DrawerDissect: Whole-drawer insect imaging, segmentation, and transcription using AI

Elizabeth G. Postema[1], Leah Briscoe[1,2], Chloe Harder[1,3], George R. A. Hancock[4], Lucy D. Guarnieri[5], Tony Eisel[1,6], Kelton D. Welch[7], Nicole Fisher[8], Christine Johnson[9], Diego Souza[1], Dexter Phillip[1], Rebekah Baquiran[1], Tatiana Sepulveda[1], Bruno A. S. de Medeiros[1]

1. Field Museum, 1400 S DuSable Lake Shore Drive, Chicago, IL 60605, USA
2. Lake Forest College, 555 N Sheridan Rd, Lake Forest, IL 60045, USA
3. Lane Tech College Prep High School, 2501 W Addison St, Chicago, IL 60618, USA
4. Centre for Ecology and Conservation, University of Exeter, Penryn, TR10 9FE United Kingdom
5. The Ohio State University, Department of Entomology, 2021 Coffey Road, Columbus, Ohio, 43201
6. Roosevelt University, 430 S Michigan Ave, Chicago, IL 60605
7. Ecdysis Foundation, Estelline, SD 57234, USA
8. CSIRO, National Research Collections Australia, Clunis Ross St, Canberra, ACT 2601, Australia
9. American Museum of Natural History, 200 Central Park West at 79th Street, NY 10024, USA

Corresponding Authors

Elizabeth G. Postema (egpostema@fieldmuseum.org)

Bruno A.S. de Medeiros (bdemedeiros@fieldmuseum.org)

INTRODUCTION

Natural history collections connect the past to the future, preserving unique ecological and evolutionary histories in specimens that can maintain their scientific value for centuries. Central to biological taxonomy (Kapun et al. 2025) and critical for taxonomic and evolutionary studies that rely on genomic data (de Medeiros et al., 2025), collections-based research is now being used to reveal impacts of environmental change (Sanders et al., 2023), elucidate ecological drivers of morphological traits (Cooney et al., 2022; Crowell et al., 2024; Holmes et al., 2016; Wu et al., 2019), and track the spread of pathogens worldwide (Colella et al., 2021). However, this work is plagued by a persistent digitization bottleneck, whereby the acquisition of metadata and specimen images is slow and labor-intensive, particularly for hyper-diverse groups, such as insects, with over 1 million estimated species (Stork, 2018; Cobb et al., 2019). Given the ecological significance of insects, and their sensitivity to anthropogenic change (Wagner, 2020; Yang et al., 2021), accelerating digitisation is increasingly urgent.

There is likely no one-size-fits all solution to digitizing insect collections, given the varied goals, sizes, infrastructures, budgets, and conventions of different institutions. Some collections manually digitize or database only specimen metadata; other techniques involve high-throughput imaging to quickly acquire metadata and basic specimen information (Welch et al., 2026; Wührl et al., 2022; Blagoderov et al., 20112; Dietrich et al., 2012; Johnson et al., 2026; Picturae; Steinke et al., 2024). Mass-imaging of specimens for research purposes has often been a secondary goal, as extracting quality information from specimen images involves considerable pre-curation and pre-processing (East et al., 2025; van den Berg et al., 2020) and images with little metadata have historically fewer uses. Advances in computer vision (CV) and other AI techniques have shifted this balance by making image-processing steps far more efficient (Borowiec et al., 2022; Johnson et al., 2026; Lürig, 2022; Pichler & Hartig, 2023; Zhao et al., 2024). With these new tools, images are not just byproducts of digitization with limited uses, but high-quality sources of data.

As computer vision tools continue to make images more fertile sources of information (Høye et al., 2021; Liu et al., 2025), one reasonable strategy for collections digitization is to prioritize bulk image capture. Whole-drawer imaging offers one of the fastest approaches to photographing the greatest number of pinned specimens with the least amount of per-specimen-handling (Holovachov et al., 2014; Mantle et al. 2012). Previous methods for parsing whole-drawer images, both automated (e.g. InSelect: Hudson et al., 2015) and non-automated (Blagoderov et al. 2012) are suitable for well-curated drawers with spatial standardization, but less effective for insect drawers that are not specifically manually arranged for imaging; the latter represents the state of the majority of insect drawers.

To answer the need for a scalable, image-focused method of specimen digitization, we developed DrawerDissect: an open-source AI-driven pipeline that uses advanced computer vision tools to provide research grade information about specimens and acquire associated metadata *en masse* (Fig. 1). DrawerDissect combines custom-trained computer vision models for image parsing with large language model prompts for text transcription at the tray and specimen levels. The pipeline works with images produced by common whole-drawer imaging methods (from cellphone panoramas to automated imaging rigs) to generate specimen-level images, specimens with the background removed ("masked" specimens), and transcribed metadata, without extensive specimen handling or standardization. DrawerDissect is designed to provide images and associated metadata organized by taxonomic identity for research, curation, and further databasing in repositories like EMu (a common collections management system) and GBIF.

To validate our pipeline, we imaged the entire Field Museum (FMNH) collection of tiger beetles (Cicindelidae), which consists of 44 drawers and over 13,000 specimens from around the world. Tiger beetles are some of the fastest moving animals on Earth, and their striking color patterns and ecological significance (Pearson et al. 2006) made them optimal organisms to validate the utility of DrawerDissect. We also present three use cases for DrawerDissect's automated outputs, primarily using the tiger beetle collection. In Use Case 1, we analyzed >300 masked specimen images of the data-rich big sand tiger beetle, *Cicindela formosa* Say, 1817, to demonstrate DrawerDissect's compatibility with rapid phenotyping projects. *Cicindela formosa* shows significant variation along geographic clines in both elytra color and pattern; by analyzing specimen images, we could determine the interaction among taxonomic, geographic and climatic drivers of color variation and the strength of those interactions (French et al., 2021; Schultz & Bernard, 1989; Schultz & Hadley, 1987; Yamamoto & Sota, 2020). In Use Case 2, we used masked specimen images of two disparate genera, the colorful Cicindela Linnaeus, 1758, and the less visually distinct rove beetle genus, Platydracus Thomson, 1858 (Staphylinidae) to demonstrate the ability of DrawerDissect to contribute to species identification (Høye et al., 2021; Shirai et al., 2022; Wührl et al., 2022). This application shows that DrawerDissect can readily supply taxonomic identification models with the large amounts of training data they require (Borowiec et al., 2022; Gao et al., 2024; Lertrusdachakul et al., 2025; Spiesman et al., 2021; Welch et al., 2026). Finally, in Use Case 3, we tested the ability of DrawerDissect to provide a country-level inventory of weevil specimens in the subfamily Apioninae (family Brentidae). For this task, we used low quality images (cellphone panoramas) to highlight DrawerDissect's capacity to accelerate labor-intensive curation tasks in less well-curated parts of insect collections.

MATERIALS & METHODS

Our workflow has three steps: (1) image drawers, (2) process images with DrawerDissect, and (3) shape data for downstream applications. For (3), we also break down three sub-workflows: (a) rapid phenotyping, (b) species identification, and (c) databasing/curation.

First, we describe our imaging set-up using a GIGAMacro Magnify2 system (Four Chambers Studio, https://gigamacro.com/), though DrawerDissect can work with alternate whole-drawer imaging methods (include cellphone panoramas; see Use Case 3). We then give a broad overview of DrawerDissect in terms of installation, configuration, and model training/selection. Finally, we describe the steps that occur during a standard DrawerDissect run, and present three example use cases for DrawerDissect's outputs.

Whole-Drawer Imaging

DrawerDissect was designed with images produced by a GIGAMacro Magnify2 imaging system, and uses FMNH conventions for insect drawers where specimens are grouped into trays ("unit trays") based on shared taxonomic identity and geographical origin (Fig. 1b-c). To make the tray-level data visible for imaging, we developed pop-up headers that display the same information; the pop-up portion is visible during imaging (Fig. S2e), but folds down flat when trays are stored. This step is optional. Most pinned insect collections are organized by taxonomy and many already use top-down visible labels affixed within unit trays, which DrawerDissect will detect and transcribe (Fig. 1c). For collections with unit tray headers affixed to the inside of unit trays, one option to make them visible from a top view and amenable to automated data retrieval supported by DrawerDissect is to lay them flat beneath their respective trays while imaging. While we describe our digitization methods based on the FMNH standard, DrawerDissect can be adapted to various configurations. For further details on our imaging set-up, see supplemental materials Section 1.1.

Overview of DrawerDissect

DrawerDissect is a Python-based pipeline available from GitHub, including documentation on installation and usage at github.com/EGPostema/DrawerDissect. The repository includes test data to familiarize new users with the pipeline's steps and organizational structure. We aimed to make the pipeline accessible to users who are not expert programmers. Only a rudimentary familiarity with their computer's command-line interface is required to run the script, and no knowledge of deep learning is necessary.

Models Used in DrawerDissect

By default, users are set up with public FMNH models via Roboflow (Dwyer at al. 2025) for image processing, while all transcription steps are done with Anthropic LLMs. All models are accessible to users in a single YAML file that can be edited as needed. By default we use

subscription-based services with API (Application Programming Interface) functionality, rather than open source models, because the former are more accessible to users that may lack the coding experience or computing infrastructure to run powerful AI models locally. However, we include the option to run all AI-based steps with the users' preferred method of deployment, including open-source LLMs. Specific instructions for configuring these deployment methods are detailed in the github's README file under the section "Deployment Options."

Training computer vision models for detection and segmentation tasks requires a set of images that have been labeled with the desired output. See supplemental Table S1 for the taxonomic groups used to train each of our most recent public models, as well as the sources of these images. Once annotated, we split all images into training, validation, and test sets at a standard ratio of 70:20:10. Training images are used by the model to learn a given task; validation images are a separate set used by the model during training to fine-tune the model and prevent overfitting; and test images are used post-training to calculate performance metrics (Borowiec et al., 2022). We also employ image augmentations to increase model generalizability. We trained our models using Roboflow servers, which can take anywhere from a few minutes to several hours depending on the size and complexity of the model. 4-5 hours was typical for our largest models with datasets of >3,000 images.

When the model is finished training, Roboflow reports the model's rates of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as well as precision (TP/(TP + FP)) and recall (TP/(TP +FN)) (Grandini et al., 2020; Table 1). We have provided public access to all the models used in our pipeline at the time of manuscript submission, including downloadable annotated datasets, at (universe.roboflow.com/field-museum). The weights files for the most recent version of each model are also included in the github repository, which are used automatically for non-API runs.

Running DrawerDissect

DrawerDissect consists of a series of steps (Fig. 1) that can be run automatically, in sequence, with a single command: *python process_images.py all*. The github documentation includes a full list of command-line steps, optional arguments for running specific step combinations, how to process specific drawer(s), and instructions on how to change models, LLM prompts, and type of deployment (i.e. through AI company servers versus the user's preferred infrastructure).

Object Detection, Segmentation and Classification

Currently, DrawerDissect uses model architectures available through Roboflow (Yolo v11, Roboflow 3.0 Object Detection, Roboflow 3.0 Instance Segmentation, and DINOv3; Table 1) to find, outline, and classify objects in images, described below. An asterisk marks models used by

default in the current version of DrawerDissect; each of these models can be replaced by user-created models as long as they are Roboflow-hosted or, if deploying locally with user-provided weights, YOLO-based.

trayfinder-base *
This model detects unit trays from drawer images (Fig. 1a).

trayfinder-popup
This model detects unit trays (with visible header labels above each tray) from drawer images.

labelfinder *
This model can detect the locations of taxonomic information, tray-level barcodes, and geocodes. This model can detect both pop-up label text and taxonomic text within trays (Fig. 1b).

bugfinder-kdn9e *
This model detects pinned insects (Fig. 1c).

bugmasker-all *
This model outlines the main body of insects, excluding legs and antennae (Fig. 1d, Fig. 2).

pinmasker *
This model outlines the specimen pin (Fig. 1d), if present.

bugcleaner *
This classification model detects the presence of visible text in specimen dorsal images.

All CV models output coordinates in a JSON file that are used by DrawerDissect to crop, measure, or mask images (with the exception of *bugcleaner*, which outputs a CSV file with a record of its classifications). Due to file size and format constraints, all TIFs are temporarily converted to JPGs before processing, and large images are resized to fit within a 1000×1000px square. During cropping, the coordinates from reduced-size images are always rescaled to the original, full-sized image.

Masking and Measuring Specimens

Coordinates from *bugmasker-all* and *pinmasker* are used to generate binary masks: black and white images (PNGs) where all background pixels are black and all body pixels are white. A filtering step finds and fills in any erroneous partial segmentations, if present, with black pixels. To create the final masked image, the white pixels clip out the full-color specimen, while the black pixels become transparent. For each specimen mask, we measure len1, which is the greatest distance (in pixels) between any two points on the outline, and len2, the maximum distance perpendicular to len1 (Fig. 1e). We also calculate the area, in pixels squared, of the body. DrawerDissect can output a map of len1 and len2 for each specimen mask if desired (Fig.

1e). Because all of our images were taken with the same combination of camera body and telecentric lens, calculating absolute measurements from pixels is straightforward.

LLM-Based Transcription: Unit Tray-Level

To transcribe different kinds of text from images, we use Anthropic's API with customizable prompts by default. However, users can use built-in toggles to deploy open-source LLMs, or a different API-based LLM, for the transcription steps. To digitize the FMNH tiger beetle collection, we used the multimodal LLM *claude-sonnet-4-6* with a maximum token limit of 600 for simple tray-level transcriptions and 15,000 for more complex specimen-level information. Both handwritten and typed material can be transcribed (Fig. 1b-c), though typed text generally results in fewer errors.

Unit tray label text is transcribed from cropped images generated by *labelfinder* (Fig. 1b). We have three separate prompts for each type of transcription (taxonomy, barcode, and geocode), which are tailored to how these different types of information are structured. These prompts can be edited by users for their specific label structure. Most users will want to keep the taxonomy-transcribing step on while turning off the barcode- and geocode-transcribing steps, as the latter are FMNH-specific conventions. We evaluated all tray-level transcriptions manually.

LLM-Based Transcription: Specimen-Level

In many specimen images, partial text from the top label is visible. We tested a novel method of metadata extraction based solely on this fragmentary information. Given the inherent variation in label text visibility from a dorsal-only photograph, our goal with this step was not to perfectly replace manual transcription (for examples of methods for fully automated transcription on a specimen-by-specimen basis, see Johnson et al., 2026; Salili-James et al., 2023). Instead, we designed this step to extract as much information as possible from the images our workflow already produces with minimal intervention.

