DiversityScanner: Robotic discovery of small invertebrates with machine learning methods

Running title: DiversityScanner: an invertebrate sorting robot

Lorenz Wührl[1], Christian Pylatiuk[1,*], Matthias Giersch[1], Florian Lapp[1], Thomas von Rintelen[3], Michael Balke[4], Stefan Schmidt[4], Pierfilippo Cerretti[5], and Rudolf Meier[2,*]

[1]**Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany**
[2]**Department of Biological Science, National University of Singapore (NUS), Singapore**
[3]**Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany**
[4]**SNSB – Zoologische Staatssammlung München, Munich, Germany**
[5]**Sapienza University of Rome, Rome, Italy**
[*]**Correspondence: *pylatiuk@kit.edu & Rudolf.Meier@mfn.berlin***

ABSTRACT

Invertebrate biodiversity remains poorly explored although it comprises much of the terrestrial animal biomass, more than 90 % of the species-level diversity, and supplies many ecosystem services. The main obstacle is specimen- and species-rich samples. Traditional sorting techniques require manual handling and are slow while molecular techniques based on metabarcoding struggle with obtaining reliable abundance information. Here we present a fully automated sorting robot which detects each specimen, images and measures it before moving it from a mixed invertebrate sample to the well of a 96-well microplate in preparation for DNA barcoding. The images are used by a newly trained convolutional neural network (CNN) to assign the specimens to 14 particularly common “classes” of insects (N = 14) in Malaise trap samples. The average assignment precision for the classes is 91.4 % (75 - 100 %). In order to obtain biomass information, the specimen images are also used to measure specimen length and estimate body volume. We outline how the “DiversityScanner” robot can be a key component for tackling and monitoring invertebrate diversity. The robot generates large numbers of images that become training sets for CNNs once the images are labelled with identifications based on DNA barcodes. In addition, the robot allows for taxon-specific subsampling of large invertebrate samples by only removing the specimens that belong to one of the 14 classes. We conclude that a combination of automation, machine learning, and DNA barcoding has the potential to tackle invertebrate diversity at an unprecedented scale.

Keywords: biodiversity, classification, convolutional neural network, insects, machine learning

1 INTRODUCTION

Biodiversity science is currently at an inflection point. For decades, biodiversity declines had been mostly an academic concern although many biologists already predicted that these declines would eventually threaten whole ecosystems. Unfortunately, we are now at this stage, which explains why the World Economic Forum considers biodiversity decline one of the top three global risks based on likelihood and impact for the next 10 years (World Economic Forum's Global Risk Initiative, 2020). This new urgency is also leading to a reassessment of research priorities. Biologists traditionally focused on charismatic taxa (e.g., vertebrates, vascular plants, butterflies) with a preference for endangered species. However, with regard to quantitative arguments, many of these taxon biases have been poorly justified. For example, if one were to adopt a biomass point of view to terrestrial animal diversity, invertebrates would receive most of the attention because they contribute >45 times the biomass of wild vertebrates (Table S23 in Bar-On et al. (2018)), contain >90 % of the species diversity, and much of the functional and evolutionary diversity. This means that efficient tools for assessing and monitoring invertebrates are urgently needed because little has changed since Robert May (2011) declared 10 years ago: "We are astonishingly ignorant about how many species are alive on earth today, and even more ignorant about how many we can lose (and) yet still maintain ecosystem services that humanity ultimately depends upon." Much of the undiscovered and undescribed animal diversity is in what is now referred to as "dark taxa". Hartop et al. (2021) recently defined these clades as those "for which the undescribed fauna is estimated to exceed the described fauna by at least one order of magnitude and the total diversity exceeds 1.000 species." Species discovery in these taxa is particularly difficult because it requires species-level sorting of thousands of small specimens that frequently need dissection for identification with morphological traits.

Fortunately, there are three technical developments that promise relief. The first is cost-effective DNA barcoding with 2nd and 3rd generation sequencing technologies (Hebert et al., 2018; Srivathsan et al., 2019a; Wang et al., 2018). In particular, portable nanopore sequencers by Oxford Nanopore Technologies are in the process of democratizing access to DNA sequence data (Srivathsan et al., 2021; Watsa et al., 2020; Pomerantz et al., 2018). The two remaining developments remain underutilized. They are automation and data processing with neural networks. Currently, automation mostly exists in the form of pipetting robots in molecular laboratories, while data processing with neural networks is only widely used for monitoring charismatic taxa. Bulk invertebrate samples have benefited very little (but see Ärje et al. (2020b)) although thousands of samples are collected every day. They include plankton samples in marine biology, macroinvertebrate samples used for assessing freshwater quality, and mass insect samples (Karlsson et al., 2020; Brown et al., 2018; Borkent et al., 2015; Brown, 2005). Automation and data processing with artificial intelligence have the potential to greatly increase the amount of information that can be obtained from such samples (Ärje et al., 2020b). The desirable end goal should be convolutional neural nets that use images (1) to identify the specimens to species, (2) provide specimen and species counts, (3) measure biomass, and (4) compare the results to samples previously obtained from the same sites.

