

A digitised specimen-level database of the Arthur Bott Lepidoptera collection

Felix Weber

University of Würzburg

Johannes Balkenhol

University of Würzburg

Aman Akash

University of Würzburg

Mirko Wölfling

bio-advice

Thomas Dandekar

`thomas.dandekar@uni-wuerzburg.de`

University of Würzburg

data-descriptor

Keywords:

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9575316/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Historical insect collections hold century-scale records of species distributions and phenotypic traits, yet the majority remain undigitised. We present a specimen-level database of 5,974 Lepidoptera records from the private collection of Arthur Bott (Poppenhausen, Lower Franconia, Germany), covering the period 1900–2002 and encompassing 712 taxa across 351 genera. Records include 129 specimens from 19 species classified as extinct in Bavaria and 163 specimens from 30 endangered species, as well as 156 specimens of 14 species listed in Annex II and 293 specimens of 19 species listed in Annex IV of the EU Fauna–Flora–Habitat Directive. The data are organised in a normalised SQLite schema that separates specimen, species, location and trait information, with stable identifiers linking each record to external resources (GBIF, BOLD Systems, EOL and Wikipedia). Data are exportable in Darwin Core and ABCD XML formats. Image-derived colour brightness values are provided for 77 specimens of *Euphydryas aurinia* sampled across an Alpine altitudinal gradient from 600 to 2,300 m.

Background & Summary

Insect collections represent one of the most extensive biological archives on Earth, with an estimated half a billion specimens held worldwide¹. These collections encode century-scale records of species distributions, population traits and environmental conditions that cannot be replicated by contemporary sampling alone. Such data underpin work on climate-driven range shifts, phenological change and conservation planning^{2–4}.

Digitisation of insect collections also contributes to the assessment of species loss and ongoing biodiversity decline^{5–7}, and supports studies of pest population dynamics in applied agricultural and forestry entomology^{8–9}. Long-term changes in community composition can be tracked by combining historical collection data with contemporary empirical records^{10–12}.

Large-scale digitisation initiatives are underway at major institutions such as the Naturalis Biodiversity Center, the Natural History Museum London and the Museum für Naturkunde Berlin, yet progress remains slow. The Museum für Naturkunde Berlin reports that approximately 1.7% of its 30 million objects had been digitised by 2021; at current rates, completion would span many decades. The Natural History Museum, London faces a comparable challenge, with approximately 80 million objects including over half a million British and Irish Lepidoptera. Private and regional collections, which may hold irreplaceable local records, face even greater barriers due to limited resources for software, infrastructure and personnel^{13–14}. Lightweight, open-source solutions are therefore particularly valuable for institutions and private collectors with limited resources.

Following the FAIR principles — Findability, Accessibility, Interoperability and Reusability¹⁵ — maximises the scientific value of digitised collections by enabling integration across biodiversity data infrastructures. Beyond occurrence data, dried Lepidoptera specimens preserve phenotypic information (wingspan, colour intensity, sex ratios, chemical residues) that records conditions at the time of

collection^{16–19}. Intraspecific morphological variation across environmental gradients can reveal local adaptation processes^{20–21}.

Wing colouration in Lepidoptera is among the most tractable phenotypic traits measurable from museum specimens, and it carries direct ecological significance. The thermal melanism hypothesis predicts that ectotherms at higher, cooler elevations should exhibit darker colouration to enhance solar heat absorption and maintain body temperature for flight^{22–24}. This prediction has been supported in several butterfly taxa^{25–26} but rarely tested using systematic private collections, which can provide dense altitudinal sampling over decadal timescales.

The dataset described here was generated from the historical Lepidoptera collection of Arthur Bott (1939–2022), a systematic amateur entomologist based in Poppenhausen, Lower Franconia, Germany. The collection is taxonomically well organised and contains precise locality and date records spanning six decades of active collecting (1956–2002). The total collection comprises 5,974 specimens distributed across 47 insect boxes, with collection dates ranging from 1900 to 2002. It encompasses 712 taxa including 680 species across 351 genera, representing a substantial proportion of the Central European Lepidoptera fauna. Figure 1 shows a representative insect box from the collection.

[INSERT Fig. 1 HERE]

Figure 1. Representative insect boxes from the Bott Lepidoptera collection, showing pinned specimens of the family Nymphalidae. The collection spans 47 boxes with specimens dating from 1900 to 2002.

The collection includes all European Rhopalocera families (Lycaenidae, Riodinidae, Hesperidae, Papilionidae, Nymphalidae, Pieridae) as well as a wide range of Heterocera families (Brahmaeidae, Drepanidae, Endromidae, Erebidae, Euteliidae, Lasiocampidae, Noctuidae, Nolidae, Notodontidae, Saturniidae, Sphingidae, Limacodidae). Specimens from the Cossidae and Geometridae families are physically present but not yet identified to species level and are therefore not included in the current dataset. Specimens originate from 30 countries, mainly in Europe; most were collected near Poppenhausen, the collector's residence. While most specimens were wild-caught, a small number of bred individuals are also included.

The collection was assembled by Arthur Bott and is currently held and curated by Dr. Mirko Wölfling. Arthur Bott collected over half the specimens himself between 1956 and 2002, using the identification guides of Forster & Wohlfahrt. The collection is organised into day-flying butterflies (Rhopalocera) and moths (Heterocera), sorted by family, with original labels attached.

Data are organised in a FAIR-aligned relational schema implemented in SQLite and served through an R/Shiny web application. Records carry stable identifiers and are linked to external biodiversity resources (GBIF, BOLD Systems, EOL, Wikipedia). Conservation-status information from the EU Fauna–Flora–Habitat Directive and the Red List of Bavaria is integrated at species level. Image-derived colour brightness values are provided for 77 specimens of *Euphydryas aurinia* sampled along an Alpine

altitudinal gradient, demonstrating how image-derived phenotypic traits can be attached to specimen records and used in downstream analyses. The database is designed as a transferable blueprint for the digitisation of similar Lepidoptera or other insect collections, in particular for small institutions and private collectors whose specimens remain largely undocumented^{13,14}.

Methods

Collection digitisation

In a preliminary digitisation step, specimen data were transferred into a spreadsheet. The initial spreadsheet was in German and contained the columns Family, Genus, Species, combined genus/species, German name, Location, Country, Altitude (m), Date of collection, Date accuracy, Breeding, Collector and Digitizer's notes. The following columns were subsequently added: Suborder (Rhopalocera or Heterocera, following the organisation of the physical collection), GPS coordinates of the collection site, Total number of individuals of a species, status under the Fauna–Flora–Habitat Directive and status under the Red List of Bavaria.

GPS coordinates for individual locations were obtained through Google Maps. Uncertain or imprecise georeferences were flagged in the Digitizer's notes column. Locations with missing, ambiguous or untraceable names were left without coordinates. Total specimen counts per species were computed retrospectively. FFH Directive and Bavarian Red List status were joined into the main table from two separate reference spreadsheets using the XLOOKUP function. The general lookup formula is retained in cell C3 of the FFH and Bavarian Red List reference spreadsheets provided with the code repository.

[INSERT Fig. 2 HERE]

*Figure 2. Example of a catalogued specimen from the Bott collection: *Iolana iolas* (Ochsenheimer, 1816) mounted on an insect pin. The attached label provides metadata such as country, state, region or city, date of collection and collector's name, which are transcribed into the specimen record.*

Photography

A representative photograph was taken for each species. In cases of sexual dichroism or dimorphism, separate images of both sexes were taken where available and combined into a single image, with each sex labelled (♂ ♀). Files were named by genus, species and, where applicable, subspecies.