The specimen transcription step involves both CV and LLM models, and combines multiple streams of information (Fig. 3). First, specimen dorsal images are sent to the classification model *bugcleaner* to identify images where label text is visible. Then, these images are sent to the LLM along with an image of the tray it belongs to. The LLM is instructed to transcribe each specimen's labels independently, parse the text into Darwincore (DwC) fields defined within the prompt, and then cross-reference the labels across the tray based on visual and textual cues to group similar label information. The final step is designed to mimic the way human transcribers mentally group specimens with the same label information. The end product is a CSV file that can then be reviewed and mapped to DwC-comptabile databases. DrawerDissect will attempt to fill the following DwC fields when possible: *country, stateProvince, county, municipality,*

verbatimLocality, locality, islandGroup, island, waterBody, verbatimElevation, habitat, samplingProtocol, recordedBy, verbatimEventDate, and *identifiedBy*. For text it classifies as a name but cannot determine that person's role (collector, identifier, etc.), it places the name in a non-DwC field *possibleName* for later review. All verbatim, unparsed text the LLM transcribes are also retained in the field *verbatimText*.

To evaluate the outputs of specimen-level transcription for the FMNH Cicindelidae, we analyzed a set of 1,404 specimens that had both an LLM-transcribed record and an existing human-produced (HP) record. The HP records were generated by trained interns and volunteers for separate projects. We first normalized all location fields in R (version 4.5.0) to account for inconsistencies in spelling, accenting, level of abbreviation, and language. We also standardized across known historical or language-based toponyms: Kohchun to Hengchun, Tranquebar to Tharangambadi, Norsta Auren to Norsta Aura, Poonmudi to Ponmudi, Hollandia to Jayapura, and Formosa to Taiwan. Then, for all fields except for *verbatimEventDate*, we split the normalised strings into individual words by whitespace. For each word in the HP string, we calculated the maximum Levenshtein similarity (a score derived from the minimum number of single-character edits to change one text string to another; Levenshtein, 1966) to any word in the corresponding LLM string using the R package stringdist (van der Loo, 2014). This approach is essentially a modified word-level recall metric, where instead of exact word matching (as in ROUGE recall; Lin, 2004) we use maximum Levenshtein similarity to account for acceptable transcription inconsistencies. A composite similarity score for each field was calculated as the mean of these per-word maximum similarities across all HP words, such that scores penalise missing information (HP word with no close LLM equivalent in the corresponding field) but do not penalise the LLM for providing additional detail (e.g. if the LLM records "Padre Island" when the HP record has only "Padre" in the *island* field). Scores range from 0 (no similarity) to 1 (perfect match). We evaluated the following fields, which were present in both LLM-filled and HP records: *country*, *stateProvince*, *county*, *municipality*, *locality*, *islandGroup*, *island*, *verbatimElevation*, *habitat*, *samplingProtocol*, and *recordedBy*. We evaluated *eventDate* separately by calculating the proportion of exact string matches, as even a single incorrect digit can result in an unusable record for dates. To create *eventDate* records that were comparable between LLM and HP, we normalized all *verbatimEventDate*s to YYYY-MM-DD format following DwC standards and used these values to fill an *eventDate* column.

There is no universal standard for what is an acceptable Levenshtein similarity score, as this threshold depends on the discipline, task, and user end goal. In the case of collections, even with DwC standardization, two different transcribers might input the same record in two different ways that are both acceptable for use in research or databasing. Therefore, we determined our own threshold by manually checking a random sample of fields at different similarity levels (0.50, 0.60, 0.70, 0.75, 0.80, and 1.0; n=106 records). We found that scores ≥0.75 generally represented an acceptable match. For example, this pair of *locality* records resulted in a score of

0.751: LLM =“S. slope Mt. Balabag”, HP = “Mount Balabag, South Slope, Mantalingajan Range”. While the HP record provides additional higher-level geographic information (Mantalingajan Range), the LLM captures the more critical, nested piece of information (the south slope of Mount Balabag), and the two pieces of data would ultimately result in the same set of coordinates on a map. Scores between 0.50 and 0.75 generally included some correct information but were missing critical data that were present in the HP. Below 0.50, the information was almost always too incomplete to be useful for georeferencing. For scores below 0.75, we also evaluated whether the LLM was putting accurate information in a different field than the HP. To do so, we calculated additional token-level Levenshtein similarities for the focal field against all other fields in the human-produced specimen record; if the maximum score of any of these comparisons was ≥0.75, we counted the field as misplaced.

Advanced Databasing and Archiving Functions

The final step of the default DrawerDissect pipeline (merge_data) compiles a time-stamped folder with data summaries at the drawer, tray, and specimen level for a given run. To collate data across drawers, users can run the advanced master-merging script. This produces drawer, tray, and specimen-level data for all drawers (or a specified set of drawers, e.g., all drawers with the prefix “cicindelidae_”). Data can then be further processed with scaffolding functions to prepare for manual validation and databasing. The outputs of scaffolding steps are spreadsheets organized by specimen_id (Fig 1c), pre-filled with any existing data, that human transcribers can validate and/or complete. After validation of LLM-generated metadata, and manual transcription of any missing fields, users can run formatting steps to reshape the specimen-level data for databasing in EMu and/or GBIF.

Methods for bulk-uploading data and images to these repositories varies by institution, and often involves further processing, IRN generation, and field-matching. Therefore, DrawerDissect’s current data-formatting steps are meant to produce flat, flexible datasheets that are easy for existing upload pipelines to parse. Once the specimen records and images are databased, users can use the archive function to compress projects: archiving converts the original full-drawer image, as well as all CSV/JSON outputs of inference, transcription, and measurement, into a .zip file. This step deletes all image outputs (specimen dorsals, tray maps, etc.) by default. However, these can be regenerated from the original full drawer image using the preserved JSON coordinates if the user restores the compressed file and re-runs the main DrawerDissect pipelines.

One of the many uses for DrawerDissect outputs is to train species identification models. To facilitate this process, DrawerDissect includes a specimen-sorting function that users can run after the master-merging step. This step copies masked specimen images into standard train/validation/test splits (70:20:10 ratio by default) and into subfolders by taxon.

Augmented Transcription

To investigate the potential of LLM transcription as a way to augment manual transcription, we performed timed transcriptions of the same tray for three example trays (65 specimens total) and compared the speed of human-only versus human + LLM transcription. For the former, the transcriber was given an empty DwC template with pre-filled specimen IDs. For the latter, the transcriber was given a DwC template with the same specimen IDs, as well as any fields that had been pre-filled by the LLM. The time it took the LLM to process each tray was included in the human + LLM time. We used the same transcriber for each tray, and always performed the human + LLM trial before the human-only trial to ensure that any time savings in the augmented condition could not be attributed to prior familiarity with the labels.

Pipeline Validation and Use Cases

To validate DrawerDissect, we photographed and processed the Field Museum's (FMNH's) entire 44-drawer collection of pinned tiger beetles. Then, we tested DrawerDissect's ability to integrate with existing batch color analysis pipelines (Hancock et al., 2025) by using the masked images generated by DrawerDissect to investigate geographic and climatic patterns of color diversity in two subspecies of a well-represented species in the collection, *Cicindela formosa* (**Use Case 1**). We also demonstrate the utility of DrawerDissect's masked images as training data for a species identification model (**Use Case 2**). Lastly, we investigate DrawerDissect's ability to work with cellphone-quality images of whole drawers for basic curatorial tasks (**Use Case 3**).

Use Case 1: Phenotypic analysis of *Cicindela formosa* subspecies

Due to a previous digitizing effort, the FMNH has records of the collection locations for all specimens of two *C. formosa* subspecies: *C. f. formosa* and *C. f. generosa*. To analyze phenotypic variation between and among the two subspecies, we used an existing ImageJ batch-analysis pipeline (Hancock et al., 2025) to quantitatively measure color and pattern geometry in 374 *C. formosa* specimens. For each specimen, we measured the mean for each CIELab channel (L*, a* and b*) and quantified the shape and contrast of patterning by using Gabor filters set to 4 different orientations (0, 45, 90, 135 degrees) and 4 spatial scales relative to the width of the tiger beetle (1/16x, 1/8x, 1/4x, 1/2x the width). The output for each scale and orientation were then combined to produce image statistic maps for the periodicity (spatial frequency with the highest energy for each pixel), opponent orientation vectors 'Vertical-Horizontal' (VH: 0 - 90 degrees) and 'Orthogonal-Acute' (OA: 135 - 45 degrees), and the anisotropy/directionality of patterning (VH-OA). Maps were produced for each CIELab channel and we measured both the mean and standard deviation.

As our image analysis method outputs 36 different metrics, we used principal component analysis (PCA) both to reduce the dimensionality of our dataset and to identify the main components of variation for our two tiger beetle subspecies. All statistical analyses were carried out using R version 4.3.2. To test whether or not sympatric specimens converged in phenotype, we used the latitude and longitude data for each specimen to create a binary variable for sympatry (no/yes). Specimens were considered to be sympatric if they were within 3-degrees of one another. We then used linear mixed models with PCs 1-3 as the dependent variables and both subspecies and sympatry as independent variables. Only adult specimens with longitude and latitude data were used in the analysis (n =347).

To compare the effect of latitude, mean annual temperature (celcius), and mean annual precipitation (mm) on the mean CIELab L* value (higher = lighter) and the mean CIELab a* value (higher = redder), we again used linear mixed models. Mean annual temperature and precipitation measurements were collected from a guide to North American level III ecoregions (Wiken et al., 2011); we then mapped each specimen's collection location (coordinates) to ecoregion using the Ecoregion Locator tool (https://bplant.org/). For all of our analyses, numeric variables were rescaled to Z-scores. The year the specimens were collected (to account for wear) and subspecies were used as random effects for the ecocline analyses. Residual plots were used to verify that all models met assumptions, and where required, data was transformed to match a normal distribution, e.g. square-root transformation.

Use Case 2: **Training species identification models**

We trained a single taxonomic classification model for genus, species, species groups and subspecies of two divergent beetle taxa well-curated in the Field Museum pinned collections. These are the genus *Cicindela* (Cicindelidae), including 6,611 total images of 114 species/subspecies and the genus *Platydracus* (Staphylinidae), including 2,280 total images of 60 species/subspecies. *Cicindela* is a good example of insect with variation in visually striking features, though some species are diagnosed by more detailed characters such as chetotaxy of the labrum (e.g. *Cicindela punctulata* Olivier, 1790*;* Pearson et al. 2006). *Platydracus*, on the other hand, is a genus of medium- to large-sized rove beetles that includes multiple cases of species difficult to distinguish using external characters alone. For example, species in the *femoratus* group (e.g. *Platydracus femoratus* Fabricius, 1801*, Platydracus championi* Sharp, 1884*, Platydracus sallaei* Sharp, 1884) typically require dissection for identification (A. Newton, *personal communication*). The collection of *Cicindela* in the Field Museum was last thoroughly curated in the 1990s (M. Kippenhan, *personal communication*), and most specimens are identified to species. For *Platydracus*, we used a synoptic curated research collection organized by species groups and including all species from the Americas, with specimens both from the Field Museum and loans from other institutions (A. Newton, *personal communication*). In both

cases, we have not attempted to update identifications and used species names obtained from unit trays without any taxonomic name resolution, reflecting the expert knowledge when labels were created. This is not necessarily the most recent taxonomy, but it is useful for the task we are testing for: to recognize species limits as they are set by human experts based on masked dorsal images alone.

To train and evaluate a taxonomic classification model for these genera we split all identified specimens into training, validation and test sets using a 70:20:10 ratio for each species or subspecies. Single-image taxa were only included in the training set and not in the validation or test sets. Taxa with 2 images were included in the training and testing sets but not the validation set. The validation set was used to monitor training progress while all results reported here refer to the test set. Our goal was to train a multilabel model that could simultaneously identify a given specimen to genus (*Cicindela* or *Platydracus*), species group (for *Platydracus*), species, and subspecies ("label" in this context means the taxonomic name at each of these levels). The input data to the models were individual masked beetle images with white backgrounds, associated with all genus, species group, species and subspecies labels. We primarily used the python libraries pytorch (Paszke et al., 2019), fastai (Howard & Gugger, 2020) and timm (Wightman, 2019) for deep learning, and computations were performed on 2 NVIDIA RTX A5000 GPUs with combined 48GB of memory. For the model architecture, we used EVA02 (specifically, eva02_large_patch14_448; Fang et al., 2024) as the base, one of the most accurate pretrained image classification models available through the timm library at the time of our study. Training images of all specimens were rescaled to 448x448 pixels, and, for data augmentation, we applied the following random transformations at training time: rotation, zooming, warping, brightness and contrast. We additionally employed MixUp augmentations (Zhang et al., 2018).

One of the primary challenges with any biological dataset for taxonomic identification is data imbalance: some species are very abundant and represented by hundreds of specimens in our dataset, while others are rare and represented by unique specimens. To mitigate data imbalance effects, we applied two methods from our prior work (de Medeiros et al., 2025): (1) an asymmetric loss function (Ben-Baruch et al., 2021) with focusing exponents gamma_pos and gamma_neg of 2 and clipping at 0.05 when performing multilabel classification and (2) improved regularity in multilabel training by first pre-training a single-label model with cross-entropy loss and then fine-tuning the multilabel model. During each epoch (a full pass through the training set), images were randomly sampled in batches of size 16 with higher probability for rare taxa. Specifically, the probability was proportional to the average square root of the inverse frequencies of labels associated with each image. The learning rate was obtained with fastai's learning rate finder (Smith, 2017). We initially pre-trained a single-label model with a cross–entropy loss function, sample weighting and MixUp as described above and a single classification head, assigning the most specific taxonomy label available to each specimen. For this, we used the model pretrained weights from the timm library, training for 5 epochs with

frozen body weights and 5 epochs with free weights. We then transferred the body weights to a new model, this time using two to four labels per specimen (i.e. genus, species and, when available, species group and subspecies). Given the dissimilarity between the two genera, we isolated their effects on training by modifying the model architecture. The model body used the architecture and pretrained weights from the single label regime, but we implemented 3 classification heads: one classifying the genus, one classifying species and subspecies of *Cicindela* and another classifying species group, species and subspecies of *Platydracus*. Ground-truth labels for genus-level were used to mask the loss function for each genus, so that images from a given genus only contributed to the loss of its own classification head and the genus classification head. At inference time, routing depends on the genus-level prediction, as labels are unknown. All images contributed to the loss of body weights. We additionally modified the MixUp callback to only mix images from the same genus. A cross-entropy loss function was applied to the genus-level classification (since it was a single-label task) and the asymmetric loss described above was applied to the species-level classification. This new model was trained for three epochs with frozen body weights and 30 epochs with free weights. We evaluated the final model using the test set, with a confidence threshold of 0.5. We also predicted labels for specimens in the FMNH collection that had been identified to genus but not yet to species, and thus were not used to train the model (12 for *Cicindela* and 3 for *Platydracus*) but had been imaged along with the other specimens. We then compared those predictions to new manual identifications made using dichotomous keys for *Cicindela* or to tentative identifications by A. Newton for *Platydracus*. To avoid biasing the manual identifications, these were completed prior to running inference with the classification model.