Currently, the most popular way to process bulk invertebrate samples is with metabarcoding but the technique is affected by taxonomic bias and struggles with yielding abundance information (Creedy et al., 2019). Fortunately, computer-based identification systems for invertebrates that could be used for specimen-based approaches are starting to yield promising results (Perre et al., 2016; Feng et al., 2016; Knyshov et al., 2021). Particularly attractive are deep convolutional neural nets with transfer learning (Ärje et al., 2020b), but they require reasonably large sets of training images which are hard to obtain for invertebrates given that most species are undescribed and/or difficult to identify. It is here that robotics can have an impact. For example, one recently developed system can size and identify stoneflies (Plecoptera) (Sarpola et al., 2008) that are routinely used for freshwater quality assessment. Another system is designed for processing samples consisting of soil mesofauna (Chamblin et al., 2011). However, they use a robotic arm, which makes them comparatively expensive. Other robots have been designed for specific insect sorting purposes. One can separate intact mealworm larvae (*Tenebrio molitor*) from skins, feces, and dead worms. Another can sort mosquitoes (Lepek et al., 2020) and is capable of distinguishing males from females. However, all these machines lack the ability to recognize a wide variety of insect specimens in bulk invertebrate samples. The machine closest to this capability is the BIODISCOVER by Ärje et al. (2020a), which can identify ethanol-preserved specimens which, however, have to be fed into the machine manually one by one. After identification all specimen are returned into the same container.

We here describe a new system that overcomes some of these shortcomings. It recognizes insect specimens based on an overview image of a sample. Specimens below 3 mm body length are then imaged and moved into the wells of a 96-well microplate. We here demonstrate that the images are of sufficient quality for using convolutional neural nets (CNNs) for classifying the specimens into 14 common groups of insects (usually family-level). Furthermore, the images yield length measurements and an estimation of biomass based on specimen volume.

2 CONCEPT AND METHODS

The aim of the project was to develop an insect classification and sorting robot that is compact and that works reliably. Note that we here use to the term ”classification” in the machine learning context as assigning objects to different “classes”; i.e., the term “class” is here not used as a rank class in a Linnean classification and our “classes” may even be monophyletic groups (e.g., acalyptrate flies). Our robot consists mostly of standard components that all connected via parts that can be produced by a standard 3D printer. The basic design with a cube-shaped frame (50 x 50 x 50cm) and 3 linear drives with accurately positioning stepper motors is based on a zebrafish embryo handling robot (Pfriem et al., 2012). The robot was equipped with two high-resolution cameras with customized lenses, LED lighting and image recognition software. Furthermore, a transport system based on a suction pump was integrated to transfer detected insects into the wells of a standard 96-well microwell plate. Thus, the robot system can be divided into: (1) the Transport System, (2) the Image Acquisition System, (3) the Image Processing System, and (4) a touch screen with graphical user interface (GUI; Figure 1).