Images were acquired with a Nikon Coolpix S9500 digital camera in macro mode. A cube-shaped softbox provided even diffuse lighting for the photographed specimens. A custom wooden stand was built to position the pinned specimens close to the camera: a base plate with a balsa-wood strip attached at a 90° angle, drilled at the top with two holes of different diameters to accommodate insect pins of different sizes. A sheet of white paper served as the background. ISO sensitivity, white balance

and softbox illumination were held constant across all sessions to minimise variability in apparent colour brightness between images. The setup with a mounted specimen is shown in Fig. 3.

[INSERT Fig. 3 HERE]

*Figure 3. Setup for photographing pinned butterflies, exemplified by *Papilio machaon* (Linnaeus, 1758). The specimen is mounted on a pin within a standardised lighting chamber to minimise illumination variance. The setup enables consistent comparisons of specimens from different localities, altitudes or collection dates.*

Relational database schema

To ensure long-term accessibility, interoperability and reusability, the specimen data, image-derived trait data and auxiliary reference data were transferred into a normalised relational database implemented in SQLite. The schema (Fig. 4) is organised around six core tables: specimen, species, location, species_traits, location_traits and colour_brightness. The separation reflects conceptual entities and minimises redundancy.

The specimen table contains records of individual specimens, including collection date and location references, individual count, collector metadata, references to the associated species record, image references and digitiser's notes. The species table holds taxonomic information (genus, species, family, suborder) and assigns a unique species_id. Conservation-status fields at species level link to the Red List of Bavaria (organised into Rhopalocera, Macroheterocera and Microlepidoptera sections with separate revision dates) and the EU Fauna–Flora–Habitat Directive (Annexes II and IV).

The species_traits table holds one-to-many mappings between species and external resources (Wikipedia, BOLD Systems, EOL, GBIF), with persistent identifiers and URLs for molecular and ecological trait integration, as well as curated fields for breeding behaviour. The location and location_traits tables separately store geographical coordinates and derived environmental variables (elevation, annual mean temperature, annual precipitation). The colour_brightness table captures quantitative image-derived colour measurements per specimen, linked by specimen PID.

All data entities are assigned stable identifiers (specimen PID, species_id, location_id). Foreign-key relationships enforce referential integrity. External identifiers (EOL, BOLD, GBIF) are stored as typed fields so that the records can be integrated into global biodiversity information systems. The schema is extensible to additional data types (morphometric, genomic, citizen-science contributions) and can be deployed within wider biodiversity data pipelines or registered in domain-specific repositories such as GFBio.

[INSERT Fig. 4 HERE]

Figure 4. Relational schema of the database. Normalised SQLite database structure linking specimen records to taxonomic, spatial, trait and image-derived data. Core tables include specimen, species and location, with associated traits stored in species_traits, location_traits and colour_brightness. External resources (EOL, BOLD, GBIF, Wikipedia) are integrated via persistent identifiers.

Data integration and provenance

Primary input data originated from four separate spreadsheets: the specimen list, the colour brightness dataset and two conservation-status sources (the FFH Directive and the Red List of Bavaria). These were harmonised into the relational schema through controlled vocabularies and consistent identifier assignment. Taxonomic consistency was ensured by the use of accepted names and persistent taxon identifiers. Trait and conservation information were linked to species via species_id; visual trait data were linked to specimens via PID.

Image-derived colour brightness measurements

Image-derived colour brightness values were computed for 77 specimens of *Euphydryas aurinia* (Rottemburg, 1758) sampled across an Alpine altitudinal gradient of 600–2,300 m. The species was selected because a sufficient number of specimens covering a broad altitudinal range was available in the combined material. To expand the dataset, specimens from an additional historical collection were included in the colour-brightness analysis. Because the analysis was conducted on a single species, only one species name appears in the brightness plots; however, the species is split into distinct geographic populations so that brightness variation can be evaluated separately per subpopulation.

Specimens were entered into a dedicated colour-brightness spreadsheet. Where the collection-site elevation was not recorded on the label, it was determined from the georeferenced locality. Images were acquired under the standardised photography setup described above (Fig. 3). ISO sensitivity, white balance and softbox illumination were held constant across all specimens.

Image analysis in Fiji

Image analysis was performed in Fiji²⁷. Each image was opened in Fiji, a threshold was applied to separate the specimen from the background, a selection was generated from the thresholded area, and the Analyze Particles command was used to extract the pixel properties of the selected region. The resulting mask was overlaid on the image for quality control (Fig. 5). Mask errors, typically caused by within-specimen pattern elements being incorrectly included in the background, were corrected by manual threshold adjustment.

[INSERT Fig. 5 HERE]

Figure 5. Mask applied for determining colour brightness in Euphydryas aurinia (Rottemburg, 1758). (A) binary mask defining the area of interest; (B) mask overlaid on the specimen image; (C) an example of a masking error, in which complex wing-pattern elements are incorrectly included in the background region.

Due to the large number of images, a macro was written in the Fiji macro editor to automate the colour-brightness extraction across all specimens. The macro is available in the digital appendix under "\Colour Brightness". Summary output of all analysed images was exported as a single spreadsheet, also provided in the digital appendix.

The primary brightness metric reported is the Mean value returned by Fiji, i.e. the average of the per-pixel grayscale values within the mask²⁸. Values are dimensionless and lie on an 8-bit grayscale from 0 (black) to 255 (white).

Threshold sensitivity testing

To characterise the effect of threshold choice on the derived brightness values, all specimen images were combined into a single image stack and analysed using a further macro that iterated across threshold values. Testing was initiated at extreme (very low and very high) thresholds to establish the upper bound of the effect of threshold choice; intermediate thresholds were then evaluated in fixed increments. For each threshold, the percentage of dark (below-threshold) area relative to the total specimen area was recorded, alongside the mean brightness of the selected region. The resulting summary table is available in the digital appendix under "\Colour Brightness" as "Summary colour brightness Euphydryas aurinia population Alps".

The complex colour patterns of *E. aurinia* contribute to variability in derived brightness values. Figure 6A shows an example specimen from 599 m elevation; Fig. 6B shows a specimen from 2,300 m, in which the orange hue appears visibly paler and the dark wing pattern more extensive. To better resolve the dark component of the wing pattern specifically, an alternative mask was constructed at an elevated threshold to isolate the melanised (dark) area alone (Fig. 6C). This allows independent assessment of the dark area fraction as a complementary metric to mean brightness.

[INSERT Fig. 6 HERE]

Figure 6. Complexity of butterfly wing patterns motivates the use of multiple thresholds. Specimens of Euphydryas aurinia (Rottemburg, 1758) collected at (A) 599 m and (B) 2,300 m above sea level. (C) Mask with an elevated threshold, designed to isolate the dark (melanised) wing area as a complementary metric to overall mean brightness.

Software architecture

The application was implemented in R/Shiny with a SQLite backend. The full source code, the SQLite schema definition, the Fiji macro, the reference lookup tables for the FFH Directive and the Bavarian Red List, and the Docker container definition are publicly available on GitHub (https://github.com/johannes-balkenhol/butterfly_database). The project is organised in a modular layout:

The Database layer (data/) contains the central SQLite database (butterfly_collection.sqlite) and four original spreadsheets used during development: butterfly_database.xlsx, colour_brightness_butterfly.xlsx, FFH_List.xlsx and Red_List_Bavaria.xlsx. High-resolution specimen images and image-derived colour data are stored separately for performance and modularity.