Use Case 3: **Using cellphone pictures to augment curation tasks**

To improve DrawerDissect's ability to generalize with respect to both taxonomic identity and image quality, we took cellphone panoramas (Samsung Galaxy S25; Fig. 4) of seven additional insect drawers across seven orders: Coleoptera (Cicindelidae, duplicate of a GIGAMacro-imaged drawer), Diptera (Syrphidae), Hymenoptera (Formicidae), Mantodea (Liturgusidae), Orthoptera (Gryllidae), and Phasmotodea (Diapheromeridae). We visually evaluated the initial ability of DrawerDissect to mask specimens from these drawers, retrained *bugmasker-all* with 103 new specimen dorsals (12-17 images per drawer), and re-evaluated the masks generated by the new version to track any improvements in performance (Fig. 4b-d).

Following a request received by the insect collection, we then tested whether DrawerDissect could assist in the routine curatorial task of summarizing the contents of the collection for a taxon of interest containing mostly undigitized specimens. We imaged all pinned specimens in drawers containing Apioninae (Brentidae), through the drawer glass while handholding drawers in the collection. Images were taken with a Samsung Galaxy S26 cellphone using the standard camera app (supplemental Fig. S4). We purposefully replicated the image quality expected in

this kind of routine work that requires speed rather than high-resolution data. As the photos were mainly of trays rather than full drawers, we created a new folder in the drawers directory and manually added the images directly to the "trays" subfolder. These images were then processed with DrawerDissect: only the specimen-level transcription step was toggled on, and we did not mask or measure the specimens, as the goal was simply to acquire per-Country specimen counts. We manually validated transcription accuracy for all specimens using the tray maps produced by DrawerDissect to match specimen ID to label (Fig. 1b).

RESULTS

Pipeline Validation: Digitizing the FMNH Tiger Beetle Collection

With DrawerDissect, we were able to produce a complete inventory of the Field Museum's tiger beetle collection. We processed the collection into 13,496 separate images of intact specimens obtained from 44 drawers of pinned tiger beetles from around the world (Fig. 5). All specimen photos are linked to taxonomic and biogeographic metadata. The majority of these specimens (13,484) were successfully masked and measured (length, width, and area, in mm). Our collection contains 58 genera and 663 species of tiger beetle. This represents nearly a quarter of the ~2,300 known species globally (Gough et al., 2019).

Unit Tray Cropping and Transcription

Out of 940 trays of tiger beetles, *trayfinder-popup* made no detection errors. 36 trays (3.9%) required manual transcription of one or more pieces of tray label text, which could be extracted quickly from tray images. For the majority of trays, missed transcriptions occured when *labelfinder* failed to detect text, which is automatically flagged by DrawerDissect in the final data-merging step. In four instances, label information was detected, but the bounding box failed to fully encompass the text. We observed only two instances of taxonomic transcription typos: *Manticora* misspelled as "*Mantichora*", and *Distipsidera* misspelled as "*Distintipsidera*". For barcodes, there were three instances where the numbers 5 and 6 were erroneously switched. For geocodes, there were 9 transcription errors that resulted in the incorrect biogeographical realm. All tray-level text for these drawers was typed.

Specimen Detection and Masking

Overall, *bugfinder-kdn9e* made 13,662 specimen detections. 153 (1.1%) of these were false positives, where non-specimens were erroneously detected. This is a higher rate of false positives than the model's roboflow test set (0.02%; Table 1). No specimens were missed by the model, meaning the rate of false negatives was 0% (lower than the Roboflow-calculated rate of 0.02%). Thus, *bugfinder-kdn9e* is more likely to detect non-specimens than to miss true specimens.

Out of the 13,662 cropped specimen photos generated from *bugfinder-kdn9e*, *bugmasker-all* found masks for 13,501 images (supplemental Fig. S5). Of the 161 images where no mask was found, 154 were true negatives, meaning they did not contain a specimen, or contained a damaged specimen. Therefore, DrawerDissect's masking steps can help to mitigate false positive errors in specimen detection. Only 7 were false negatives, where no mask was found despite the image containing a complete, intact specimen. In the masked set, we found 12 false positives, where a mask was applied to a non-specimen (3) or heavily damaged specimen (9). In 5 cases a mask did not fully outline an intact specimen. Our actual rates of false positives (0.09%) and false negatives (0.05%) for our set of tiger beetles were slightly higher than the rates reported for the model using the test set in Roboflow (0.02% for both; Table 1).

Specimen-level Label Transcription

Bugcleaner filtered out 5,270 specimen images with no visible text, sending 8,089 specimens to the LLM for transcription. The LLM filled in a total of 17,264 fields (~2/specimen) across 12 DwC categories. For our sample of 1,404 specimens with matching human-produced (HP) records (Fig. 6a), the LLM filled in 1,422 fields. The detection rate (# of LLM filled fields / # HP filled fields for matched records) was 18.2%; the false positive rate (# of LLM filled fields / # HP empty fields for matched records) was 1.8%; and the rate of false negatives (# of LLM empty fields / # HP filled fields for matched records) was 81.8%.

For *eventDate*, only 8 LLM-filled fields were a perfect match to the HP (14.8% of LLM-filled fields); 13 captured year accurately but did not retrieve month or day (24.1%); and 33 resulted in an incorrect date (61.1%). The mean word-level Levenshtein similarity score across the other 11 DwC fields where both the LLM and HP contained information was 0.886 ± 0.007 (Fig. 6b). Overall, 982 LLM-filled fields (82.6%) achieved a similarity score ≥ 0.75 when compared to the corresponding HP-filled field. There were 131 cases (9.2% of LLM-filled fields) of information placed in a field that differed from the HP, which we considered "misplaced" (Fig. 6a).

Augmented Transcription

The average time gain across the three trays (65 specimens) was 3 minutes 50 seconds, or ~8.2 seconds per specimen (Table 2). For two of the three trays, the combined human-LLM approach was faster than human-alone: 8m 25s saved for a 33-specimen tray, and 3m 05s saved for a 20-specimen tray. The smallest (12-specimen) tray resulted in no time gains as the LLM did not pre-fill any DwC fields for this tray.

Size Measurements

We were able to measure all masked specimens (13,484), generating length, width, and area in mm/mm^2. To assess the accuracy of the digital measurements, we compared them against a random subset of specimens (n = 79) that we hand-measured using calipers (Fig. 7a-b). We found that the digital and manual measurements were strongly correlated (length: $R^2 = 0.989$, width: $R^2 = 0.961$). Body area, length, and width varied both within and across genera (Fig. 7c).

Total Time and Cost

Imaging all 44 drawers took ~70 hours, including moving drawers, inventorying unit tray labels, and printing and replacing header labels, with the latter two steps being optional. On a Windows computer with an AMD RyzenTM 7 Processor 7800X3D CPU and 32GB of RAM, DrawerDissect is able to fully process an 8GB drawer image in 16 minutes. For 44 drawers it took ~10 hours to fully process, with the advantage of parallel processing. This translates to approximately 21.4 seconds per specimen from the beginning of imaging to the final data-merging step (Table 3). The total cost of LLM calls and Roboflow API calls to process all 13,484 Cicindelidae specimens was $650 (Table 3).

The cost for whole-drawer imaging and DrawerDissect processing will depend on the deployment method used. LLM transcription and computer vision processing costs are highly variable, as DrawerDissect supports both API calls to paid AI services and the use of open-source (free) models (including access to our own CV model weights). For non-API approaches, the cost will ultimately be determined by the cost of the user's computing infrastructure. At FMNH, we currently use Anthropic models for LLM steps, which do not require a subscription for API calls; this method is relatively inexpensive ($50 for 44 drawers; Table 3). For computer vision tasks, we paid $600 for a mid-tier Roboflow account, which provides enough credits for processing a large number of insect drawers. Processing 44 drawers of Cicindelidae took 42,712 total inferences, which is only 7% of all inferences available annually at that tier. Additionally, Roboflow provides a free tier that includes 360,000 annual inferences. This means that a user could process >170,000 specimens, annually, at no cost. As of submission of this manuscript, the platform additionally offers free research credits (https://research.roboflow.com/credits).

***Use Case 1*: Results of Phenotypic Analysis**

In our PCA, 53% of the variation in color and pattern metrics were accounted for by PCs 1-3. PC1 (22%) primarily consisted of CIE L* (luminance), with higher values indicating lighter, more contrasting maculations; PC2 (19%) consisted of CIE a* (red-green) with higher values indicating redder coloration; and PC3 (12%) consisted of patterns with lower spatial frequency. Both subspecies varied considerably in PC1 (Fig. 8a-b). For PC2, *C. f. generosa* were less red than *C. f. formosa* ($\beta = -5.12$, SE = 0.369, $t\text{-value}_{304} = -13.93$, $p=<0.001$). For PC1, sympatry had a significant effect ($\beta = 1.68$, SE = 0.62, $t\text{-value}_{287} = 2.72$, $p=0.007$), likely due to increased

maculation size in northern regions where populations overlapped (Fig. 8a). Individuals of both subspecies were less red in sympatric regions (β = -2.74, SE = 0.430, t-value$_{336}$ = -6.37 , p=<0.001), but less so for *C. f. generosa* (β = 2.77, SE = 0.604, t-value$_{342}$ = 4.60 , p=<0.001). For PC3, no significant differences were observed between or within subspecies. Luminance was not significantly affected by any of the geographic or environmental factors included in our analysis, while redness increased in regions with lower mean annual precipitation (β = -0.53, SE = 0.123, t-value$_{245}$ = -4.33 , p=<0.001) and higher mean temperature (β = 0.53, SE = 0.133, t-value$_{259}$ = 3.97 , p=<0.001) (Fig. 8c).

Use Case 2: **Results of Species Classification Model**

Our model classified genus with 100% accuracy. Species and subspecies of both genera could be identified with high accuracy even for modest sample sizes in the training set (Fig. 9c). The average F1 score, which balances precision and recall, increases with availability of samples in the training set (Fig. 9b), with *Cicindela* showing generally higher scores than *Platydracus* for a given number of training specimens. Both asymptote above 90% for more than ~10 specimens (*Cicindela*) or ~20 specimens (*Platydracus*) per label. Lower F1 scores for labels with small sample size are generally caused by low recall instead of low precision (Fig. 9c). Interestingly, the *Platydracus* species group *femoratus*, which includes particularly similar species generally requiring dissection for identification (A. Newton, personal communication), showed similar results, with only one case of very low F1 score due to 0% recall over the two samples available in the test set (*Platydracus pluviasilva* Newton, 1974). Specifically, we observed 97.2% precision and 93.7% recall for species when averaged over specimens for *Cicindela*, and 93.6% precision with 88.3% recall for *Platydracus* (Fig. 9c). When averaged over species labels rather than specimens, these numbers decrease to 96.9% precision and 77.1% recall for *Cicindela* and 90.1% precision and 62.3% recall for *Platydracus* (Fig. 9c). Performance was slightly better for *Platydracus* species groups and slightly worse for subspecies (Fig. 9). This means that most specimens can be predicted with very high accuracy, while the model may fail to recognize rare species and subspecies.

Using the model to identify 12 *Cicindela* samples in the collection that had not been identified to species previously confirms this pattern (Table 4). Among those specimens, 10 were common species that were predicted correctly (Table 4). One sample that we identified as *Cicindela sylvicol*a Dejean, 1822, was misclassified as *Cicindela hybrida* Linnaeus, 1758. The trained model has a moderate accuracy in identifying *C. sylvicola* (75% precision, 60% recall), so this result was a true misclassification. The final specimen elicited no species-level prediction by the model. We determined that this was a subspecies (*Cicindela transbaicalica japanensis* Chaudoir, 1863) not found elsewhere in the FMNH collection and therefore not present in the training set. The model correctly failed to make a species prediction in this case. The model was also able to make closely matching identifications for three *Platydracus* specimens in the *femoratus* species

group that has been tentatively marked as either *Platydracus championi* or *Platydracus sallaei* in the collection. The model labeled two of the specimens as either *Platydracus championi* or *Platydracus femoratus,* and the last specimen as belonging to the *femoratus* group (but did not provide a species label; Table 4).

Use Case 3: Results of Cellphone Whole-Drawer Imaging

Adding 12–17 cellphone images (median 14) was sufficient to increase masking success for a given taxon (mean +48.4% ± 8.6% usable masks; Table 5). No additional training for *bugmasker-all* was required to evaluate the Apioninae request, as these specimens did not need to be masked. *Bugfinder-kdn9e* (v21) was able to detect 596 out of 606 total specimens, and provided accurate country information for 533 specimens (supplemental Table S2). The specimens that lacked country information were either not transcribed (n=63) or not detected (n=10).

DISCUSSION

Using DrawerDissect, we imaged and processed the whole tiger beetle collection (Cicindelidae) at the Field Museum. This included 13,484 pinned insect specimens at a rate of ~21.4 seconds and ~$0.50 per specimen (from the start of imaging to the end of DrawerDissect's full set of processing steps; Table 3) with 97.27% recall and 97.52% precision averaged across computer vision models (Table 1). All specimens include individual dorsal images with corresponding tray maps, masks, taxonomic identity, biogeographical region, and basic measurements (Fig. 1b-e). We successfully used these outputs for phenotypic analysis (Fig. 8), species identification (Fig. 9; Table 4), and a subfamily inventorying request (supplemental Table S2; Fig. S4).