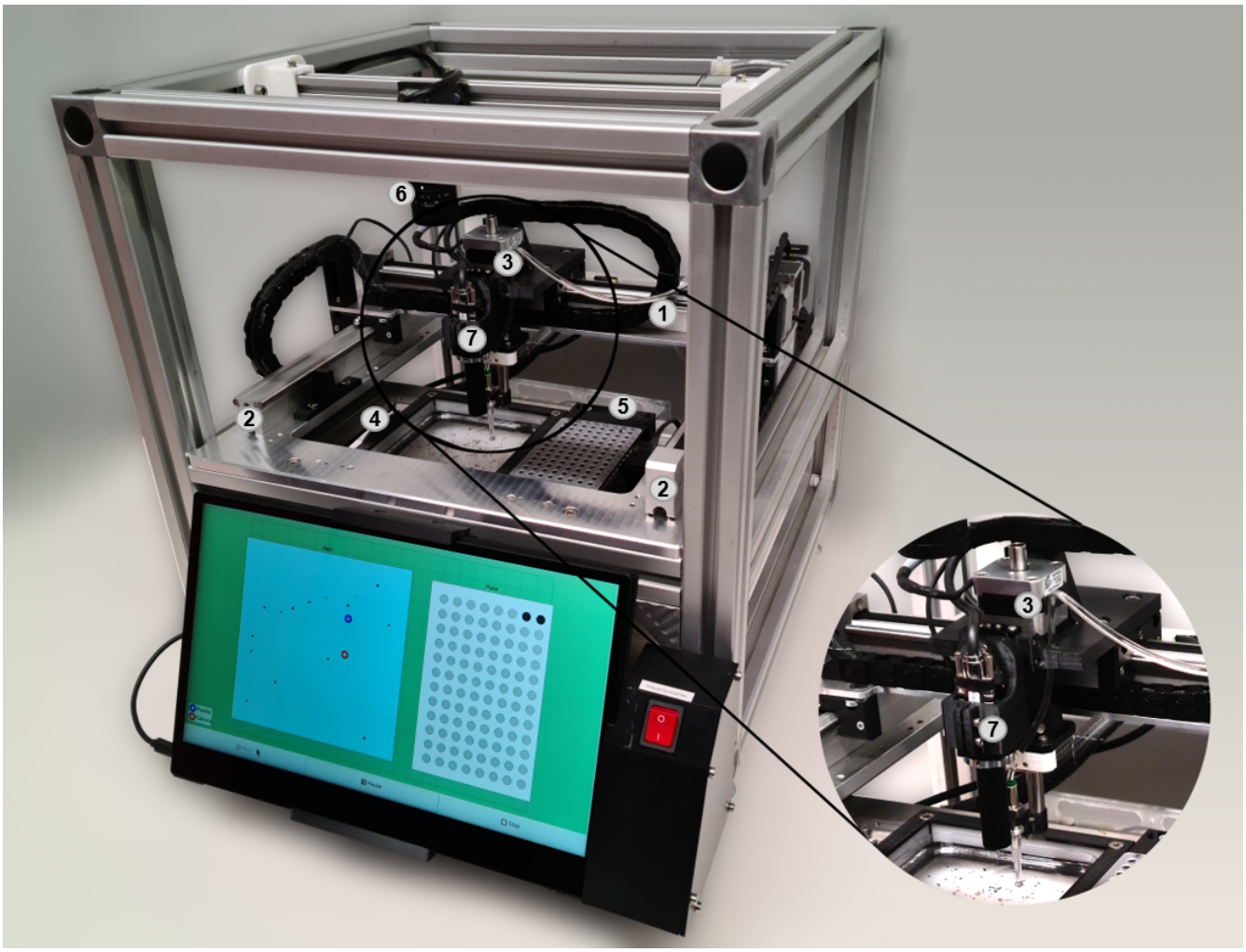


Figure 1. The DiversityScanner with 1: x-axis; 2: y-axis; 3: z-axis; 4: Petri dish; 5: 96-well microplate; 6: Overview camera (C1), 7: Detail camera (C2). The electronics box with Raspberry Pi, motor control unit, and the syringe pump are in the lower part of the sorting robot and therefore not visible in this view. The status of both, insect position determination and status of the sorting process are displayed on a touch screen, where the sorting process can also be started and stopped.

2.1 Transport System

The x- and y-axes of the robot are realised by LEZ1 linear drives (Isel AG, Eichenzell, Germany) and connected to the outer frame of the robot at half height. Both linear drives are driven by high-precision stepper motors with little tolerance to ensure good positioning accuracy. The y-axis is connected orthogonally to the shaft slide of the x-axis and is transported by it. The shaft slide of the y-axis transports the camera (C2) and the z-axis with the suction hose. In order to move the suction hose in the z-direction (= up and down) the z-axis is driven by an AR42H50 spindle drive with stepper motor (Nanotec Electronic GmbH & Co. KG, Feldkirchen, Germany). All three axes are controlled by a single TMCM-3110 motor controller (Trinamic, Hamburg,

Germany) that allows for precise, fast, and smooth movements.The motor controller is protected from water and ethanol droplets by locating it in a box at the bottom of the robot. The transport system is controlled by a Raspberry Pi single-board computer that was programmed in Python, specially developed for the sorting robot. In order to pick up insects from a Petri dish and discharge them in a well of a 96-well microplate a suction hose with a pipette tip is positioned by the transportation system. The hose is connected to a LA100 syringe pump (Landgraf Laborsysteme HLL GmbH, Langenhagen, Germany), that is also controlled by the Raspberry Pi. The sorting process is illustrated in Figure 2.