Application code is separated into global.R, a server/ directory and a ui/ directory, following the standard R/Shiny separation of logic, display and controls. The scripts/snippets/ directory contains database formatting and helper functions for table handling. The www/ directory hosts more than 400 specimen images used in the application interface, together with custom map icons. A Docker container (docker/entrypoint.sh) ensures reproducible deployment across systems and supports long-term preservation, including containerised archiving for GFBio registration. The architecture is transferable: collections with a similar structure (images plus tabular metadata) can adopt the same codebase with minimal adaptation.

Data Records

The specimen-level dataset (5,974 Lepidoptera records) is deposited in Zenodo²⁹, in Darwin Core format (occurrence.txt) with an accompanying EML metadata file. A copy of the dataset in ABCD XML is also provided in the same deposit. The image-derived colour brightness dataset for the 77 *Euphydryas aurinia* specimens (colour_brightness_euphydryas_aurinia.csv) is included as a separate file within the deposit, together with the threshold-sensitivity summary table.

The following fields are included per specimen record: specimen PID (stable identifier), species_id, family, genus, species, subspecies (where applicable), German vernacular name, country, locality name, latitude, longitude, elevation (m), annual mean temperature (°C, where available), annual precipitation (mm, where available), collection date, date accuracy flag, collector, bred/wild-caught flag, FFH Directive Annex (II, IV, none), Bavarian Red List category (where applicable), external identifiers (GBIF taxon key, EOL page ID, BOLD taxon ID), image file reference and digitiser's notes.

The colour_brightness file contains, per image: specimen PID, species, locality, elevation (m), mask threshold, specimen area in pixels, dark area in pixels, dark-area fraction and mean grayscale brightness (0–255).

Interactive web interface

The dataset is interactively accessible at <https://butterflies.bioinfo-wuerz.eu>. The web interface (Fig. 7) consists of a header with a title, a sidebar with filter inputs and a main display area. The sidebar allows users to refine searches through systematic-group selection (toggle between Rhopalocera and Heterocera), conditional panels for family selection, text inputs for genus, species, German vernacular name and country, an elevation slider, a date-range picker, a specimen-count slider, image display toggle, filters by FFH Directive and Bavarian Red List status, and options to include or exclude digitiser notes. A Run button executes the query.

The main display area is organised into collapsible sections: (i) a searchable results table with unique specimen identifiers and key metadata; (ii) a collapsible phylogenetic tree view; (iii) a geographic map of collection localities; and (iv) an interactive analysis panel linking elevation with image-derived colour brightness. All query results can be downloaded as CSV or, following FAIR principles, in standardised Darwin Core and ABCD XML formats.

[INSERT Fig. 7 HERE]

Figure 7. Interactive interface of the digitised Lepidoptera database. The left panel displays searchable filter controls. The main window (right) includes a taxonomic tree viewer, a results table with stable identifiers and representative images, a geographic map of collection localities, and an interactive analysis panel showing elevation–colour-brightness relationships, alongside the currently active filter metadata.

Clicking on a specimen's persistent identifier in the results table opens a detailed specimen information panel (Supplementary Figure S1). The panel is organised into three sections: specimen-level data (collection date, individual count, FFH Directive status, Bavarian Red List category, collector metadata); species-level data (full taxonomic hierarchy, vernacular names, accepted binomial nomenclature); and environmental context (coordinates, country, annual temperature and precipitation where available). In addition to local metadata, the panel provides direct outbound links to GBIF, BOLD Systems, EOL and Wikipedia for global occurrence records, DNA barcode and sequencing data, ecological trait profiles and taxonomic summaries.

The taxonomic tree viewer (Fig. 8A) visualises the hierarchy of the collection from order (Lepidoptera) down to species level, allowing users to navigate by evolutionary relationship. The interactive geographic map (Fig. 8B) uses distinct icons to differentiate Rhopalocera and Heterocera, with family-level colour coding. Users can click icons to open popups containing specimen metadata (genus, species, collection data); zooming into clustered markers separates them dynamically to display individual specimens. A legend in the lower left corner decodes icon colours and shapes.

[INSERT Fig. 8 HERE]

Figure 8. Features of the web interface. (A) Taxonomic tree viewer showing the hierarchy of the collection from order to species level. (B) Interactive geographic map displaying collection localities, with icons

differentiating Rhopalocera and Heterocera and colour-coded by family. Clicking an icon retrieves specimen-level metadata.

Data Overview

As a demonstration of the analytical content of the dataset, we examined the relationship between image-derived colour brightness and collection elevation in *Euphydryas aurinia*. Normalised colour brightness was computed for each specimen as described in the Methods and imported into the web application, where exploratory scatter plots and linear regression can be generated interactively.

Figure 9 shows the relationship between elevation and normalised colour brightness across selected taxa and elevation ranges, with Pearson correlation coefficients and linear regression lines calculated for each subset. Across the Alpine *E. aurinia* sample ($N \approx 77$, elevation range 600–2,300 m), brightness decreases significantly with increasing elevation (Pearson's $r \approx -0.60$, $p < 0.001$). Over this range, the average normalised brightness decreases from approximately 75 at lower elevations to 65 at higher altitudes (mean elevation 1,600 m).

[INSERT Fig. 9 HERE]

Figure 9. Relationship between normalised colour brightness and collection elevation across Lepidoptera populations in the dataset. Scatter plots with linear regression lines show the direction and strength of altitude–brightness correlations per subpopulation. The Alpine population of Euphydryas aurinia (highlighted) exhibits a significant negative correlation; other populations show weaker or less consistent trends, largely reflecting differences in sample size and elevation range (see Technical Validation).

To reduce potential bias from geographically distinct subpopulations, the Alpine subset was refined by excluding specimens from the locality "Basses Alpes, near Digne", which is geographically isolated from the main Alpine sampling region and may represent a distinct biogeographic lineage (see Technical Validation). After exclusion, the regression line indicates a more pronounced decrease in colour brightness with increasing elevation (Fig. 10).

[INSERT Fig. 10 HERE]

Figure 10. Relationship between colour brightness and elevation in Euphydryas aurinia populations from the Alps. Normalised colour brightness values are plotted against collection elevation for E. aurinia specimens sampled across a broad altitudinal range in the Alpine region (Basses-Alpes locality excluded; see Technical Validation). The trend line indicates a significant negative correlation.

Representative images of *Euphydryas aurinia* collected at different altitudes are shown in Fig. 6. The specimen collected at 599 m appears visibly brighter than the specimen collected at 2,300 m, providing qualitative visual support for the quantitative pattern.

Technical Validation

Taxonomic validation

Species names were cross-checked against the GBIF Backbone Taxonomy³⁰ to identify typographic errors, synonyms and unaccepted names. Where discrepancies were found, the accepted name and authorship were adopted in the species table; the original label transcription is retained in the digitiser's notes. The GBIF taxon key is stored as an external identifier in the species_traits table. Additional outbound links (BOLD, EOL, Wikipedia) allow reviewers to cross-check individual records against external taxonomic, molecular and occurrence sources.

Georeference validation

Coordinates were verified against the label locality during transcription and flagged as uncertain in the digitiser's notes where the label text was ambiguous. Records without a resolvable locality carry no coordinates. Elevation values were retrieved from the georeferenced locality and compared with elevation information on the label where present.