A key innovation of DrawerDissect is its ability to produce high-quality images of specimens excluding both the background and overlapping specimens (Fig. 2). Both are significant challenges for bulk insect imaging (East et al., 2025) and necessary for efficient image-based morphological analyses (Correa-Carmona et al., 2025; Hancock et al., 2025; Mráz et al., 2023; Van Belleghem et al., 2018; van den Berg et al., 2020; Weller et al., 2024). DrawerDissect's masking step is capable of segmenting pinned insects of various shapes, sizes, colors, orientations, and degrees of overlap with other specimens (Fig. 2). Updating existing models to work with new taxa and imaging set-ups requires minimal additional data. For example, while *bugmasker-all* was originally trained on various Coleopteran families (Fig. 2a-b) primarily photographed with a GIGAMacro Magnify2 imaging rig, we were able to retrain the model to segment specimens of six new insect families (from six different orders) with a remarkably small number of cellphone photographs (Fig. 4a). Adding fewer than 20 new images per taxon to the training set (12-17, median = 14) translated to significant improvements in the model's ability to mask specimens from these taxa (Fig. 4b-d, Table 5).

DrawerDissect is also capable of specimen-level data capture. We retrieved at least one DarwinCore metadata field for 5,947 specimens (44.1% of the Cicindelidae; Fig. 6a). Evaluated against a subset of 1,404 human-produced (HP) specimen records, the LLM's transcriptions were generally a close match (mean word-level Levenshtein score across fields = 0.886 ± 0.007, where scores >0.75 represent an acceptable match; Fig. 6b). This pattern was mainly driven by four well-populated (Fig. 6a) DwC fields that had high similarity scores between the LLM and HP records: *country* (0.981 ± 0.006), *county* (0.950 ± 0.022), *stateProvince* (0.940 ± 0.013), and *recordedBy* (0.750 ± 0.029). Assigning information to the wrong DwC field was uncommon (9.3% of LLM-filled records) and reflected the tendency for the LLM to use *locality* as a catch-all for more specific fields like *stateProvince, municipality*, or *island,* as well as using *island* or *islandGroup* where *country* would be more appropriate (e.g. the Philippines; supplemental Fig. S6). Further prompt refinement or post-processing scripts would likely resolve many of these discrepancies, though human transcribers may mix up fields in similar ways. The prompt could also be modified to reference an existing library of common locations or collector names to further improve metadata recovery for these fields. The LLM performed notably poorly in capturing *eventDate*: only 14.8% of LLM-transcribed dates were accurate when compared to HP records. This was likely because a single missed or mistranslated character in a date string results in a fundamentally different date, whereas fragmentary location information can often be resolved with surrounding context. Additional data-cleaning scripts to ensure that the numbers being returned are reasonable dates (for example, automatically standardizing to a YYYY-MM-DD format, and returning blank cells for unparseable dates) would likely reduce the number of errors for this field.

One limitation of whole-drawer or whole-tray imaging is the inconsistent visibility of specimen-level label information. Other methods for high-throughput digitization avoid this problem by removing labels and laying them flat (Johnson et al., 2026), or taking images from multiple angles (Salili-James et al., 2023). Here, we show that multimodal large language models (LLMs) can augment the transcription process (Stenhouse & Thrall, 2025) without label removal or additional imaging by generating DwC-compatible datasets with all visible fields pre-filled. Dorsal-only transcription adds little in terms of time and cost (Table 2), and is largely accurate when sufficient label text is visible (Fig. 6b). Though metadata retrieval was incomplete for the majority of specimens, the pre-filled fields provided by LLM transcription augmented manual data transcription for a small set of example trays. Manually completing pre-filled metadata sheets was ~4min faster per tray (or ~8.2s/specimen) compared to filling in the same set of fields with no LLM assistance (n = 3 trays, 65 specimens). The greatest time gain came from the most densely packed tray (33 specimens; Table 2). Depending on the density of each tray and the proportion of specimens with visible label text, this could translate to hours of time saved if deployed at scale. As insect collections tend to have more metadata records than specimen images (Cobb et al., 2019), another fruitful strategy may be to target previously digitized groups

that lack images (>40 million specimens worldwide on GBIF). DrawerDissect outputs spatial maps (Fig. 1c) to facilitate matching specimen images to existing records.

As a demonstration of this approach (Use Case 1), we combined 374 masked images of two subspecies of the big sand tiger beetle (*Cicindela formosa*) with previously digitized metadata to analyze this group's color diversity (Fig. 8). The results of this analysis confirm the ability of the pipeline to extract meaningful phenotypic data, as there were significant differences in the two subspecies' color patterns that matched differences described in field guides (Pearson et al., 2006). However, we also found that the color traits used to distinguish the two subspecies (maculation lightness [PC1] and elytra redness [PC2]; Fig. 8b) show convergence in sympatric populations, as well as significant relationships with abiotic variables (Fig. 8c). This result raises further questions about the genetic basis of this subspecies delineation: are the two *C. formosa* subspecies genetically distinct with convergent morphologies in similar environments (as is the case with *C. formosa gibsoni* Brown, 1940 relative to *C. formosa gaumeri* French, 2021; French et al., 2021), have the sympatric populations hybridized, or some combination of these mechanisms? We show this example case to highlight how sampling broadly across taxa and deeply within each taxon is essential for teasing apart the complex phylogenetic and environmental factors that shape phenotypic diversity (Pennell et al., 2016; Postema et al., 2023). DrawerDissect readily produces inter- and intraspecifically rich datasets, making it a suitable tool for investigating the evolution of diverse insect traits at a scale that has previously only been feasible for less speciose groups of animals (Cooney et al., 2022; Crowell et al., 2024).

While there are limits to the types of data that can be extracted from dorsal-only images and masks, such images can be surprisingly useful for tasks that traditionally require more information, like species identification (Fig.9; Table 4). For example, we trained an image classification model (Use Case 2) for the taxonomic identity of 158 species and 81 subspecies in the genus *Cicindela* housed at the Field Museum (~19% of global species diversity; Vogler et al., 2005), as well as 62 species and 3 subspecies in a synoptic collection of the rove beetle genus *Platydracus* (~21% of of global species diversity; Newton, 2022). The model performed well for both *Cicindela* (precision: 96.9%, recall: 77.1%) and *Platydracus* species (precision: 90.1%, recall: 62.3%). For both genera, the model returned no prediction more often than an inaccurate prediction for a species it could not determine, which is optimal behavior. Scarce data generally limits species-level classification models, but ours is comparable in performance to classifiers for beetle genera (Lertrusdachakul et al., 2025) and bumble bees (Spiesman et al., 2021).

The strong performance of our model is likely not an artifact of the model recognizing irrelevant parts of the image, as the background of each image was removed and specimens were photographed under standardized lighting conditions. Instead, the model is successfully learning to associate visual traits on the dorsal aspect of each specimen with a taxonomic identity, irrespective of the characters a human identifier using a key might use. This is less surprising for

Cicindela species, as features of the elytra are commonly used to distinguish many (but not all) species in this group. For *Platydracus*, the model's accuracy is impressive given that the visual characters used for this genus are generally more subtle, may come from the ventral side of the organism, and certain species (ex. *femoratus* group) rely on dissection rather than characters visible to the naked eye (A. Newton, *personal communication*). Therefore, DrawerDissect is likely capable of producing datasets for taxonomic identification models for a variety of insect groups; even those that are challenging for a human to identify. We note that expert human taxonomists are also often able to make correct assignments without relying on characters typically used in keys and diagnoses, which implies subtle characters exist and may be learned by vision models. Such models could be used to augment the identification specimens in existing collections (Table 4), to work through the massive backlogs produced by biodiversity surveys (Welch et al., 2026; Wührl et al., 2022) or even improve models designed for field identification (given appropriate pre-processing, e.g. copy-paste augmentations: Mesnage et al., 2025; Sun et al., 2021).

DrawerDissect users should carefully consider what image capture method will work best for them, given the taxonomic group(s) they work with, the end goal of the workflow (Fig. 10), and the inherent constraints of top-down imaging. For example, we observe a tradeoff between specimen size and data type: large specimens provide higher-resolution phenotypic data (Fig. 8) at the cost of automatic retrieval of label metadata (Figure 5), while the balance shifts for small specimens in which the label is more visible (supplemental Fig. S4, Table S2). High-quality images that are best suited for phenotypic analysis (Fig. 10a) and species identification (Fig. 10b) may require more expensive equipment, though there are many options for whole-drawer imaging at a variety of price points (Holovachov et al., 2014). Cellphone images are low-cost and helpful for streamlining simple curation tasks as we show with Use Case 3 (Fig. 10c; supplemental Fig. S4), but may be less tenable for morphological studies of tiny specimens.

Natural history collections represent centuries of human effort. The specimens that comprise these collections are valuable not just as a record of the world's biodiversity, but as the foundation for impactful research programmes across a variety of scientific disciplines (Holmes et al., 2016; Kapun et al., 2025; Sanders et al., 2023). We designed DrawerDissect to make accessing, distributing, and preparing these specimens for research as efficient as possible, with a focus on images and the data that can, increasingly, be extracted from them via computer vision (East et al., 2025; Høye et al., 2021). While our pipeline can already perform a wide range of useful functions for research and curation (Fig. 6-9), we also intend for DrawerDissect to be an entry point for other image analysis pipelines (Fig. 10a; Lürig, 2022; Mráz et al., 2023; Schwartz et al., 2026.; Welch et al., 2026). For example, trays of Coleoptera specimens produced by DrawerDissect could be further processed with BeetleFlow (Liu et al., 2025) for more fine-grained morphological segmentation and measurement; similarly, DrawerDissect-generated Lepidoptera dorsals could be analyzed with LEPY (Correa-Carmona et al., 2025) to extract

information about specimen wing morphology. In terms of taxonomic identification, DrawerDissect's masked images may be useful as a source of training data for ID models that can reduce identification bottlenecks in biodiversity inventories (East et al., 2025; Pollock et al., 2025; Welch et al., 2026) or backlogs of unsorted specimens in collections (Table 4) when time or expertise is limited. Finally, within DrawerDissect, there are still steps that could be made more efficient via automation, such as the detection of color standards for automated color-correction, or a georeferencing function that automatically assigns coordinates to location metadata.

ACKNOWLEDGEMENTS

We thank the FMNH Collaborative Imaging Lab and its manager, Stephanie Ware, for imaging space and support. We also thank Gene Cooper, Jacob Witt, Janeen Jones, and Matt Wadleigh for their invaluable technical assistance and programming advice. We are grateful to Rick Ree for his insightful comments on the manuscript, to Tom Schultz for his expertise on tiger beetle coloration, Al and Margaret Newton for their help with Staphylinid identification, to Maureen Turcatel for help with collection management, to several FMNH volunteers (Francis Dela Pena, Ian Chiu, and Lily Bazan) for specimen barcoding, and to Matt Buffington for early discussions on the value of whole-drawer imaging. Finally, we thank our reviewers for their excellent suggestions for improving the manuscript, as well as the countless curators and collectors of the collections we imaged, who have made this work possible.

REFERENCES

Ben-Baruch, E., Ridnik, T., Friedman, I., Ben-Cohen, A., Zamir, N., Noy, A., & Zelnik-Manor, L. (2021). Multi-label classification with partial annotations using class-aware selective loss (arXiv:2110.10955). arXiv. https://doi.org/10.48550/arXiv.2110.10955

Blagoderov, V., Kitching, I., Livermore, L., Simonsen, T., Smith, V.S. (2012). No specimen left behind: industrial scale digitization of natural history collections. Zookeys, 209, 133-146. 10.387/zookeys.209.3178

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022). Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution, 13(8), Article 8. https://doi.org/10.1111/2041-210X.13901

Cobb, N. S., Gall, L. F., Zaspel, J. M., Dowdy, N. J., McCabe, L. M., & Kawahara, A. Y. (2019). Assessment of North American arthropod collections: Prospects and challenges for addressing biodiversity research. PeerJ, 7, e8086. https://doi.org/10.7717/peerj.8086

Colella, J. P., Bates, J., Burneo, S. F., Camacho, M. A., Bonilla, C. C., Constable, I., D'Elía, G., Dunnum, J. L., Greiman, S., Hoberg, E. P., Lessa, E., Liphardt, S. W., Londoño-Gaviria, M., Losos, E., Lutz, H. L., Garza, N. O., Peterson, A. T., Martin, M. L., Ribas, C. C., … Cook, J. A. (2021). Leveraging natural history biorepositories as a global, decentralized, pathogen surveillance network. PLOS Pathogens, 17(6), e1009583. https://doi.org/10.1371/journal.ppat.1009583

Cooney, C. R., He, Y., Varley, Z. K., Nouri, L. O., Moody, C. J. A., Jardine, M. D., Liker, A., Székely, T., & Thomas, G. H. (2022). Latitudinal gradients in avian colourfulness. Nature Ecology & Evolution, 6(5), 622–629. https://doi.org/10.1038/s41559-022-01714-1

Correa-Carmona, Y., Böttger, D., Korsch, D., Holzmann, K., Alonso-Alonso, P., Pinos, A., Yon, F., Keller, A., Steffan-Dewenter, I., Peters, M., & Brehm, G. (2025). LEPY: A Python-Based Pipeline for Automated Morphological Trait Analysis of Lepidoptera Images. https://doi.org/10.32942/X2WS78

Crowell, H. L., Curlis, J. D., Weller, H. I., & Davis Rabosky, A. R. (2024). Ecological drivers of ultraviolet colour evolution in snakes. Nature Communications, 15(1), 5213. https://doi.org/10.1038/s41467-024-49506-4

de Medeiros, B. A. S., Cai, L., Flynn, P. J., Yan, Y., Duan, X., Marinho, L. C., Anderson, C., & Davis, C. C. (2025). A composite universal DNA signature for the tree of life. Nature Ecology & Evolution, 9(8), 1426–1440. https://doi.org/10.1038/s41559-025-02752-1

Dietrich, C., Hart, J., Raila, D., Ravaioli, U., Sobh, N., Sobh, O., & Taylor, C. (2012). InvertNet: A new paradigm for digital access to invertebrate collections. ZooKeys, 209, 165–181. https://doi.org/10.3897/zookeys.209.3571

Dwyer, B., Nelson, J., Hansen, T. et al. Roboflow (Version 1.0) [Software]. https://roboflow.com (2025).