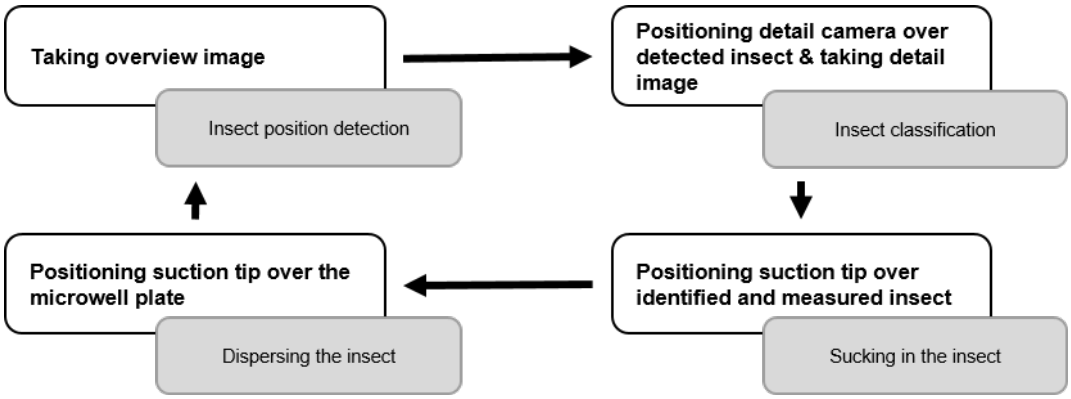


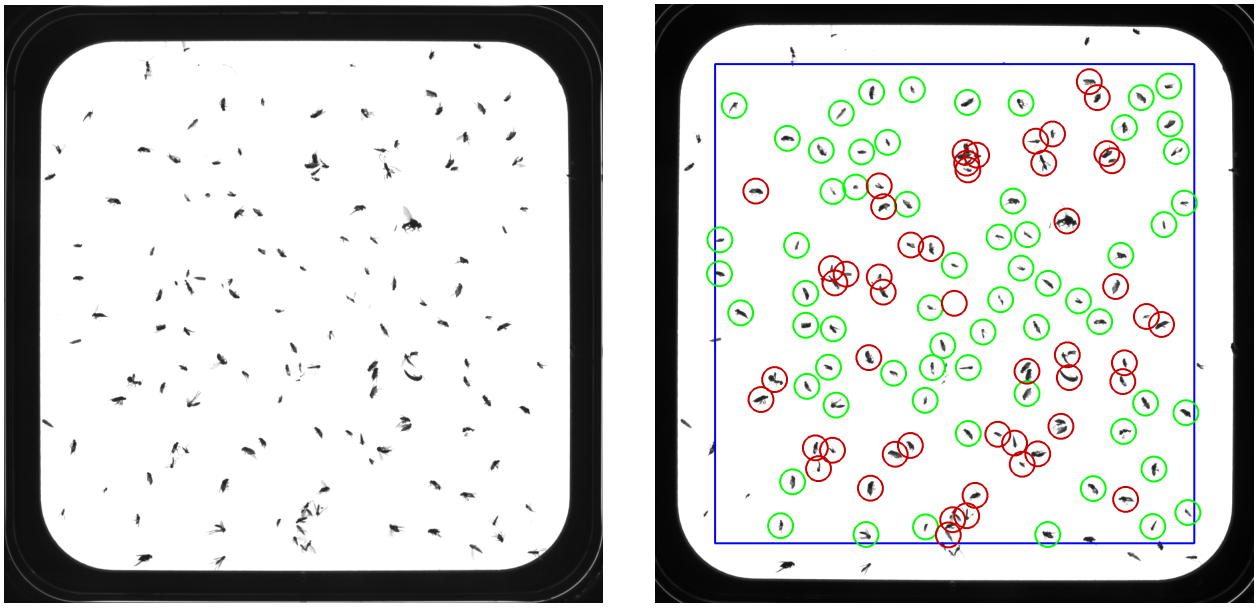
Figure 2. Process-chain for the classification and sorting process.

The sorting system includes two cameras with different lenses: the overview camera (C1) and the detailed view camera (C2). The first camera (C1) is a Ximea MQ042CG-CM camera with a CK12M1628S11 lens (Lensation GmbH, Karlsruhe, Germany) with a focal length of 16mm and an aperture of 2.8 is positioned directly above the Petri dish to take a detailed overview image of all insects inside. This image is used for detecting insects and their position within the Petri dish for the sorting process (see Figure 3a). The second camera (C2) is a Ximea MQ013CG-E2 with a telecentric Lensation TCST-10-40 lens with a magnification of 1x. This camera has to be moved by the x and y axes of the robot above the position of an insect to take a detailed image for classification and measuring (Figures 4 (a) and 6).

2.2 Image Processing Software

Three different software algorithms are used: The first algorithm determines the position of each object within the square Petri dish. The second one measures the length and volume of each insect. The third algorithm is based on an artificial neural net to classify insects into different classes.

Determination of Object Position: Most objects are insects, or parts of insects, but there can also be debris or other objects. After the overview image is taken, various image processing operations have to be performed to detect the objects: (1) A median filter removes noise from the image, (2) a conversion from a RGB-image to a gray scale image is performed, (3) an adaptive threshold filter segregates the objects, and (4) a contour finder identifies the boundaries of all objects. Two conditions must be



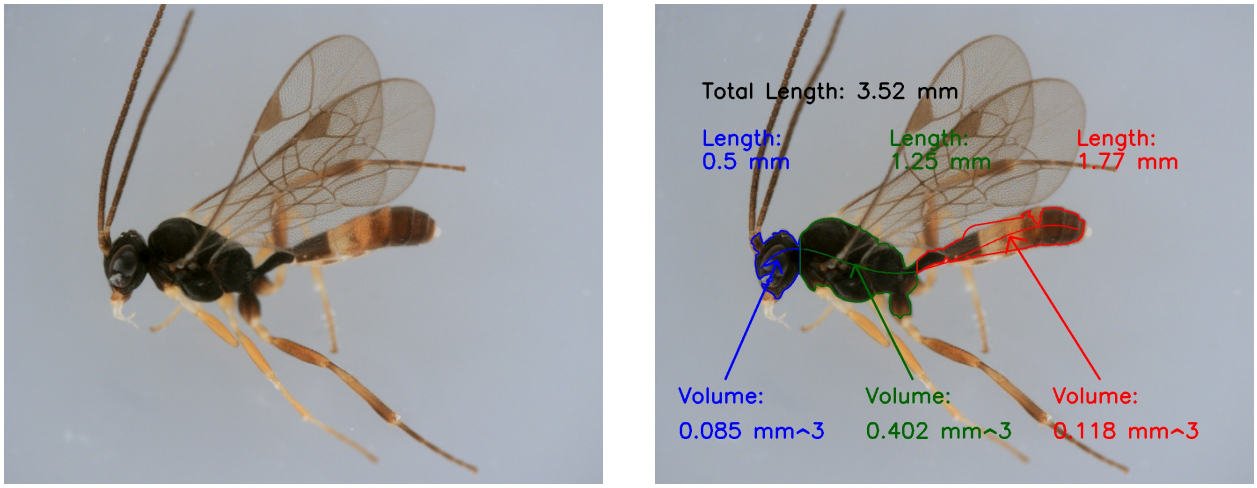
(a) The square Petri dish has a size of 120x120mm

(b) Blue line defines the area in which the object positions are determined (10 mm from edge); circles represents detected objects (green=meet size and distance conditions for imaging and movement.

