhao Z, Lu Y, Tong Y, Chen X, Bai M. 2023. PENet: A phenotype encoding network for automatic extraction and representation of morphological discriminative features. Methods in Ecology and Evolution 14: 3035–3046.