East, A., Campolongo, E. G., Meyers, L., Rayeed, S. M., Stevens, S., Zarubiieva, I., Fluck, I. E., Girón, J. C., Jousse, M., Lowe, S., Perry, K. I., Betancourt, I., Charney, N., Donoso, E., Fox, N., Landsbergen, K. J., Nepovinnykh, E., Ramirez, M., Singh, P., … Record, S. (2025). Optimizing image capture for computer vision-powered taxonomic identification and trait recognition of biodiversity specimens. Methods in Ecology and Evolution, 16(10), 2260–2275. https://doi.org/10.1111/2041-210X.70140

929 French, R., Bell, A., Calladine, K., Acorn, J., & Sperling, F. (2021). Genomic distinctness
930 despite shared color patterns among threatened populations of a tiger beetle.
931 Conservation Genetics, 22, 1–16. https://doi.org/10.1007/s10592-021-01370-1
932 Gao, Y., Xue, X., Qin, G., Li, K., Liu, J., Zhang, Y., & Li, X. (2024). Application of machine
933 learning in automatic image identification of insects—A review. Ecological Informatics,
934 80, 102539. https://doi.org/10.1016/j.ecoinf.2024.102539
935 Global Biodiversity Information Facility. GBIF Occurrence Download
936 https://doi.org/10.15468/dl.5c629n (2025).
937 Gough, H. M., Duran, D. P., Kawahara, A. Y., & Toussaint, E. F. A. (2019). A comprehensive
938 molecular phylogeny of tiger beetles (Coleoptera, Carabidae, Cicindelinae). Systematic
939 Entomology, 44(2), 305–321. https://doi.org/10.1111/syen.12324
940 Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: An
941 Overview (arXiv:2008.05756). arXiv. https://doi.org/10.48550/arXiv.2008.05756
942 Hancock, G., Cuthill, I., & Troscianko, J. (2025). Shining a light on camouflage evolution: Using
943 genetic algorithms to determine the effects of geometry and lighting on optimal
944 camouflage. https://doi.org/10.1101/2025.03.01.640995
945 Holmes, M. W., Hammond, T. T., Wogan, G. O. U., Walsh, R. E., LaBarbera, K., Wommack, E.
946 A., Martins, F. M., Crawford, J. C., Mack, K. L., Bloch, L. M., & Nachman, M. W.
947 (2016). Natural history collections as windows on evolutionary processes. Molecular
948 Ecology, 25(4), 864–881. https://doi.org/10.1111/mec.13529
949 Holovachov, O., Zatushevsky, A., & Shydlovsky, I. (2014). Whole-Drawer Imaging of
950 Entomological Collections: Benefits, Limitations and Alternative Applications. Journal of
951 Conservation and Museum Studies, 12(1). https://doi.org/10.5334/jcms.1021218
952 Høye, T. T., Ärje, J., Bjerge, K., Hansen, O. L. P., Iosifidis, A., Leese, F., Mann, H. M. R.,
953 Meissner, K., Melvad, C., & Raitoharju, J. (2021). Deep learning and computer vision
954 will transform entomology. Proceedings of the National Academy of Sciences, 118(2),
955 e2002545117. https://doi.org/10.1073/pnas.2002545117
956 Hudson, L. N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B. W., Walt,
957 S. van der, & Smith, V. S. (2015). Inselect: Automating the Digitization of Natural
958 History Collections. PLOS ONE, 10(11), e0143402.
959 https://doi.org/10.1371/journal.pone.0143402
960 Johnson, C. A., Willigan, E. M., Hwang, E. Y., Austin, P. M., Chang, K. L., Daly, P., Davis, A.,
961 Gonzalez, J. M., Kao, C.-N., Mifsud, I. E. J., Miller, K. A., Nicolescu, M., Ruiz, A. M.,
962 Ruo, E. Y., Schwartzman, J. L., Seeman, A. B., Stenzel, J., Zurita, N. C., Berniker, L. F.,
963 & Rodríguez, E. (2026). Unlocking biodiversity data with robotic imaging and AI-driven
964 transcription of natural history collections. Science Advances, 12(18), eaeb6975.
965 https://doi.org/10.1126/sciadv.aeb6975
966 Kapun, M., Schwentner, M., Haring, E., Akkari, N., Kroh, A., Kruckenhauser, L., Palandačić,
967 A., & Vohland, K. (2025). Museomics, the Extended Specimen and Collectomics – how
968 to frame and name the diversity of information linked to specimens in natural history
969 collections. Natural History Collections and Museomics, 2, 1–21.
970 https://doi.org/10.3897/nhcm.2.161331
971 Lertrusdachakul, I., Leni, P.-E., Gschwind, R., & Bertheau, C. (2025). Deep learning for genera-
972 level bark beetle taxonomic classification. Forest Ecology and Management, 598,
973 123190. https://doi.org/10.1016/j.foreco.2025.123190

Levenshtein, V.I. (1966). Binary codes capable of correction deletions, insertions, and reversals. *Soviet Physics Doklady,* 10, 707-710.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, 74–81. https://aclanthology.org/W04-1013/

Liu, F., Rayeed, S. M., Stevens, S., East, A., Chiang, C. H., Lee, C., Yi, D., Yang, J., Naik, T., Wang, Z., Kilrain, C., Buckwalter, E. H., Hou, J., Bueno, S. I., Wang, S., Ma, X., Liu, Y., Tao, Z., Zhang, Z., … Chao, W.-L. (2025). BeetleFlow: An Integrative Deep Learning Pipeline for Beetle Image Processing (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2511.00255

Lürig, M. D. (2022). phenopype: A phenotyping pipeline for Python. Methods in Ecology and Evolution, 13(3), 569–576. https://doi.org/10.1111/2041-210X.13771

Mantle BL, La Salle J, Fisher N (2012) Whole-drawer imaging for digital management and curation of a large entomological collection. ZooKeys 209: 147–163. https://doi.org/10.3897/zookeys.209.3169

Mesnage, C. S., Corbett, A., Curry, J., Rowlands, S., Dutta, A., Everson, R., & Simmons, B. I. (2025). Copy-Paste Augmentation Improves Automatic Species Identification in Camera Trap Images. Ecology and Evolution, 15(11), e72357. https://doi.org/10.1002/ece3.72357

Mráz, R., Štěpka, K., Pekár, M., Matula, P., & Pekar, S. (2023). MAPHIS—Measuring arthropod phenotypes using hierarchical image segmentations. Methods in Ecology and Evolution, 15. https://doi.org/10.1111/2041-210X.14250

Pearson, D. L., Knisley, C. B. & Kazilek, C. J. A Field Guide to the Tiger Beetles of the United States and Canada (Oxford University Press, 2006).

Pearson DL, Vogler AP. 2001. *Tiger Beetles: The Evolution, Ecology, and Diversity of the Cicindelids*. Ithaca: Comstock Publishing Associates.

Pennell, M. W., FitzJohn, R. G., & Cornwell, W. K. (2016). A simple approach for maximizing the overlap of phylogenetic and comparative data. Methods in Ecology and Evolution, 7(6), 751–758. https://doi.org/10.1111/2041-210X.12517

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution, 14(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Pollock, L. J., Kitzes, J., Beery, S., Gaynor, K. M., Jarzyna, M. A., Mac Aodha, O., Meyer, B., Rolnick, D., Taylor, G. W., Tuia, D., & Berger-Wolf, T. (2025). Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. Nature Reviews Biodiversity, 1, 166–182. https://doi.org/10.1038/s44358-025-00022-3

Postema, E. G., Lippey, M. K., & Armstrong-Ingram, T. (2023). Color under pressure: how multiple factors shape defensive coloration. Behavioral Ecology, 34(1), 1–13. https://doi.org/10.1093/beheco/arac056

Picturae. Entomology Digitization. https://picturae.com/services/entomology-digitization/ (accessed 28 October 2025).

Salili-James, A., Scott, B., Livermore, L., Price, B., Dupont, S., Hardy, H., & Smith, V. S. (2023). AI-Accelerated Digitisation of Insect Collections: The next generation of Angled Label Image Capture Equipment (ALICE). Biodiversity Information Science and Standards, 7, e112742. https://doi.org/10.3897/biss.7.112742

Sanders, N. J., Cooper, N., Davis Rabosky, A. R., & Gibson, D. J. (2023). Leveraging natural history collections to understand the impacts of global change. Journal of Animal Ecology, 92(2), 232–236. https://doi.org/10.1111/1365-2656.13882

Schultz, T. D., & Bernard, G. D. (1989). Pointillistic mixing of interference colours in cryptic tiger beetles. Nature, 337(6202), 72–73. https://doi.org/10.1038/337072a0

Schultz, T. D., & Hadley, N. F. (1987). Structural Colors of Tiger Beetles and Their Role in Heat Transfer through the Integument. Physiological Zoology, 60(6), 737–745. https://doi.org/10.1086/physzool.60.6.30159990

Shirai, M., Takano, A., Kurosawa, T., Inoue, M., Tagane, S., Tanimoto, T., Koganeyama, T., Sato, H., Terasawa, T., Horie, T., Mandai, I., & Akihiro, T. (2022). Development of a system for the automated identification of herbarium specimens with high accuracy. Scientific Reports, 12(1), 8066. https://doi.org/10.1038/s41598-022-11450-y

Shiyake, S. How to Identify Tiger Beetles (Osaka Museum of Natural History, 2017).

Spiesman, B. J., Gratton, C., Hatfield, R. G., Hsu, W. H., Jepsen, S., McCornack, B., Patel, K., & Wang, G. (2021). Assessing the potential for deep learning and computer vision to identify bumble bee species from images. Scientific Reports, 11(1), 7580. https://doi.org/10.1038/s41598-021-87210-1

Steinke, D., Mckeown, J., Zyba, A., Mcleod, J., Feng, C., & Hebert, P. (2024). Low-cost, high-volume imaging for entomological digitization. ZooKeys, 1206, 315–326. https://doi.org/10.3897/zookeys.1206.123670

Stenhouse, A., & Thrall, P. H. (2025). SpeciMate: Improving metadata extraction from digitised biological specimens. Biodiversity Data Journal, 13, e160553. https://doi.org/10.3897/BDJ.13.e160553

Stork, N. E. (2018). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? Annual Review of Entomology, 63, 31–45. https://doi.org/10.1146/annurev-ento-020117-043348

Sun, J., Futahashi, R., & Yamanaka, T. (2021). Improving the Accuracy of Species Identification by Combining Deep Learning With Field Occurrence Records. Frontiers in Ecology and Evolution, 9. https://doi.org/10.3389/fevo.2021.762173

Sys, S., Weißbach, S., Jakob, L., Gerber, S., & Schneider, C. (2022). CollembolAI, a macrophotography and computer vision workflow to digitize and characterize samples of soil invertebrate communities preserved in fluid. Methods in Ecology and Evolution, 13(12), 2729–2742. https://doi.org/10.1111/2041-210X.14001

Trautner, J. & Geigenmüller, K. Tiger Beetles Ground Beetles: Illustrated Key to the Cicindelidae and Carabidae of Europe (Margraf, 1987).

Van Belleghem, S. M., Papa, R., Ortiz-Zuazaga, H., Hendrickx, F., Jiggins, C. D., Owen McMillan, W., & Counterman, B. A. (2018). patternize: An R package for quantifying colour pattern variation. Methods in Ecology and Evolution, 9(2), 390–398. https://doi.org/10.1111/2041-210X.12853

van den Berg, C. P., Troscianko, J., Endler, J. A., Marshall, N. J., & Cheney, K. L. (2020). Quantitative Colour Pattern Analysis (QCPA): A comprehensive framework for the analysis of colour patterns in nature. Methods in Ecology and Evolution, 11(2), 316–332. https://doi.org/10.1111/2041-210X.13328

van der Loo, M. P. J. (2014). The stringdist Package for Approximate String Matching. The R Journal, 6(1), 111–122. https://doi.org/10.32614/RJ-2014-011

Vogler, A. P., Cardoso, A., & Barraclough, T. G. (2005). Exploring Rate Variation Among and Within Sites in a Densely Sampled Tree: Species Level Phylogenetics of North American Tiger Beetles (Genus Cicindela). Systematic Biology, 54(1), 4–20. https://doi.org/10.1080/10635150590906028

Wagner, D. L. (2020). Insect Declines in the Anthropocene. Annual Review of Entomology, 65(Volume 65, 2020), 457–480. https://doi.org/10.1146/annurev-ento-011019-025151

Weaver, W. N., & Smith, S. A. (2023). From leaves to labels: Building modular machine learning networks for rapid herbarium specimen analysis with LeafMachine2. Applications in Plant Sciences, 11(5), e11548. https://doi.org/10.1002/aps3.11548

Welch, K. D., & Lundgren, J. G. (2024). Introducing BugBox: A Platform for AI-Assisted Bioinventories of Arthropods. American Entomologist, 70(3), 31–33. https://doi.org/10.1093/ae/tmae047

Weller, H. I., Hiller, A. E., Lord, N. P., & Van Belleghem, S. M. (2024). recolorize: An R package for flexible colour segmentation of biological images. Ecology Letters, 27(2), e14378. https://doi.org/10.1111/ele.14378

Wiken, E., Jiménez Nava, F. & Griffith, G. North American Terrestrial Ecoregions — Level III (Commission for Environmental Cooperation, 2011).

Wightman R. 2019. PyTorch Image Models. Zenodo. DOI: 10.5281/zenodo.4414861. Available at https://github.com/huggingface/pytorch-image-models (accessed 28 May 2026).