Figure 3. Sample image obtained with the detail camera (C2) before (a) and after (b) processing.

met for objects to be detected: first, their size must be within a specified interval, and second, the distance between an object and neighbouring objects must exceed a minimum threshold value. If a cluster of objects is present, then the objects in the cluster fall below the specified minimum distance and are therefore not considered until they are separated. This ensures that only one single object is picked up during pipetting. Additionally, an accessible area within the Petri dish has been defined that has a distance of ten millimeters from the edge to ensure that the insects can be reached (blue line in Figure 3 (b)). Finally, all objects are color-coded. The coordinates of the detected objects are stored in a list, which is then used to control the position of the pipetting tip. After an object is removed, a new overview image is taken to determine the new coordinates of the objects, as they might have moved due to the pipetting of an object. This position identifying process continues until no more objects are detected or all wells of the 96-well microplate are filled with one insect each.

Object Dimensions: Lengths and volume of insect bodies are useful for estimating biomass. Several image processing operations are used to make such measurements. First, the contour is determined using morphological operators. Only those surfaces are selected which have a minimum value. If more than one surface is found (e.g. two body parts of the same specimen separated by a light area), they are connected so that there is only one contour. Within this contour, points are placed randomly, which are used to create a regression. The more points are used, the more accurate the regression and thus the estimate of the insect length is. To find the dividing lines of the head, thorax and abdomen, straight lines are placed at right angles to and along the regression line. Only those points of a line are considered that lie within the contour in the process. Subsequently,



(a) Hymenoptera Ichneumonidae. The image size is 6.4x4.8mm.

(b) Wasp with body length and volume measurement of the head, mesosoma and metasoma

Figure 4. Specimen image obtained with the detail camera (C2) before (a) and after (b) processing.

the dividing line between the head and thorax or between the thorax and abdomen is determined by examining the changes in length. To estimate the volume, a straight line is drawn through each body part. After that additional perpendicular straight lines are drawn which must be within the body contour. Now the distance and length of the straight lines can be used to determine the volume slice by slice. The lengths and volumes of the individual body parts as well as the total length are displayed on the screen of the sorting robot and the measurements are stored. All operations use the free OpenCV program library (version 4.5.1) and Python scripts (version 3.8.6). Currently, volume estimates are more accurate for body parts that are rotationally symmetrical. This works relatively well for Hymenoptera, but yields less precise measurements for Diptera.

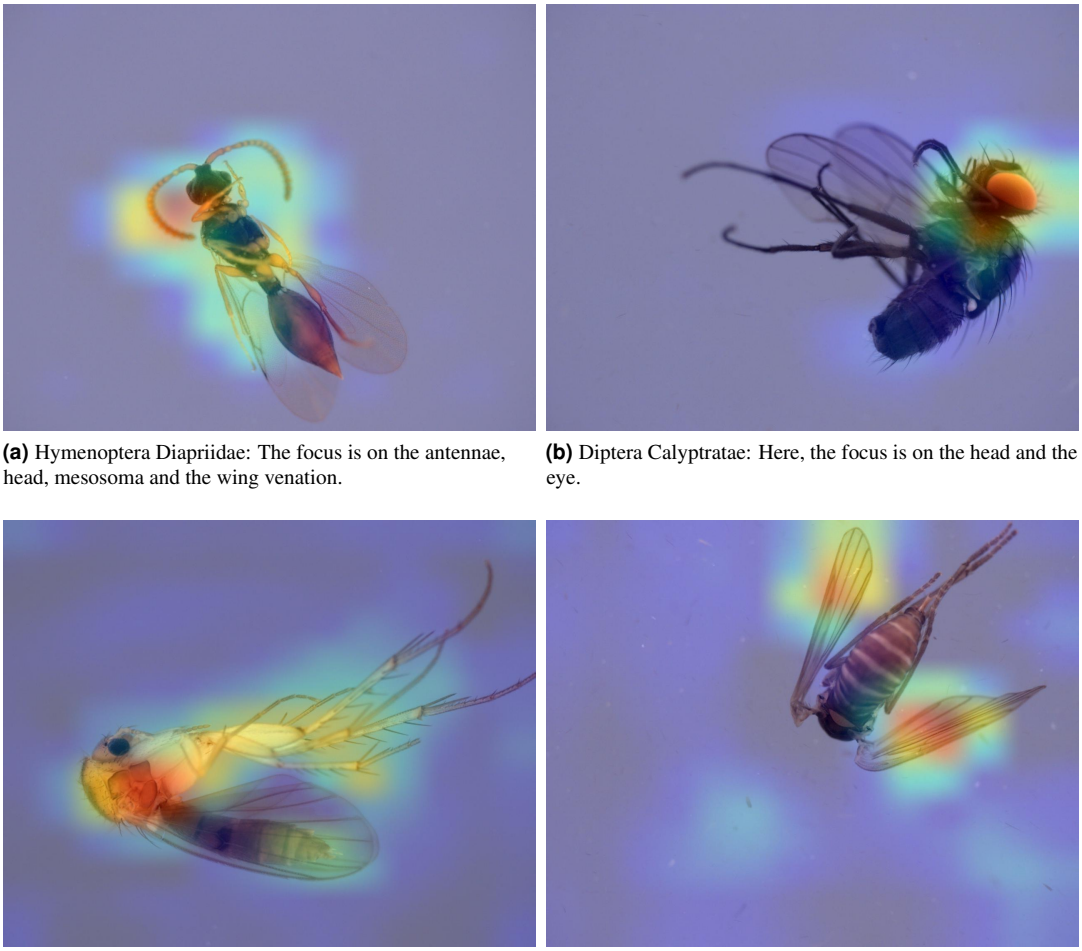
2.3 Insect Classification

In order to recognize different classes of insects and identify specimens to classes, machine learning algorithms based on convolutional neural nets (CNN) were applied.