Conservation-status validation

FFH Directive Annex II/IV status and Bavarian Red List category were joined from the respective reference lists by species name using XLOOKUP. The reference lists are included as auxiliary files in the code repository to allow reviewers to reproduce the join. The last revision dates of the three Bavarian Red List sections (Rhopalocera, Macroheterocera, Microlepidoptera) are recorded in the species table, so that users can identify records linked to outdated revisions of the list.

Colour brightness: threshold sensitivity

The derived colour brightness metric depends on the threshold used in the Fiji mask. Threshold sensitivity was characterised by re-running the mask-based extraction across a range of threshold values, from the extremes to the middle of the scale. All threshold settings, mask statistics and derived brightness values are provided in the data record.

The sign of the brightness–elevation relationship for the Alpine Euphydryas aurinia sample is stable across the tested threshold range, while the magnitude of the regression slope varies with threshold in a manner consistent with the changing share of melanised (dark) pixels captured at each threshold level. For example, a threshold of 50 yielded a regression with a negative slope of approximately one brightness unit per elevation bin; a threshold of 99.5 (used initially) yielded a slope of approximately five units; a threshold of 80 approximately ten units. The percentage of dark area in relation to total specimen area decreased by about 5% with increasing elevation at a threshold of 50; at a threshold of 80, the dark-

area fraction showed no clear trend with elevation, even though the mean-brightness trend was retained. The mid-range threshold (approximately 50) is therefore the most informative single-threshold choice for this dataset, but we provide values across the full range so users can select or combine thresholds according to the question being addressed.

Colour brightness: subpopulation structure

Differences in the strength of the brightness–elevation relationship between subpopulations (Fig. 9) are consistent with differences in the structure of each subpopulation rather than with genuine absence of an altitudinal effect. Subpopulations with narrow elevation ranges, heterogeneous sampling or limited representation at key altitude levels have reduced statistical power.

For example, the Harz subpopulation yields a regression line with a steep slope, but the specimens originate from a narrow elevation interval of approximately 20 m (255–275 m), limiting interpretability. The Alpine subpopulation, by contrast, spans an elevation gradient of approximately 1,701 m and shows the largest brightness shift between the lowest and highest points (approximately 7.2 units on the 0–255 scale). While modest in absolute terms, the combination of a wide gradient and dense sampling makes the Alpine trend both statistically robust and biologically interpretable. The Lower Franconia subpopulation shows almost no change (approximately – 0.4 units), and the pooled trend across all specimens (62.4 → 67.0) remains small, indicating that overall brightness is approximately stable across the species and that altitudinal effects are best assessed within geographically coherent subpopulations.

Specimens from "Basses Alpes, near Digne" were excluded from the Alpine subset a priori because this locality lies geographically isolated from the main Alpine sampling region and may represent a distinct biogeographic lineage. The refined dataset, restricted to contiguous Alpine localities, shows a stronger and more consistent altitudinal trend (Fig. 10). This pattern is consistent with previous reports that population structure can obscure within-species environmental gradients^{24,31}.

Colour brightness: biological interpretation and limitations

The negative brightness–elevation correlation in the Alpine *E. aurinia* subpopulation is consistent with the thermal melanism hypothesis, under which ectotherms at cooler, higher elevations are predicted to carry more melanin to enhance solar heat absorption^{22–24}. Visual inspection of the Alpine specimens (Fig. 6) suggests that high-elevation individuals show a larger dark-patterned area while maintaining comparable overall brightness values. This is consistent with a redistribution of pigmentation – an increase in melanised area accompanied by a modest decrease in the intensity of orange pigments – rather than a uniform darkening of the wing. Under this interpretation, high-elevation individuals may absorb a comparable absolute amount of heat but distribute it more evenly across the wing surface,

potentially enabling faster warm-up times; controlled thermal imaging studies would be needed to test this mechanistically³².

Several biological and methodological factors can further contribute to heterogeneity among subpopulations. Specimen sex is not recorded in the underlying collection records; sex-specific patterns are reported in some Lepidoptera species³³, and would only influence the results here if sex ratios differ systematically along elevation. Variability in historical image-capture conditions, minor inaccuracies in older collection labels and the presence of geographically isolated lineages can also introduce variation. These factors are not artefacts of the database itself; they are inherent to the underlying collection and are documented in the digitiser's notes field where relevant, so that downstream users can apply appropriate filters.

Usage Notes

The dataset can be used as a standalone source of occurrence and phenotypic data, or combined with external resources via the stored GBIF, BOLD, EOL and Wikipedia identifiers. Darwin Core exports can be ingested directly into GBIF-compatible tooling; ABCD exports support integration with collection-management infrastructures that consume the ABCD standard.

The colour brightness measurements are image-derived and dependent on the photography setup described in the Methods. Users reusing the brightness values for comparative analyses across other datasets should be aware that absolute brightness is sensitive to illumination and camera settings. Relative comparisons within this dataset are therefore more robust than direct comparisons across datasets acquired under different conditions. The per-specimen mask statistics provided alongside the mean brightness values allow users to re-derive alternative summary measures — for example the fraction of melanised area at a chosen threshold — without re-processing the source images.

Species-level conservation-status fields reflect the EU Fauna–Flora–Habitat Directive and the Red List of Bavaria at the time of digitisation. Users applying these fields in contemporary analyses should consult the most recent revisions of these reference lists; the revision dates of the three Bavarian Red List sections are stored in the species table to facilitate this.

Specimens from the Cossidae and Geometridae families are physically present in the collection but not yet identified to species level and are not included in the current dataset. Future releases of the dataset are planned to incorporate these specimens, to extend the image-derived trait layer to additional species, and to add direct links to BOLD barcode sequences and GenBank records where available. Users interested in contributing additional trait layers (e.g. morphometrics, wing pattern metrics) can use the application's modular architecture and Docker container to deploy local instances and extend the schema.

Declarations

Competing Interests

The authors declare no competing interests.

Funding

TD wants to thank Deutsche Forschungsgemeinschaft (DFG) for funding (grant 492620490 – SFB 1583/INF).

Author Contribution

F.W. and J.B. designed and implemented the relational database and the R/Shiny web application, digitised specimen records and performed image analysis; F.W. and J.B. contributed equally as first authors. A.A. contributed to image analysis and data curation. M.W. curated the physical collection and provided taxonomic guidance. T.D. conceived and supervised the project and coordinated manuscript preparation. M.W. and T.D. contributed equally as senior authors. All authors contributed to manuscript drafting and approved the final version.

Acknowledgement

We thank the Bioinformatics Core Unit at the University of Würzburg and NFDI-Biodiversity for technical infrastructure and support. We thank the digitisation team and student assistants for their work on specimen curation, imaging and data processing. We thank Arthur Bott (1939–2022) for assembling the collection that forms the basis of this dataset, and Mirko Wölfling for its continued curation and taxonomic oversight.

Data Availability

The specimen-level dataset (5,974 records, Darwin Core and ABCD XML) and the colour brightness dataset for the 77 *Euphydryas aurinia* specimens, together with the threshold-sensitivity summary table, are deposited in Zenodo²⁹ (<https://doi.org/10.5281/zenodo.XXXXXXX>). The dataset is also accessible interactively at <https://butterflies.bioinfo-wuerz.eu>.

Code Availability

The full source code of the R/Shiny application, the SQLite schema definition, the Fiji macro used for image-derived colour brightness extraction, the reference lookup tables for the FFH Directive and the Bavarian Red List, and the Docker container definition are openly available under the MIT License at https://github.com/johannes-balkenhol/butterfly_database. The Fiji macro requires Fiji (ImageJ) v1.53 or later²⁷. The R/Shiny application was tested under R 4.3, with package versions pinned in the repository.