Wu, S., Chang, C.-M., Mai, G.-S., Rubenstein, D. R., Yang, C.-M., Huang, Y.-T., Lin, H.-H., Shih, L.-C., Chen, S.-W., & Shen, S.-F. (2019). Artificial intelligence reveals environmental constraints on colour diversity in insects. Nature Communications, 10(1), 4554. https://doi.org/10.1038/s41467-019-12500-2

Wührl, L., Pylatiuk, C., Giersch, M., Lapp, F., von Rintelen, T., Balke, M., Schmidt, S., Cerretti, P., & Meier, R. (2022). DiversityScanner: Robotic handling of small invertebrates with machine learning methods. Molecular Ecology Resources, 22(4), 1626–1638. https://doi.org/10.1111/1755-0998.13567

Yamamoto, N., & Sota, T. (2020). Evolutionary fine-tuning of background-matching camouflage among geographical populations in the sandy beach tiger beetle. Proceedings of the Royal Society B: Biological Sciences, 287(1941), 20202315. https://doi.org/10.1098/rspb.2020.2315

Yang, L. H., Postema, E. G., Hayes, T. E., Lippey, M. K., & MacArthur-Waltz, D. J. (2021). The complexity of global change and its effects on insects. Current Opinion in Insect Science, 47, 90–102. https://doi.org/10.1016/j.cois.2021.05.001

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization (arXiv:1710.09412). arXiv. https://doi.org/10.48550/arXiv.1710.09412

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4), 99. https://doi.org/10.1007/s10462-024-10721-6

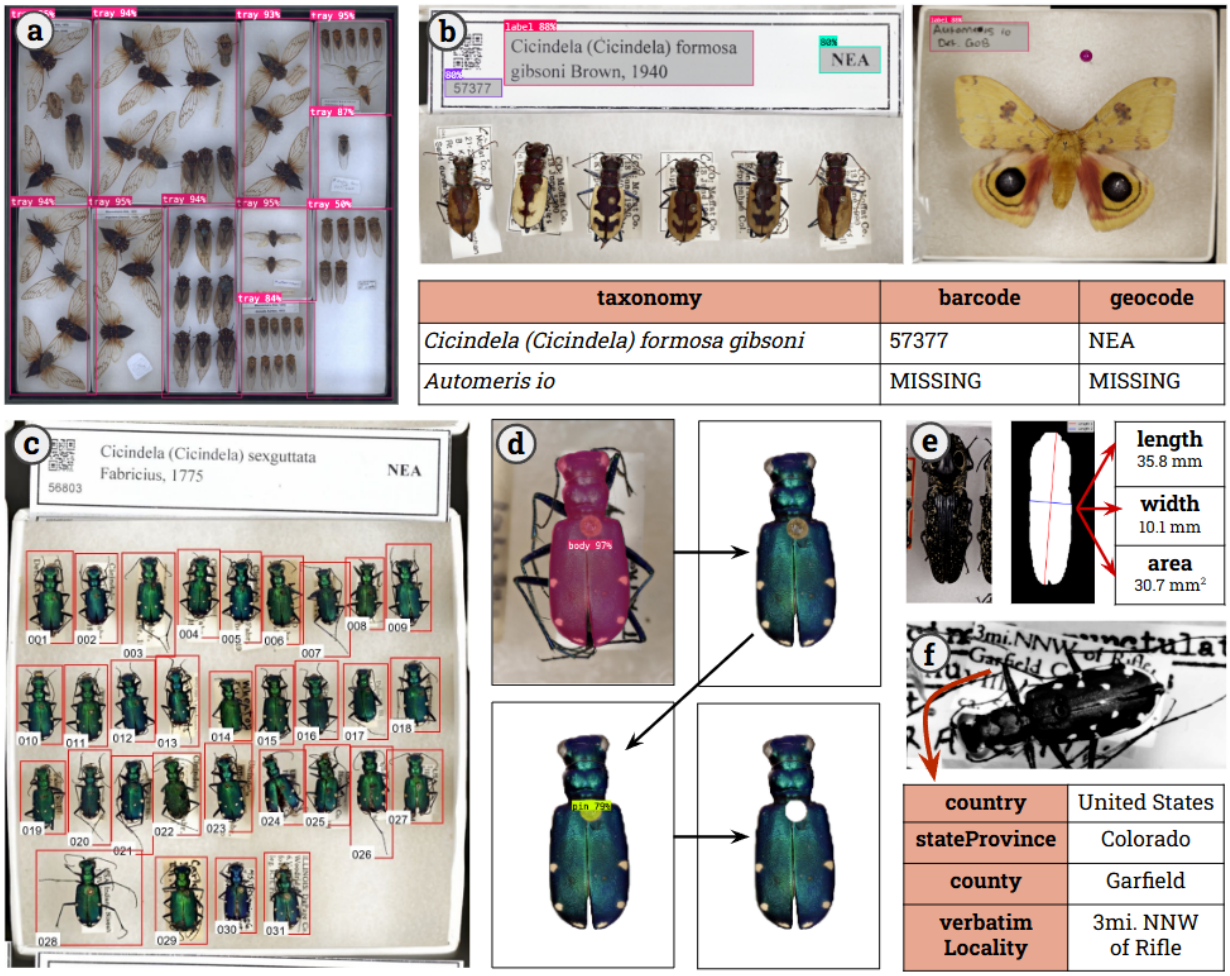

Figure 1. Overview of DrawerDissect steps. (a) Trays are detected in whole-drawer images. (b) Tray-level text is detected, cropped, and transcribed, including handwritten text. (c) Specimens are detected and cropped from trays, producing specimen-level dorsal views and numbered location maps. (d) Specimens go through two masking steps to generate images that are immediately usable for phenotypic analyses. (e) Specimen masks are measured to get specimen length, width, and area. (f) An LLM outputs DarwinCore-standard metadata fields transcribed and interpreted from any top-down visible text.

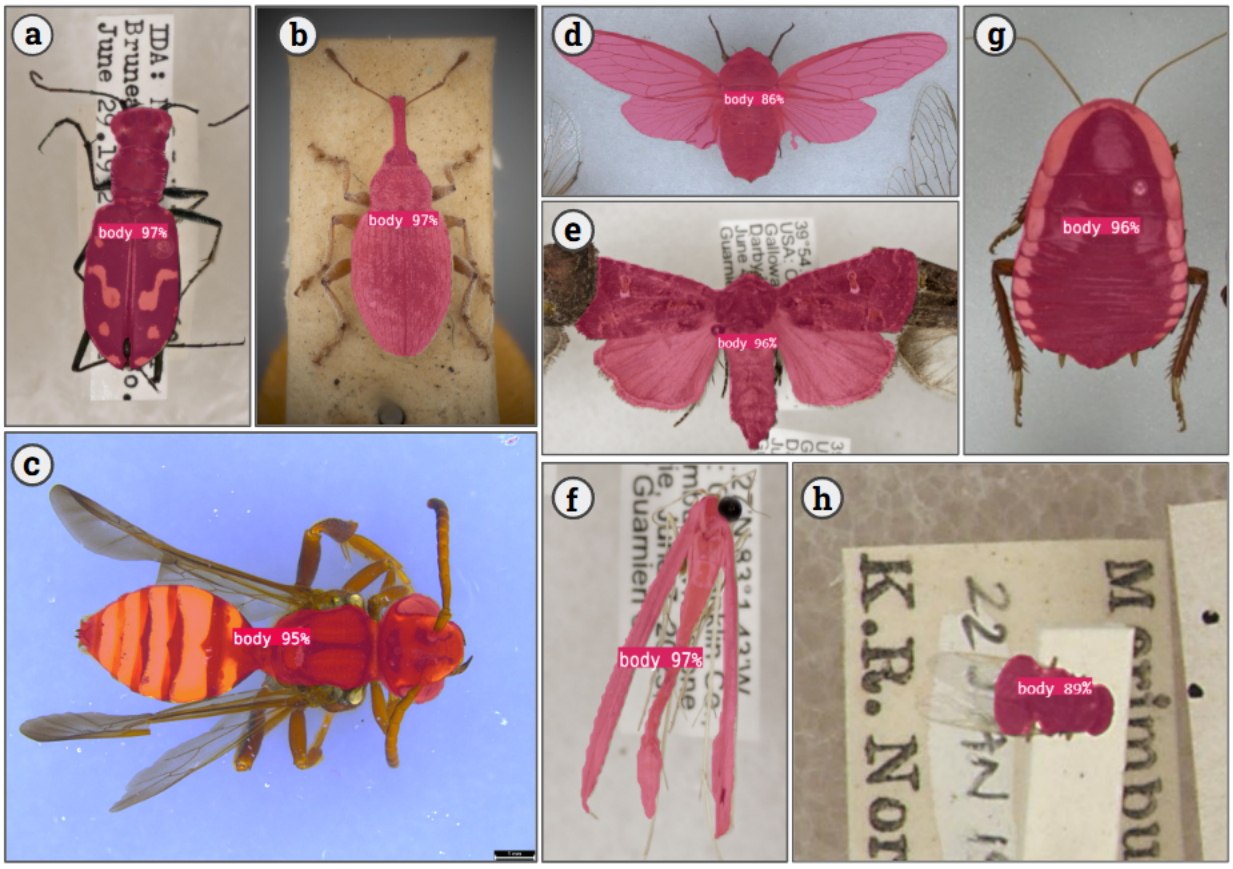

Figure 2. Visualizations of masks produced by DrawerDissect for different taxa, with model confidence as a percent. (a-b) Coleoptera, (c) Hymenoptera, (d) Hemiptera, (e-f) Lepidoptera, (g) Blattodea, (h) Diptera.

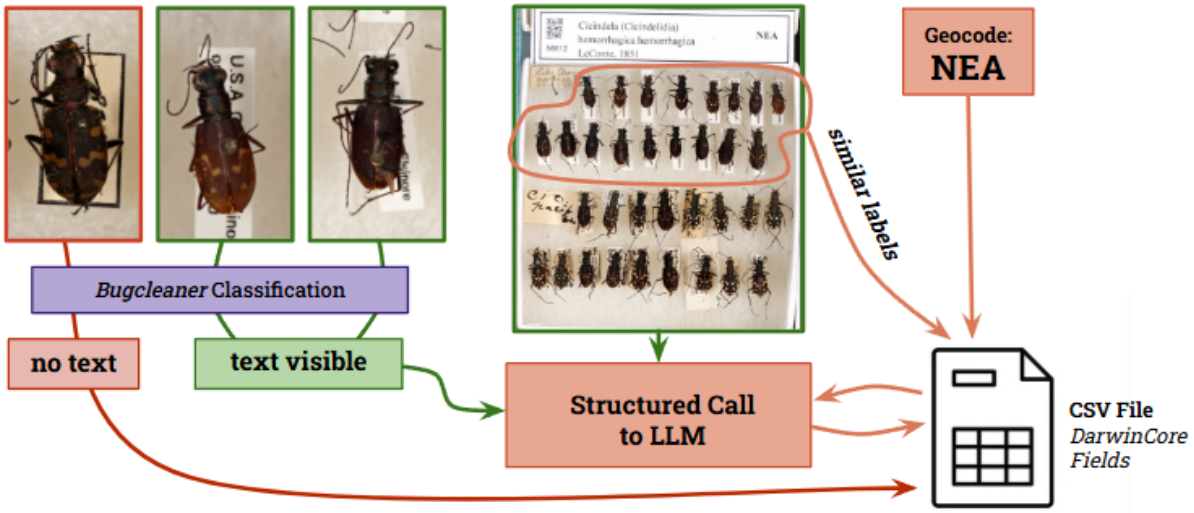

Figure 3. Visual summary of the prompt structure and information involved in specimen-level metadata transcription. Specimen dorsals are first filtered by *bugcleaner* to find specimens with visible label text; specimens with no text are flagged as 'no_text_detected' in the final CSV. Specimens with visible text are sent to an LLM with a detailed prompt, as well as unit tray images and previously transcribed tray-level geocodes (if present) for additional context. The LLM transcribes all visible text, maps the text to DarwinCore fields, flags errors, ambiguities, or conflicting information, and groups similar-looking labels (or rows with similar verbatim text) to interpolate missing information.

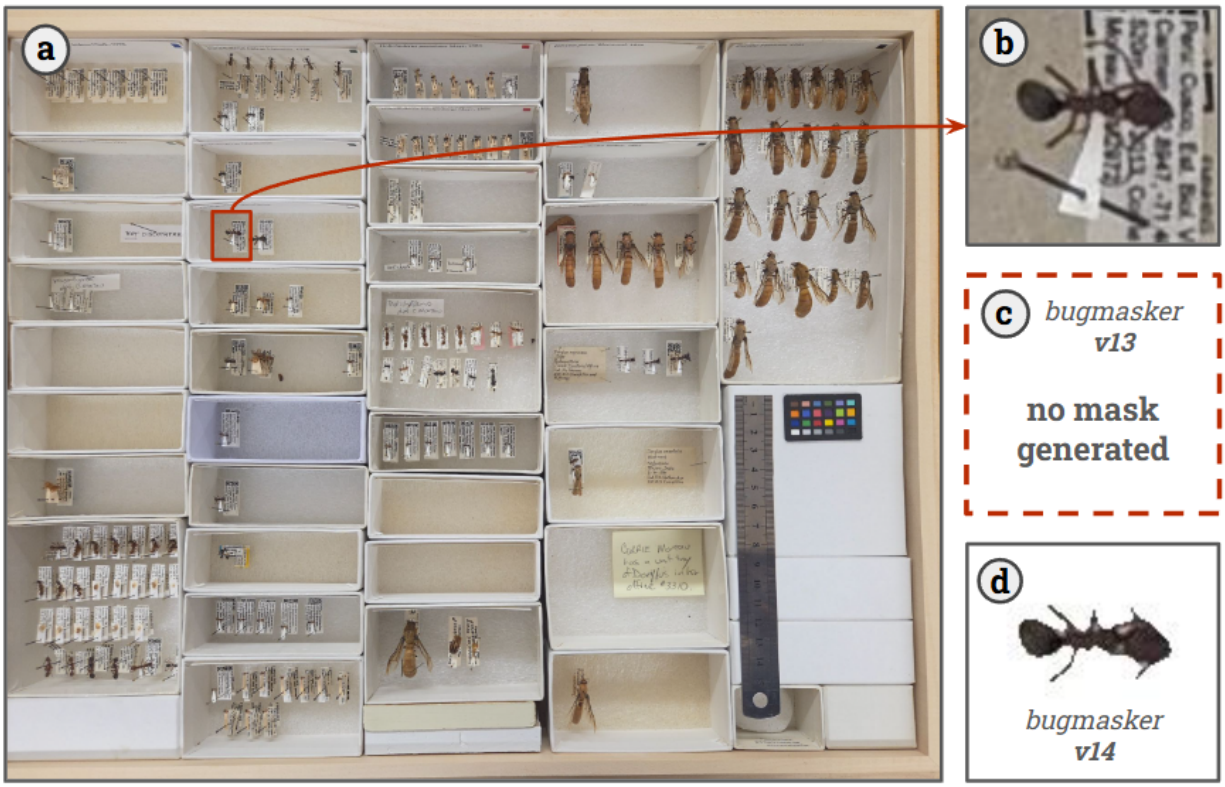

Figure 4. Performance comparison after retraining *bugmasker-all.* (a) Galaxy S25 cellphone panorama of a drawer of ants (Formicidae). (b) A specimen dorsal image produced by *bugfinder-knde9* (v21). (c) Mask produced by *bugmasker-all* (v13), which had no previous training on cellphone-quality images. (d) Mask produced for the same specimen by *bugmasker* (v14).