Data Set: We here used 4,325 color images in 15 classes for training and 1,115 images for testing (Table 1). The images were obtained with the detailed camera for insects from 5 Malaise trap samples: 3 from Germany near Rastatt, Kitzing and Framersbach and 2 from Italy (Province of L'Aquila: Valle di Teve and Foresta Demaniale Chiarano-Sparvera). The images reflected the abundances of each taxon that are typical for Malaise trap samples (Karlsson et al., 2020). Only the common classes are covered by the trained CNN. Insects that do not belong to these were assigned to residual class (N = 693), which also included images of body parts (mainly legs and wings).

Data Augmentation: Data augmentation was performed to increase the number of images and the invariance within a class. The following image processing operations were applied randomly to the images: rotation, width shift, height shift, shear, zoom, horizontal flip and fill mode nearest.

Network Architecture: The VGG19 architecture was used as a base model for classification (Simonyan and Zisserman, 2014). The model was initialized with pre-trained ImageNet weights and the last layer was removed. For the new classification layer, a global average pooling, a dense layer with 1024 units and a reLU-activation, and a linear layer with a dropout rate of 0.4 were added. For the final classification, a softmax and a L2-regularization with a value of 0.02 are applied. In total the model has about 20.5 million parameters and the input size of an image is 224x224 pixels. The number of nodes in the last layer corresponds to the number of classes in the experiment. For training, the parameters of the original model were frozen and only the classification layer was trained. Afterwards, the whole model was optimized, where training was applied to all layers. Class activation maps were obtained by a global average pooling layer to illustrate the decisive features used by the neural network (Figure 5 a-d).



(a) Hymenoptera Diapriidae: The focus is on the antennae, head, mesosoma and the wing venation.

(b) Diptera Calyptratae: Here, the focus is on the head and the eye.

(c) Diptera Keroplatidae and Mycetophilidae: The focus is on the thorax and the legs

(d) Diptera Psychodidae: In this class the focus is only on the wings

Figure 5. Class activation maps for specimens belonging to four different insect classes.

Setup: The model is implemented in Keras (version 2.4.3) based on Tensorflow (version 2.2.1) and all experiments are conducted in the Python programming language (version 3.8.6). The networks were trained on a single board computer (Nvidia, Santa Clara, California, USA) as well as on more powerful GPUs using the online tool Colabatory. The working principles of the robot are illustrated in a video clip: https://www.youtube.com/watch?v=ElJ5VSHa4OI.