References

1. Short, A. E. Z., Dikow, T. & Moreau, C. S. Entomological collections in the age of big data. *Annu. Rev. Entomol.* 63, 513–530 (2018). <https://doi.org/10.1146/annurev-ento-020117-043348>
2. Cadotte, M. W., Barlow, J., Nuñez, M. A., Pettorelli, N. & Stephens, P. A. Solving environmental problems in the Anthropocene: the need to bring novel theoretical advances into the applied ecology fold. *J. Appl. Ecol.* 54, 1–6 (2017). <https://doi.org/10.1111/1365-2664.12855>
3. Meineke, E. K., Davies, T. J., Daru, B. H. & Davis, C. C. Biological collections for understanding biodiversity in the Anthropocene. *Phil. Trans. R. Soc. B* 374, 20170386 (2019). <https://doi.org/10.1098/rstb.2017.0386>
4. Montgomery, G. A. et al. Is the insect apocalypse upon us? How to find out. *Biol. Conserv.* 241, 108327 (2020). <https://doi.org/10.1016/j.biocon.2019.108327>
5. Leather, S. R. ‘Ecological Armageddon’ – more evidence for the drastic decline in insect numbers. *Ann. Appl. Biol.* 172, 1–3 (2017). <https://doi.org/10.1111/aab.12391>
6. Seibold, S. et al. Arthropod decline in grasslands and forests is associated with landscape-level drivers. *Nature* 574, 671–674 (2019). <https://doi.org/10.1038/s41586-019-1684-3>
7. Habel, J. C., Segerer, A., Ulrich, W., Torchyk, O., Weisser, W. W. & Schmitt, T. Butterfly community shifts over two centuries. *Conserv. Biol.* 30, 754–762 (2016). <https://doi.org/10.1111/cobi.12656>
8. Meineke, E. K. & Davies, T. J. Museum specimens provide novel insights into changing plant–herbivore interactions. *Phil. Trans. R. Soc. B* 374, 20170393 (2019). <https://doi.org/10.1098/rstb.2017.0393>
9. Summerville, K. S. & Crist, T. O. Structure and conservation of lepidopteran communities in managed forests of northeastern North America: a review. *Can. Entomol.* 140, 475–494 (2008). <https://doi.org/10.4039/n07-LS06>
10. Habel, J. C., Trusch, R., Schmitt, T., Ochse, M. & Ulrich, W. Long-term large-scale decline in relative abundances of butterfly and burnet moth species across south-western Germany. *Sci. Rep.* 9, 14921 (2019). <https://doi.org/10.1038/s41598-019-51424-1>
11. Habel, J. C., Segerer, A. H., Ulrich, W. & Schmitt, T. Succession matters: community shifts in moths over three decades increases multifunctionality in intermediate successional stages. *Sci. Rep.* 9, 6915 (2019). <https://doi.org/10.1038/s41598-019-41571-w>
12. Davis, C. L., Guralnick, R. P. & Zipkin, E. F. Challenges and opportunities for using natural history collections to estimate insect population trends. *J. Anim. Ecol.* 92, 237–249 (2023).

<https://doi.org/10.1111/1365-2656.13763>

13. Hardisty, A., Livermore, L., Walton, S., Woodburn, M. & Hardy, H. Costbook of the digitisation infrastructure of DiSSCo. *Res. Ideas Outcomes* 6, e58915 (2020).
<https://doi.org/10.3897/rio.6.e58915>
14. Hedrick, B. P. et al. Digitization and the future of natural history collections. *BioScience* 70, 243–251 (2020). <https://doi.org/10.1093/biosci/biz163>
15. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
16. Johnson, J. B. Observations on the common brown butterfly (*Heteronympha merope*) in the early 1900s in Australia using digitized specimens. *J. Asia-Pacif. Entomol.* 25, 101898 (2022).
<https://doi.org/10.1016/j.aspen.2022.101898>
17. Adams, J. M., Kang, C. & Wells, M. J. Are tropical butterflies more colourful? *Ecol. Res.* 29, 685–691 (2014). <https://doi.org/10.1007/s11284-014-1154-1>
18. Davis, A. K. & Rendón-Salinas, E. Are female monarch butterflies declining in eastern North America? *Biol. Lett.* 6, 45–47 (2010). <https://doi.org/10.1098/rsbl.2009.0632>
19. Suarez, A. V. & Tsutsui, N. D. The value of museum collections for research and society. *BioScience* 54, 66–74 (2004). [https://doi.org/10.1641/0006-3568\(2004\)054\[0066:TVOMCF\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2)
20. MacLean, H. J., Nielsen, M. E., Kingsolver, J. G. & Buckley, L. B. Using museum specimens to track morphological shifts through climate change. *Phil. Trans. R. Soc. B* 374, 20170404 (2019).
<https://doi.org/10.1098/rstb.2017.0404>
21. Holmes, M. W. et al. Natural history collections as windows on evolutionary processes. *Mol. Ecol.* 25, 864–881 (2016). <https://doi.org/10.1111/mec.13529>
22. Bogert, C. M. Thermoregulation in reptiles, a factor in evolution. *Evolution* 3, 195–211 (1949).
<https://doi.org/10.1111/j.1558-5646.1949.tb00021.x>
23. Clusella-Trullas, S., Terblanche, J. S., Blackburn, T. M. & Chown, S. L. Testing the thermal melanism hypothesis: a macrophysiological approach. *Funct. Ecol.* 22, 232–238 (2008).
<https://doi.org/10.1111/j.1365-2435.2007.01377.x>
24. de Souza, A. R., Mayorquin, A. Z. & Sarmiento, C. E. Paper wasps are darker at high elevation. *J. Therm. Biol.* 89, 102535 (2020). <https://doi.org/10.1016/j.jtherbio.2020.102535>
25. Ellers, J. & Boggs, C. L. Functional ecological implications of intraspecific differences in wing melanization in *Colias* butterflies. *Biol. J. Linn. Soc.* 82, 79–87 (2004).
<https://doi.org/10.1111/j.1095-8312.2004.00319.x>
26. Kingsolver, J. G. Viability selection on seasonally polyphenic traits: wing melanin pattern in western white butterflies. *Evolution* 49, 932–941 (1995). <https://doi.org/10.1111/j.1558-5646.1995.tb02328.x>
27. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9, 676–682 (2012). <https://doi.org/10.1038/nmeth.2019>

28. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675 (2012). <https://doi.org/10.1038/nmeth.2089>
29. Weber, F., Balkenhol, J., Akash, A., Wölfling, M. & Dandekar, T. The Arthur Bott Lepidoptera collection: specimen-level dataset. Zenodo <https://doi.org/10.5281/zenodo.XXXXXXX> (2026).
30. GBIF Secretariat. GBIF Backbone Taxonomy. Checklist dataset. GBIF <https://doi.org/10.15468/39omei> (2023).
31. Tuomaala, M., Kaitala, A. & Rutowski, R. L. Females show greater changes in wing colour with latitude than males in the green-veined white butterfly, *Pieris napi* (Lepidoptera: Pieridae). *Biol. J. Linn. Soc.* 107, 899–909 (2012). <https://doi.org/10.1111/j.1095-8312.2012.01996.x>
32. Miaoulis, I. N. & Heilman, B. D. Butterfly thin films serve as solar collectors. *Ann. Entomol. Soc. Am.* 91, 122–127 (1998). <https://doi.org/10.1093/aesa/91.1.122>
33. Tuomaala, M., Kaitala, A. & Rutowski, R. L. Sex-specific patterns in wing colour of the green-veined white. *Biol. J. Linn. Soc.* 107, 899–909 (2012). <https://doi.org/10.1111/j.1095-8312.2012.01996.x>

Figures



Figure 1

Representative insect boxes from the Bott Lepidoptera collection, showing pinned specimens of the family Nymphalidae. The collection spans 47 boxes with specimens dating from 1900 to 2002.