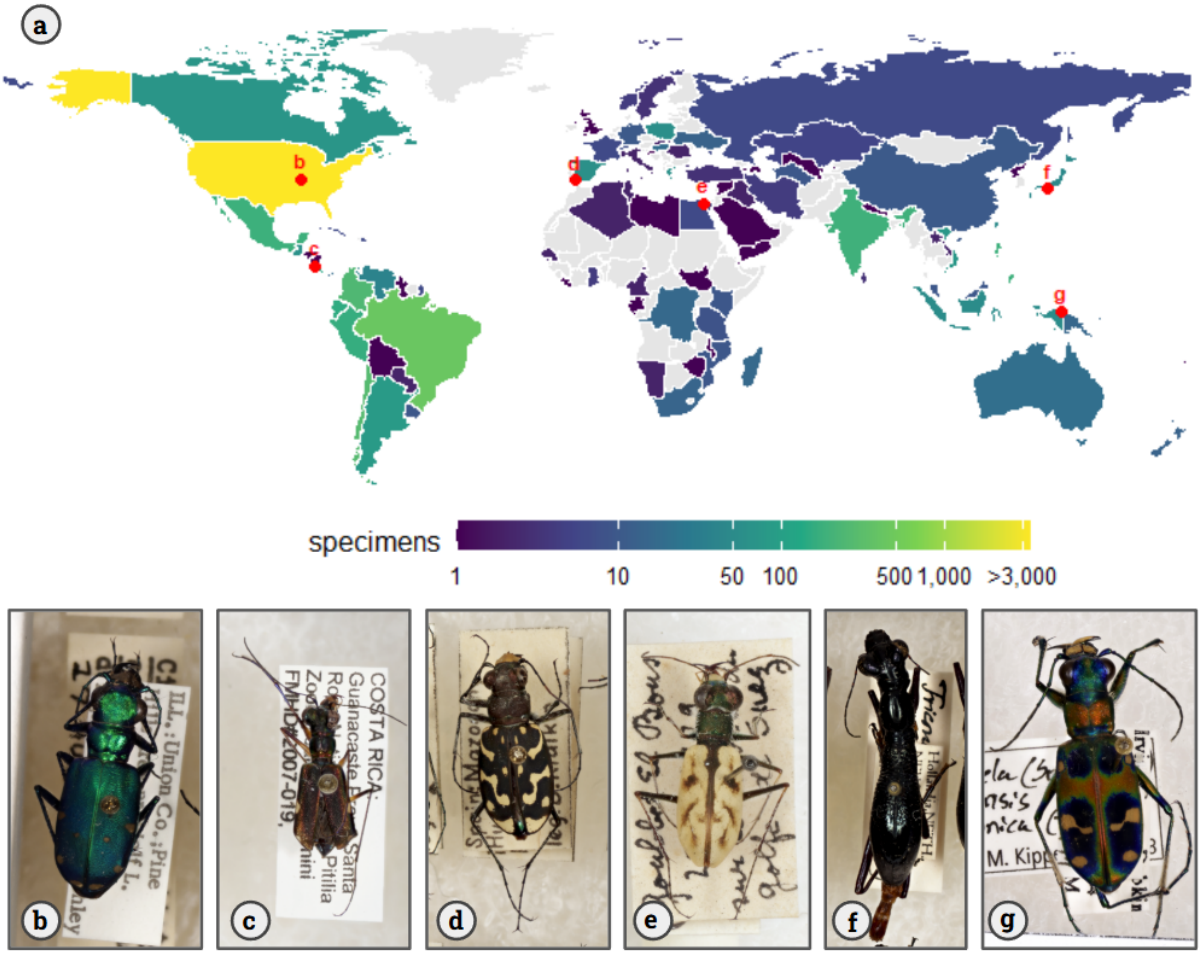

Figure 5. Distribution of tiger beetles from the FMNH Collection. (a) A world map showing the number of FMNH specimens per country from a sample of 6,190 specimens with either manually- or LLM-transcribed locations (representing 46% of the Cicindelidae collection). (b) *Cicindela sexgutatta* Fabricius, 1775 from Pine Hill, IL, USA. (c) *Odontocheila iodopleura* Bates, 1872 from the Pitilla Biological Station, Costa Rica. (d) *Lophyra flexuosa* Fabricius, 1787 from Mazagón, Spain. (e) *Hypaetha singularis* Chaudoir, 1876 from a cove on the coast of the Gulf of Suez in Egypt. (f) *Tricondyla aptera aptera* Olivier, 1790 from Jayapura, Indonesia. (g) *Cicindela chinensis japonica* Thunberg, 1781 from Ōdai, Japan

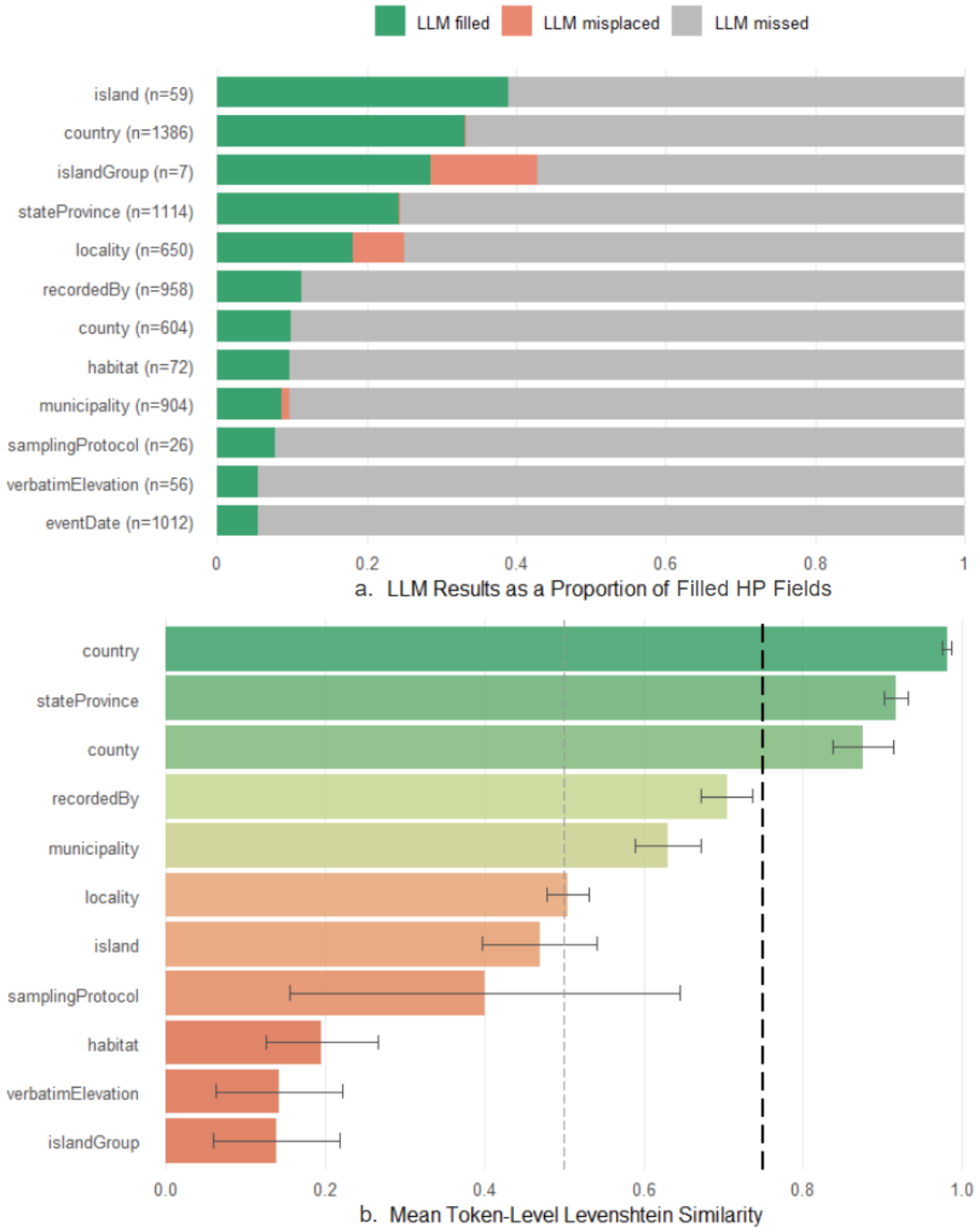

Figure 6. Evaluation of DrawerDissect's LLM-produced transcriptions. (a) Records filled, misplaced, or missed (not filled) by the LLM as a proportion of the total human-produced (HP), filled records for 1,404 specimens. N = the number of total filled HP records for each field. (b) Mean token-level Levenshtein similarity scores by DwC field, +/- SE, excluding records where both the LLM and the HP field were empty. The dashed line at 0.75 indicates the threshold where an LLM transcription is an acceptable match. The dashed line at 0.5 indicates the

threshold where transcriptions often contain some correct information, but may miss critical text compared to the HP record.

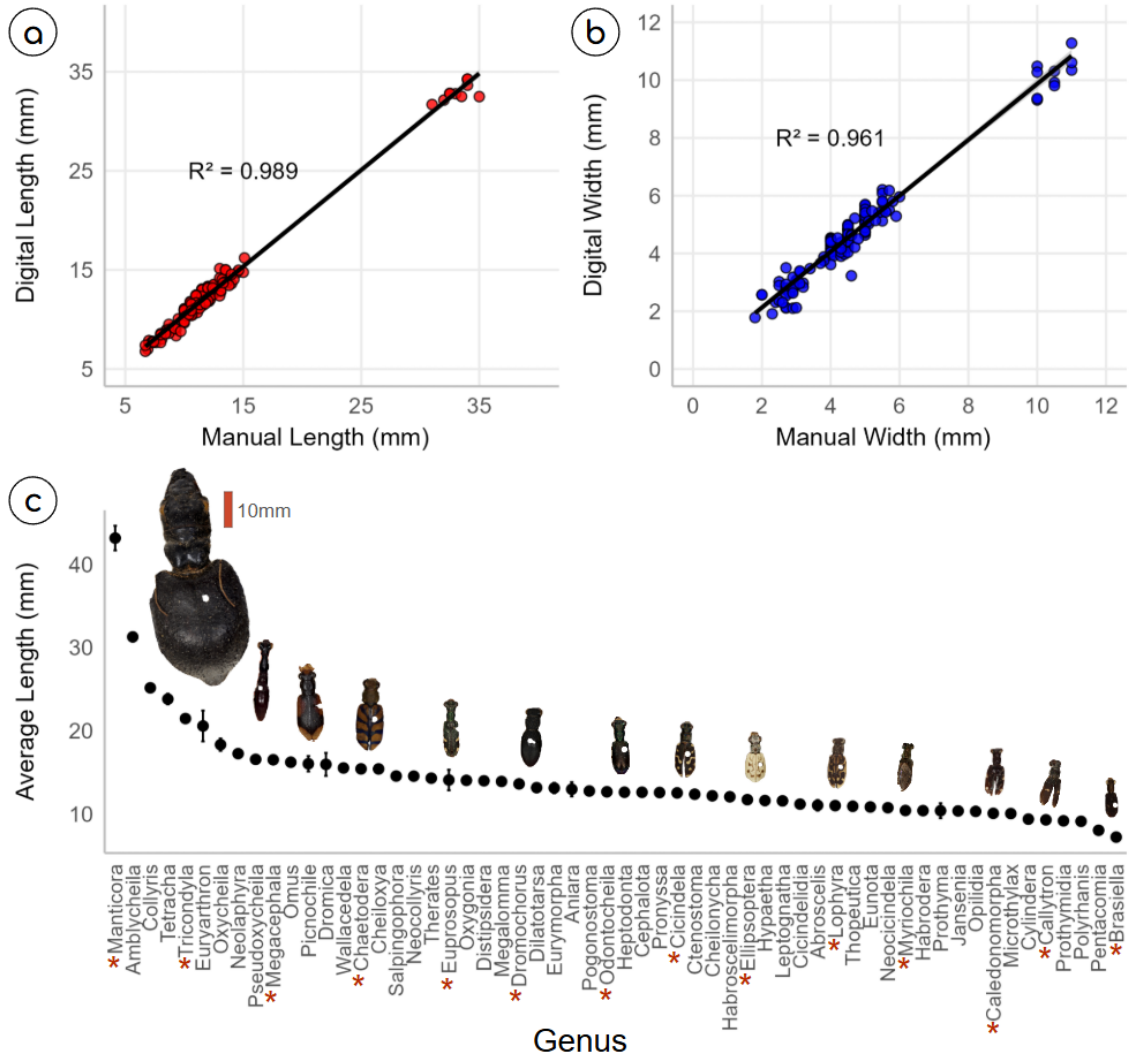

Figure 7. Validation for automated mask measurements generated by DrawerDissect. (a) The relationship between manual and digital length for a set of randomly selected specimens (n = 143), both in mm. (b) The relationship between manual and digital width for the same set of specimens. (c) The average body length, in mm, +/- SE of each genus in the FMNH collection (n = 13,484). *Indicates genus with an associated icon generated by DrawerDissect.