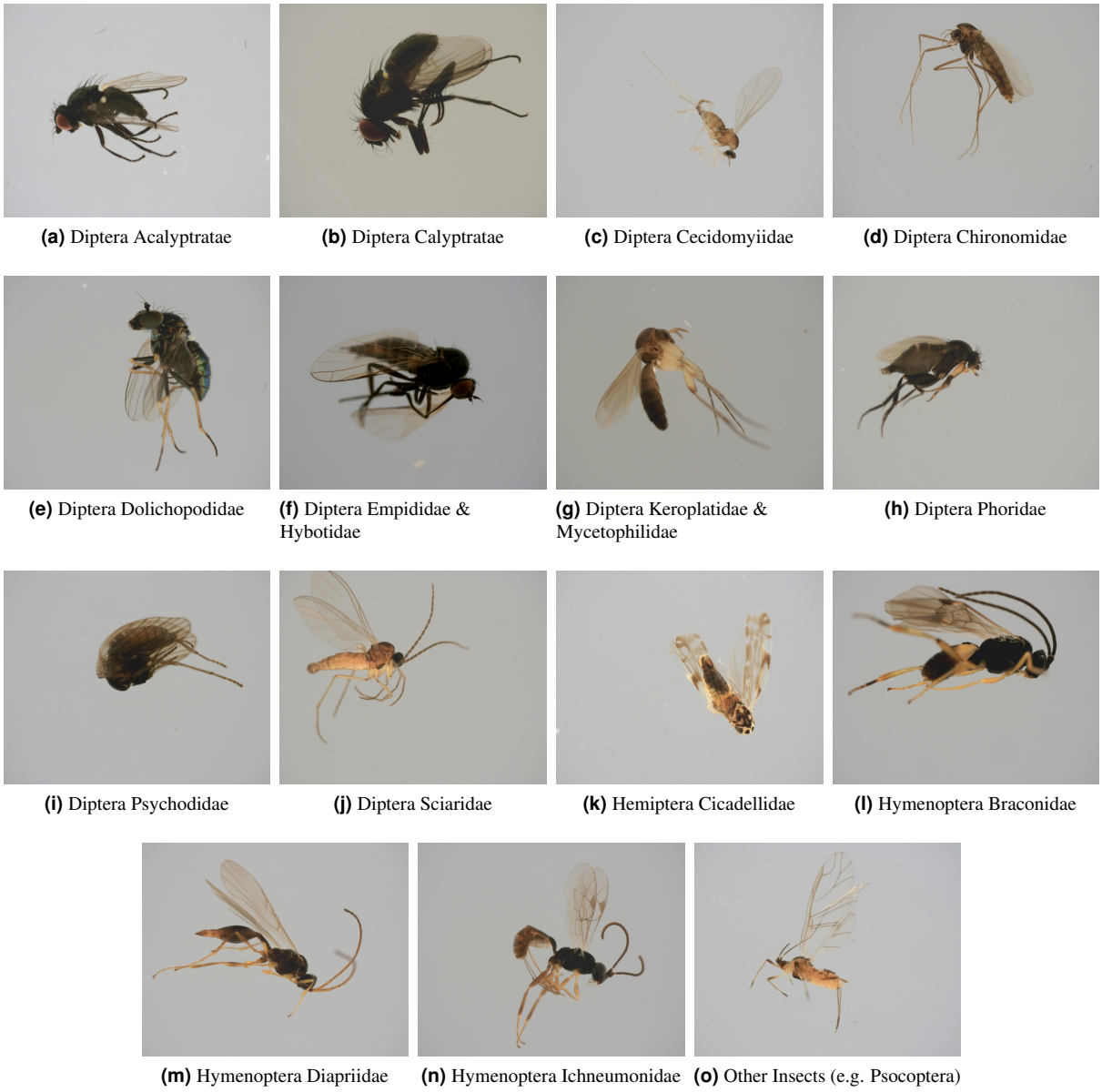
3 RESULTS

Currently, the sorting robot can pipette insects up to 3 mm length (Figure 6 a-o), as larger insects do not fit through the pipetting tip. Detected insects can be classified by the algorithm into 14 different classes of insects. All other insect classes and non-insect objects are combined in the class "other" (Table 1).

Class (Taxon)	Number of images	Result	Class (Taxon)	Number of images	Result
Diptera Acalyptratae	594	91%	Diptera Psychodidae	129	89%
Diptera Calyptratae	79	83%	Diptera Sciaridae	363	92%
Diptera Cecidomyiidae	467	91%	Hemiptera Cicadellidae	137	100%
Diptera Chironomidae	192	97%	Hymenoptera Braconidae	113	82%
Diptera Dolichopodidae	140	86%	Hymenoptera Diapriidae	255	100%
Diptera Empididae & Hybotidae	446	87%	Hymenoptera Ichneumonidae	133	75%
Diptera Keroplatidae & Mycetophilidae	440	99%	Other	693	81%
Diptera Phoridae	837	97%			

Table 1. Classification results, classes and number of images available for training, validation, and testing.

The best classification result was achieved for the classes of Hymenoptera Diapriidae and Hemiptera Cicadellidae, where all insects were correctly classified, whereas insects of the class Hymenoptera Ichneumonidae had the lowest correct classification rate. The DiversityScanner currently supports two sorting processes: Either one insect after the other can be classified and sorted until the last well of the 96-well microplate is filled or only insects of a predefined class are pipetted into the plates.



(a) Diptera Acalyptratae **(b)** Diptera Calyptratae **(c)** Diptera Cecidomyiidae **(d)** Diptera Chironomidae

(e) Diptera Dolichopodidae **(f)** Diptera Empididae & Hybotidae **(g)** Diptera Keroplatidae & Mycetophilidae **(h)** Diptera Phoridae

(i) Diptera Psychodidae **(j)** Diptera Sciaridae **(k)** Hemiptera Cicadellidae **(l)** Hymenoptera Braconidae

(m) Hymenoptera Diapriidae **(n)** Hymenoptera Ichneumonidae **(o)** Other Insects (e.g. Psocoptera)

Figure 6. Sample images for the 15 classes.

4 DISCUSSION

The use of CNNs for the identification of charismatic species is starting to be routine (Tabak et al., 2019; Stowell et al., 2019; Fairbrass et al., 2019). However, these methods are largely unavailable for small invertebrates although they comprise most of the multicellular animal species and contribute many ecosystem services. The main problem is not the availability of invertebrate samples, but the lack of CNNs which cannot be trained because there are few sets of training images. We believe that the best strategy for changing this undesirable situation is by combining automated imaging with DNA barcoding. Each

"DiversityScanner" robot can process several invertebrate samples per day. Each contains thousands of specimens that can be imaged with minimal manual labour. After imaging, the specimens are moved into microplates for DNA barcoding. Once barcoded, the images can be re-labeled with approximately species-level identifications given that most animal species have species-specific barcodes, or they can be assigned to family- or genus-level based on DNA sequence similarities. Common species, genera, and families rapidly acquire sufficiently large sets of images that can then be used for training CNNs. Indeed, for the most common "classes" of insects in Malaise traps, we already had enough images for creating such networks after partially imaging only five Malaise trap samples.