Figure 2

Example of a catalogued specimen from the Bott collection: *Iolana iolas* (Ochsenheimer, 1816) mounted on an insect pin. The attached label provides metadata such as country, state, region or city, date of collection and collector's name, which are transcribed into the specimen record.



Figure 3

*Setup for photographing pinned butterflies, exemplified by *Papilio machaon* (Linnaeus, 1758). The specimen is mounted on a pin within a standardised lighting chamber to minimise illumination variance. The setup enables consistent comparisons of specimens from different localities, altitudes or collection dates.*

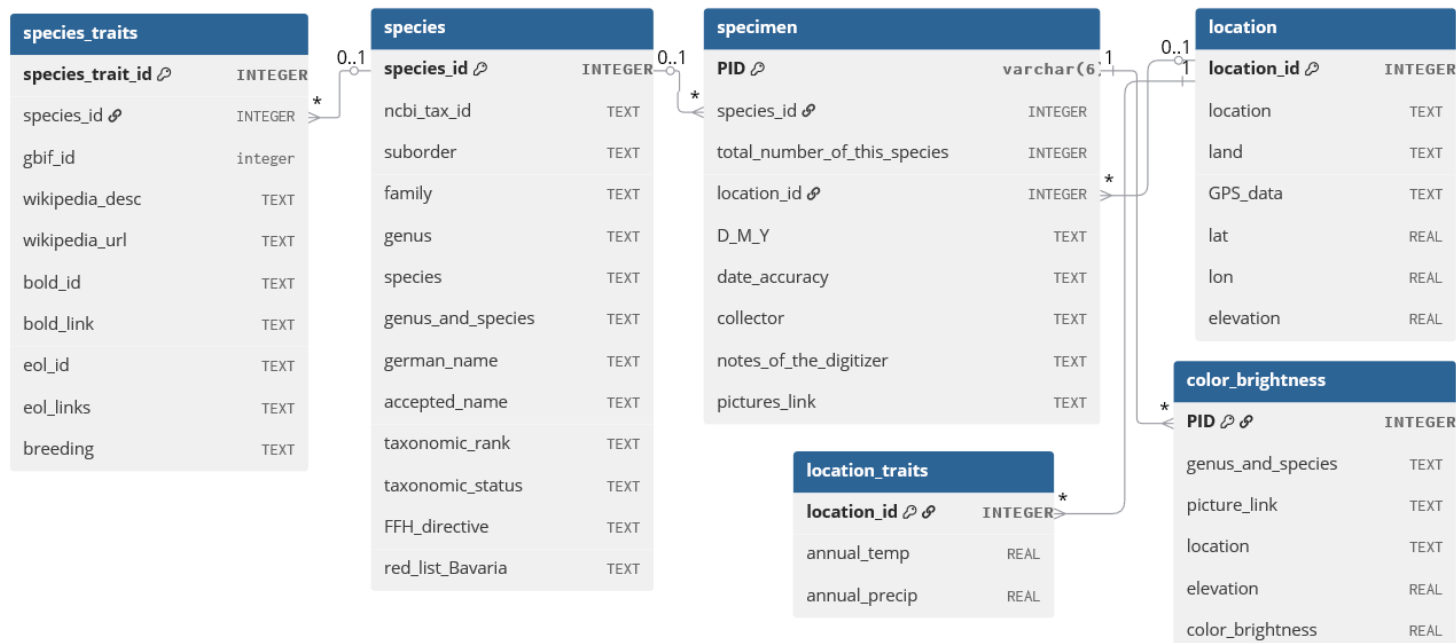


Figure 4

Relational schema of the database. Normalised SQLite database structure linking specimen records to taxonomic, spatial, trait and image-derived data. Core tables include specimen, species and location, with associated traits stored in species_traits, location_traits and colour_brightness. External resources (EOL, BOLD, GBIF, Wikipedia) are integrated via persistent identifiers.

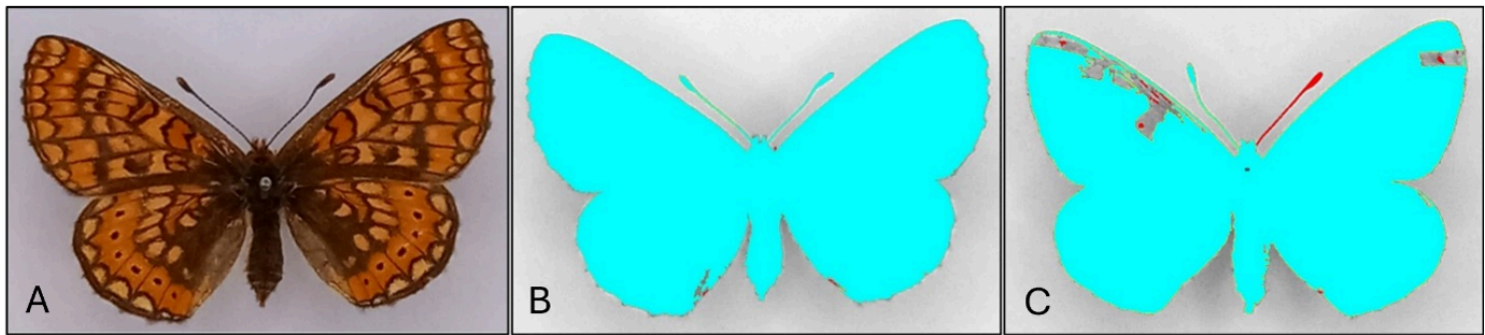


Figure 5

Mask applied for determining colour brightness in *Euphydryas aurinia* (Rottemburg, 1758). (A) binary mask defining the area of interest; (B) mask overlaid on the specimen image; (C) an example of a masking error, in which complex wing-pattern elements are incorrectly included in the background region.



Figure 6

Complexity of butterfly wing patterns motivates the use of multiple thresholds. Specimens of *Euphydryas aurinia* (Rottemburg, 1758) collected at (A) 599 m and (B) 2,300 m above sea level. (C) Mask with an elevated threshold, designed to isolate the dark (melanised) wing area as a complementary metric to overall mean brightness.