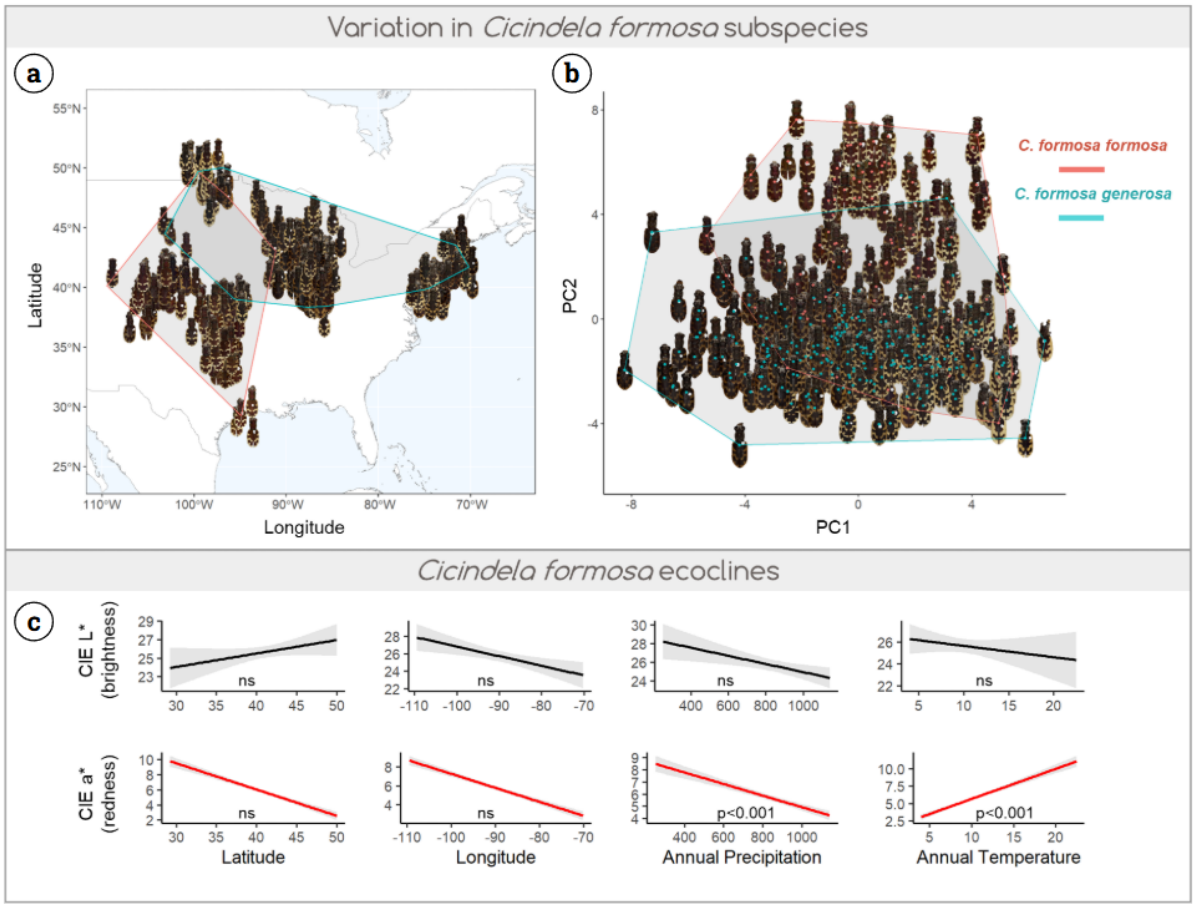

Figure 8. Geographic and environmental correlates of *Cicindela formosa* subspecies. (a) A map of the collection locations of *C. formosa formosa* and *C. formosa generosa*. (b) Specimens of each subspecies plotted by PC1 (darker to lighter) and PC2 (greener to redder). (c) Relationships between phenotypic and abiotic variables for all *C. formosa* specimens (n=374). Significant relationships are labeled within plots; ns = non-significant.

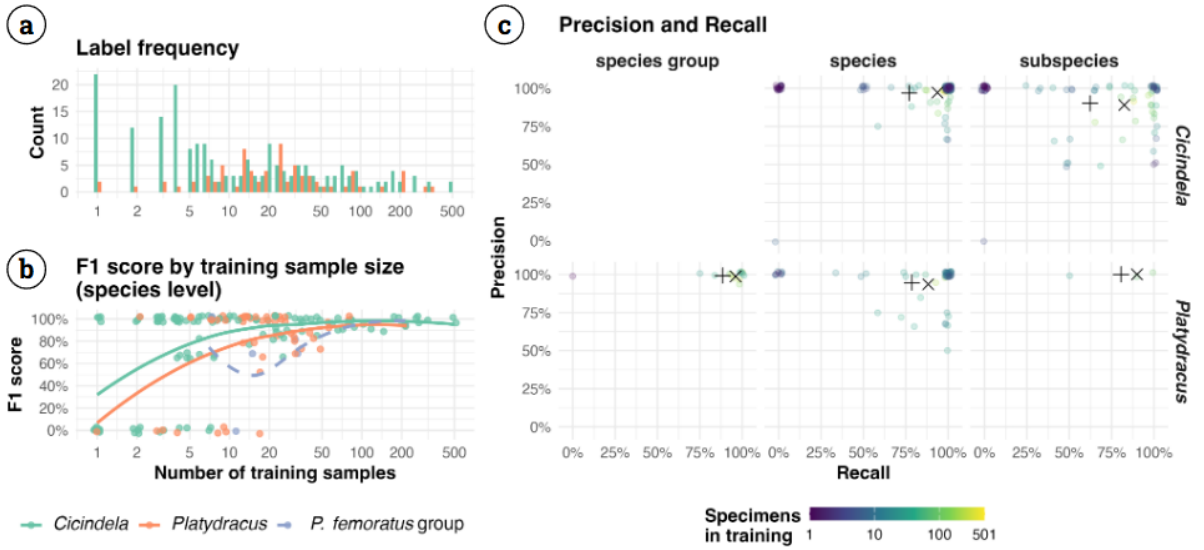

Figure 9. Taxonomic identification based on an image classification model trained on images of *Cicindela* and *Platydracus*. (a) Distribution of species and subspecies label frequencies (log scale), showing high imbalance (*P. femoratus* group included in *Platydracus*). (b) F1 score increases with sample size used in training. Points represent taxonomic labels at the species level,with trend lines showing smoothed mean scores. *Platydracus* trend line includes the *femoratus* group. (c) Precision and recall for each taxon (circles), with averages across samples (╳), and across labels (+). A small jitter was added to individual labels to allow observation of overlapping records. Point color shows the number of specimens in the training set.

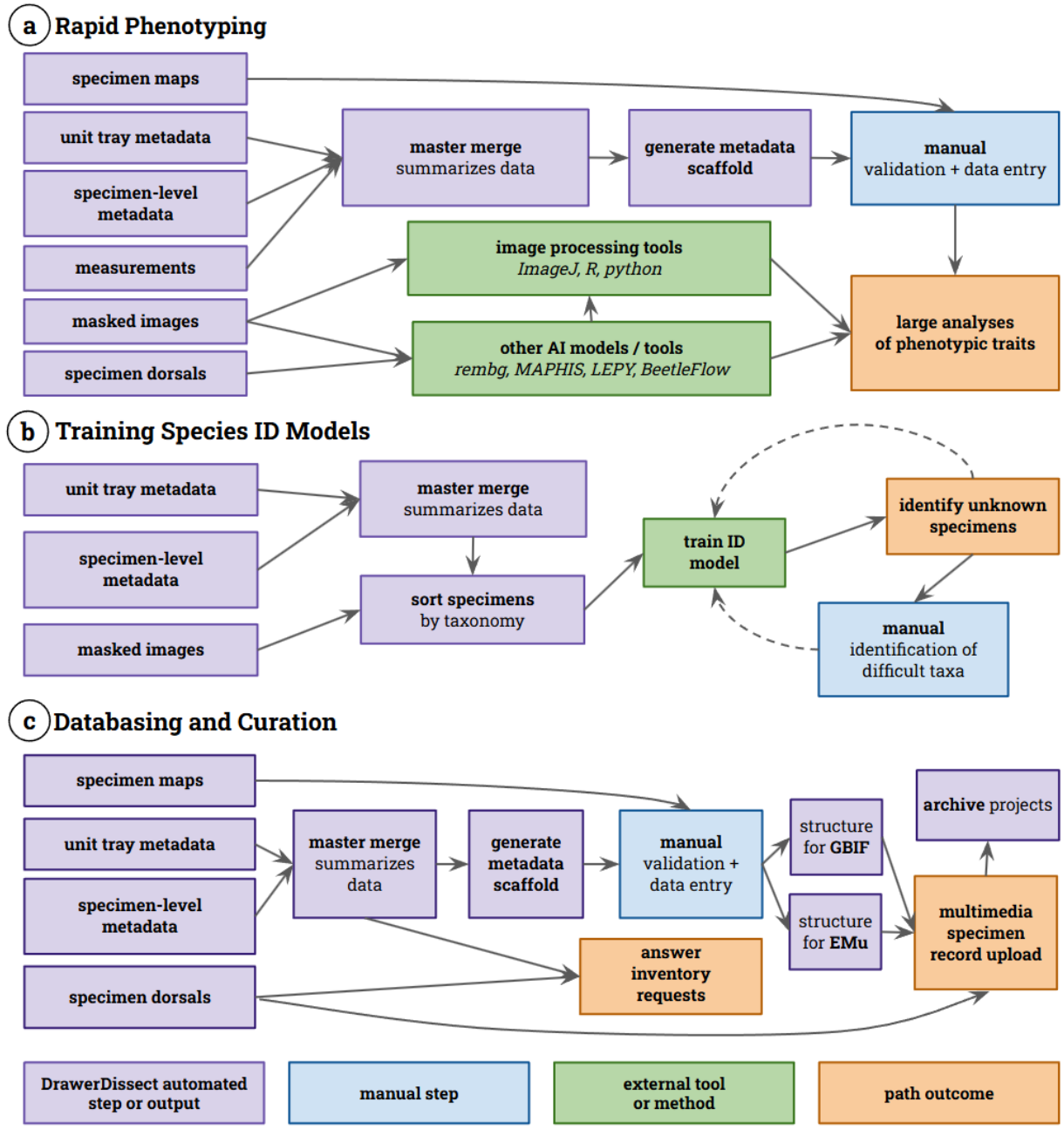

Figure 10. Overview of three example paths for using DrawerDissect outputs. (a) Rapid specimen phenotyping for large-scale comparative analyses, (b) training AI models for species identification, and (c) databasing and curation. All steps and outputs are color-coded: automatically produced by DrawerDissect (purple), manually produced (blue), or produced with an external tool/method (green). Final outcomes of each path are shown in orange. Dashed lines indicate potential iterative loops.

TABLE 1: Model Composition and Performance on Test Set

model id (*version)	# of training images**	precision	recall	false positive rate	false negative rate
trayfinder-base (5)	180	99.6%	99.1%	0.02%	0.01%
trayfinder-popup (17)	162	100.0%	100.0%	0.01%	0.00%
labelfinder (7)	201	96.0%	99.0%	0.02%	0.01%
bugfinder-kdn9e (13)	4235	97.1%	96.7%	0.02%	0.02%
bugmasker-all (5)	2035	99.9%	97.3%	0.02%	0.02%
pinmasker (6)	2210	91.0%	93.0%	0.08%	0.07%
bugcleaner (5)	3070	80.8%	80.8%	0%	0%

*The version originally used to digitize the FMNH tiger beetle collection. More recent models are available at universe.roboflow.com/field-museum.

**Includes augmented images, which increases the size of the training split x5.

TABLE 2: Evaluation of LLM-Augmented Specimen Label Transcription

tray ID	geography	# specimens	manual-only	AI-assisted	time saved
34_5_7_tray_01	United States	33	24m 06s	15m 41s	8m 25s
34_5_7_tray_03	Taiwan, India	20	11m 02s	7m 57s	3m 05s
34_5_7_tray_15	Europe, Russia	12	5m 36s	5m 46s	none

TABLE 3: Time and Cost Estimates for Processing the FMNH Cicindelidae

Step	cost per specimen	cost for all Cicindelidae	processing time per specimen	processing time for all Cicindelidae
Imaging*	-	-	18.7 seconds	70 hours
Transcription	$0.0037	$50	1.6 seconds	6 hours
Image Processing	$0.0445	$600**	1.1 seconds	4 hours
Pipeline Total	**$0.0482**	**$650**	**21.4 seconds**	**80 hours**

*Up-front costs for imaging equipment will vary.

**Calculated based on the $600 annual plan for Roboflow API calls, which includes 600,000 inferences/year.

TABLE 4. Model versus human identifications for 12 previously unidentified samples of *Cicindela* and 3 tentatively identified samples of *Platydracus*.

# samples	genus	human identification	identification source	model predictions
7	*Cicindela*	*C. ocellata ocellata*	Pearson et al., 2006	"Cicindela", "cicindela_ocellata", "cicindela_ocellata_ocellata"
2	*Cicindela*	*C. sylvicola*	Trautner & Geigenmüller, 1987	"Cicindela" "cicindela_sylvicola"
1	*Cicindela*	*C. repanda repanda*	Pearson et al., 2006	"Cicindela", "cicindela_repanda", "cicindela_repanda_repanda"
1	*Cicindela*	*C. transbaicalica japanensis*	Shiyake, 2017	"Cicindela"
1	*Cicindela*	*C. sylvicola*	Trautner & Geigenmüller, 1987	"Cicindela" "cicindela_hybrida"
2	*Platydracus*	*femoratus* group: *P. championi* or *P. sallei*	A. Newton synoptic collection	"Platydracus" "platydracusgroup_femoratus" "platydracus_championi", "platydracus_femoratus"
1	*Platydracus*	*femoratus* group: *P. championi* or *P. sallei*	A. Newton synoptic collection	"Platydracus" "platydracusgroup_femoratus"

Table 5. Ability of *bugmasker* v12 and v14 to successfully mask specimens of different insect families from cellphone panoramas.

order	family	# of specimens	% usable masks* (v12)	% usable masks* (v14)	# of images added	v12 → v14 change in % usable masks
Coleoptera	Cicindelidae	452	87.8%	97.3%	12	+9.5
Phasmatodea	Diapheromeridae	65	29.2%	98.5%	14	+69.3
Orthoptera	Gryllidae	123	55.3%	91.1%	14	+35.8
Mantodea	Liturgusidae	30	30.0%	90.0%	17	+60.0
Diptera	Syrphidae	197	25.4%	74.6%	15	+49.2
Hemiptera	Pentatomidae	236	11.4%	87.7%	17	+76.3
Hymenoptera	Formicidae	127	28.3%	66.9%	14	+38.6

*Mask quality was determined visually by a human evaluator. A usable mask is defined as one that successfully segments the organism from its background without missing key parts of the insect's body: the head, thorax, abdomen, and, for spread specimens, wings. Legs were also counted as key body parts for specimens where legs are more consistently visible (orders Phasmotodea, Orthoptera, and Mantodea).