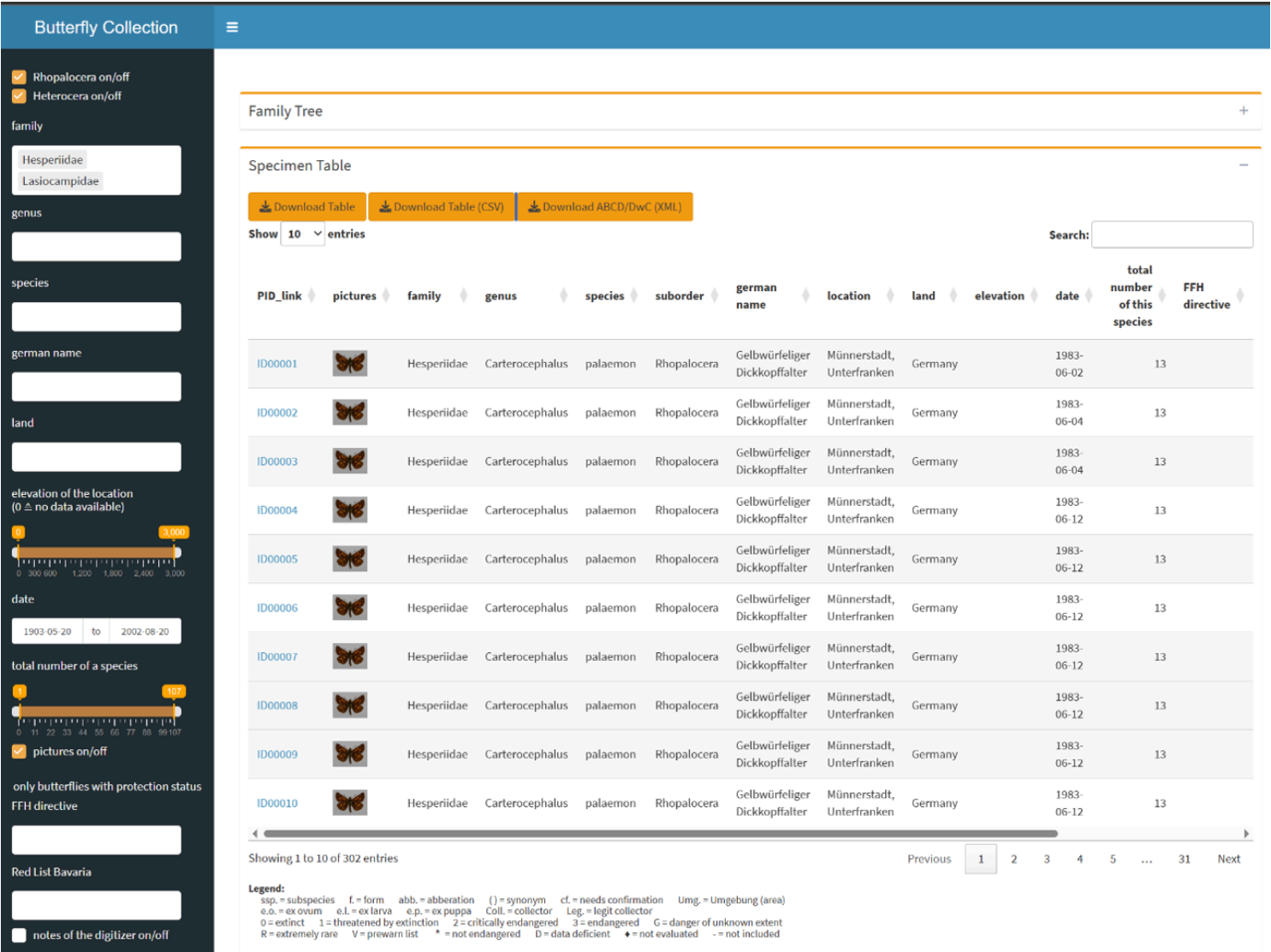


Figure 7

Interactive interface of the digitised Lepidoptera database. The left panel displays searchable filter controls. The main window (right) includes a taxonomic tree viewer, a results table with stable identifiers and representative images, a geographic map of collection localities, and an interactive analysis panel showing elevation–colour-brightness relationships, alongside the currently active filter metadata.

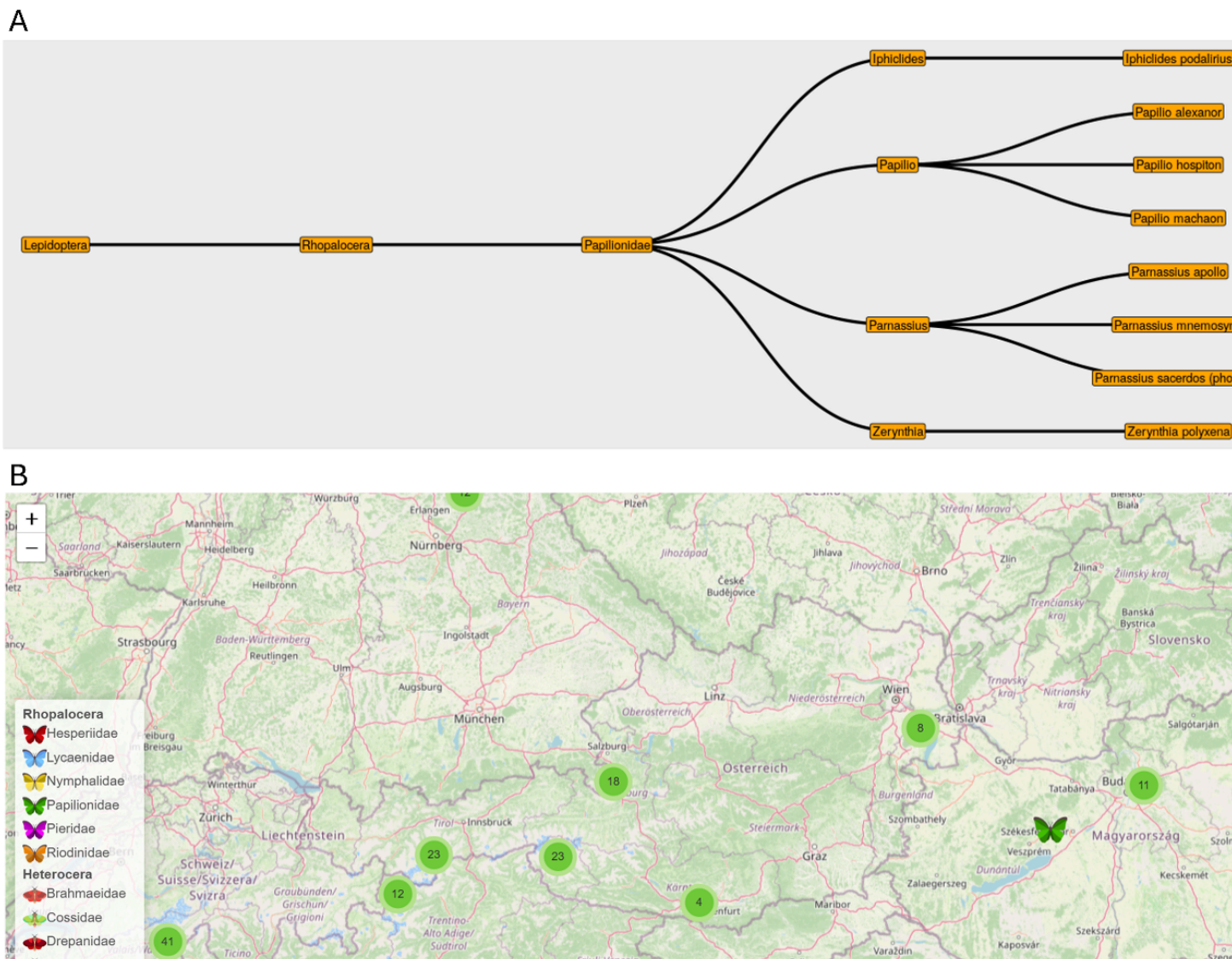


Figure 8

Features of the web interface. (A) Taxonomic tree viewer showing the hierarchy of the collection from order to species level. (B) Interactive geographic map displaying collection localities, with icons differentiating Rhopalocera and Heterocera and colour-coded by family. Clicking an icon retrieves specimen-level metadata.

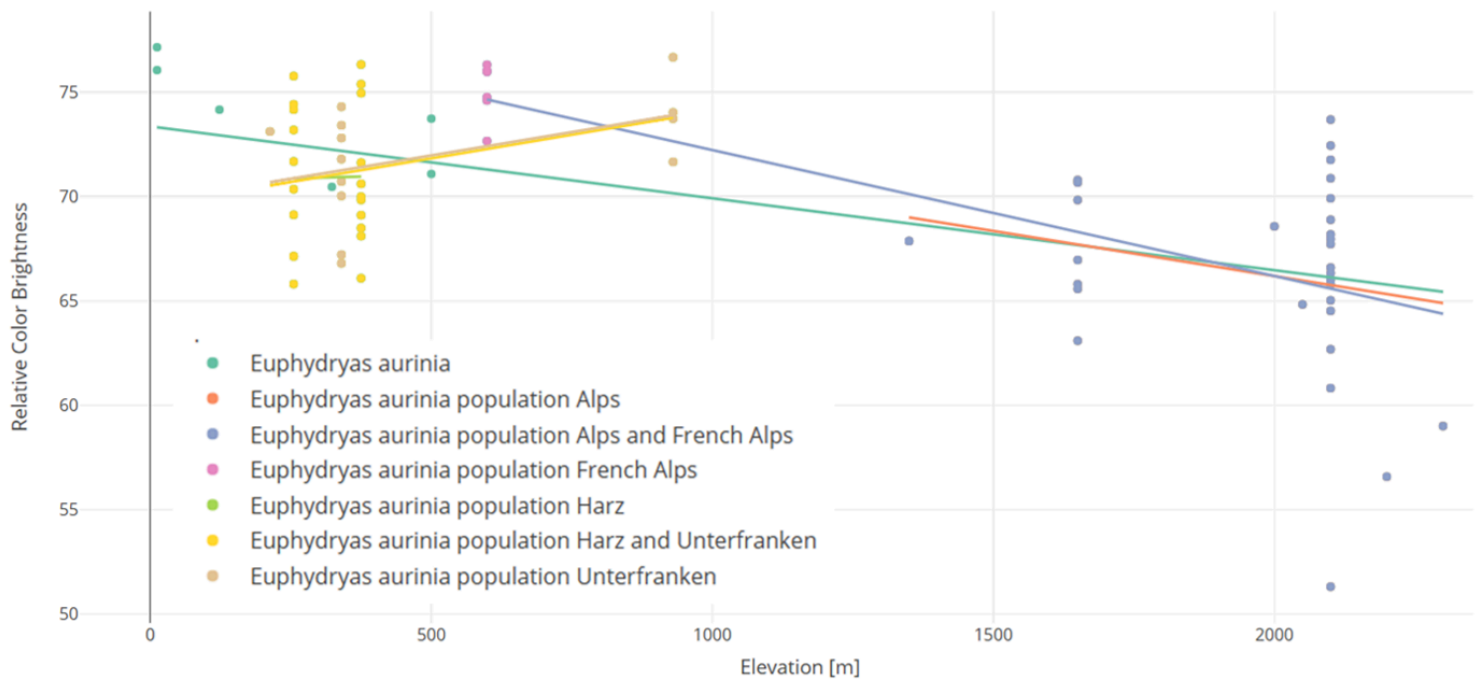


Figure 9

*Relationship between normalised colour brightness and collection elevation across Lepidoptera populations in the dataset. Scatter plots with linear regression lines show the direction and strength of altitude–brightness correlations per subpopulation. The population collection of *Euphydryas aurinia* from Alps including French Alps exhibits a significant negative correlation (blue line); other populations show weaker or less consistent trends, largely reflecting differences in sample size and elevation range (see Technical Validation).*

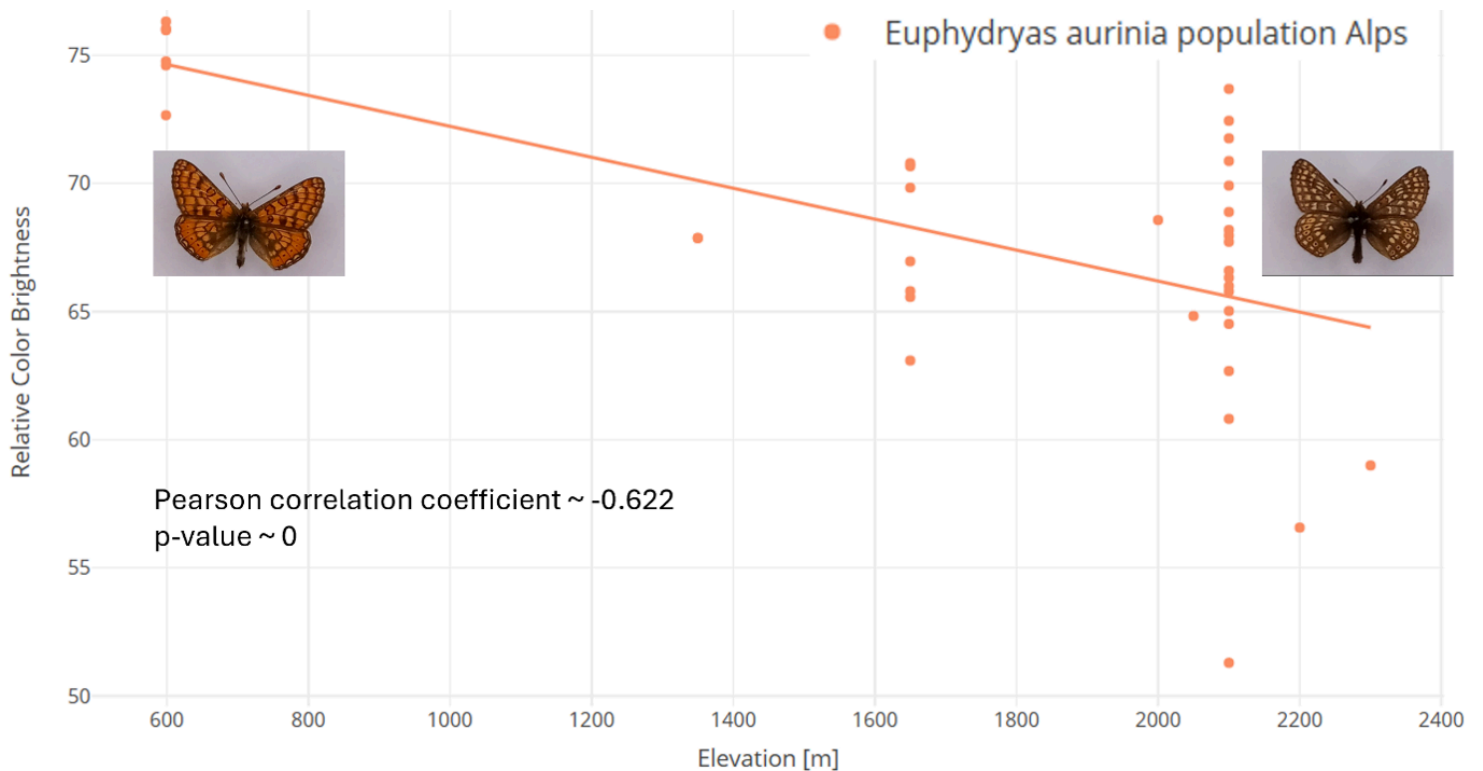


Figure 10

Relationship between colour brightness and elevation in Euphydryas aurinia populations from the Alps. Normalised colour brightness values are plotted against collection elevation for E. aurinia specimens sampled across a broad altitudinal range in the Alpine region (Basses-Alpes locality excluded; see Technical Validation). The trend line indicates a significant negative correlation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigureS1.docx](#)