Some biologists doubt that CNNs will be sufficiently powerful to yield species-level identifications for closely related species and we agree that it remains unclear whether species-level identifications can be achieved (Ärje et al., 2020b; Knyshov et al., 2021). However, we predict that the main limitation will not be the machine-learning algorithms but the number and quality of the training images. Fortunately, high-quality cameras are now readily available and it becomes feasible to obtain large numbers of images at different magnifications and in different orientations. This is particularly straightforward once specimens have been pre-sorted to putative species based on DNA barcodes. As illustrated by the BIODISCOVER robot, inserting these specimens into a cuvette allows for generating a large number of images from different angles even for rare species. Once a sufficiently large number of species are covered by the CNN, the DiversityScanner should then be able to identify many specimens based on images only. DNA barcoding would be restricted to those specimens that are not identifiable based on visual information. These are more likely to belong to rare and new species so that the DiversityScanner would also become a powerful tool for discovering new species in samples. This ability would be particularly important in the 21st century because new species continue to arrive at well-characterized sampling sites. Some of these species recently shifted their distribution in response to climate change while others may be new anthropogenic introductions. For both it would be desirable to have an early-warning system based on automated workflows.

The design of DiversityScanner focused on reproducibility and low-cost (<5,000 €), so that eventually many robots can sort a large number of insects simultaneously in many laboratories. This makes the robot an attractive alternative to manual sorting and identification. Currently, the robot only handles small invertebrates, because they are particularly abundant and pose the largest challenge for manual handling. For particularly hyperabundant taxa the DiversityScanner can be instructed to only transfer a limited number of specimens. For example, the robot could extract only 1-2 microplates' worth of non-biting midges (Chironomidae), if this taxon is too abundant. This ability to only find and move some taxa also helps with implementing clade-specific molecular recipes (e.g., different DNA extraction or PCR recipes) and restricting barcoding to either males or females given that often only one sex has species-specific morphological differences. With additional modification, the DiversityScanner will also be able to handle larger specimens. For example, the suction tip diameter can increased or one can install a gripper with a sensor-based feedback system.'

We thus believe that robots like the DiversityScanner have the potential to solve some of the problems that Robert May mentioned when he bemoaned our lack of biodiversity knowledge. Automation can expedite biodiversity discovery and monitoring of neglected "dark taxa". Of course, the DiversityScanner can only address some of the challenges. For example, newly discovered species will still need description and described species identification. Moreover, even when all species have been described, we will still know very little about the ecological roles that these species play in the ecosystems. Fortunately, molecular approaches to diet analysis and life history stage matching can help fill these gaps (Yeo et al., 2018; Srivathsan et al., 2019b). However, given that ecosystems routinely consist of thousands of species, automation and data analysis will also be needed for high-throughput species interaction research.

Acknowledgments We would like to specially thank Daniel Moser and Stefan Vollmannshauser for their support with manufacturing the mechanical parts and helping us with connecting the electronic circuits. Mr Leshon Lee prepared the video documenting the working principles of the DiversityScanner. Funding was provided by the Center for Integrative Biodiversity Discovery at the Museum für Naturkunde Berlin.

REFERENCES

Ärje, J., Melvad, C., Jeppesen, M. R., Madsen, S. A., Raitoharju, J., Rasmussen, M. S., Iosifidis, A., Tirronen, V., Gabbouj, M., Meissner, K., et al. (2020a). Automatic image-based identification and biomass estimation of invertebrates. *Methods in Ecology and Evolution*, 11(8):922–931. https://doi.org/10.1111/2041-210X.13428.

Ärje, J., Raitoharju, J., Iosifidis, A., Tirronen, V., Meissner, K., Gabbouj, M., Kiranyaz, S., and Kärkkäinen, S. (2020b). Human experts vs. machines in taxa recognition. *Signal Processing: Image Communication*, 87:115917. https://doi.org/10.1016/j.image.2020.115917.

Bar-On, Y. M., Phillips, R., and Milo, R. (2018). The biomass distribution on earth. *Proceedings of the National Academy of Sciences*, 115(25):6506–6511. https://doi.org/10.1073/pnas.1711842115.

Borkent, A., Brown, B. V., et al. (2015). How to inventory tropical flies (diptera)-one of the megadiverse orders of insects. *Zootaxa*, 3949(3):301–322.

Brown, B. V. (2005). Malaise trap catches and the crisis in neotropical dipterology. *American Entomologist*, 51(3):180–183. https://doi.org/10.1093/ae/51.3.180.

Brown, B. V., Borkent, A., Adler, P. H., de Souza Amorim, D., Barber, K., Bickel, D., Boucher, S., Brooks, S. E., Burger, J., Burington, Z. L., et al. (2018). Comprehensive inventory of true flies (diptera) at a tropical site. *Communications biology*, 1(1):1–8. https://doi.org/10.1038/s42003-018-0022-x.

Chamblin, M. A., Paasch, R., Lytle, D., Moldenke, A., Shapiro, L., and Dietterich, T. (2011). Design of an automated system for imaging and sorting soil mesofauna. *Biological Engineering Transactions*, 4(1):17–41. https://doi.org/10.13031/2013.37174.

Creedy, T. J., Ng, W. S., and Vogler, A. P. (2019). Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy. *Ecology and evolution*, 9(6):3105–3116. https://doi.org/10.1002/ece3.4839.

Fairbrass, A. J., Firman, M., Williams, C., Brostow, G. J., Titheridge, H., and Jones, K. E. (2019). Citynet — deep learning tools for urban ecoacoustic assessment. *Methods in ecology and evolution*, 10(2):186–197. https://doi.org/10.1111/2041-210X.13114.

Feng, L., Bhanu, B., and Heraty, J. (2016). A software system for automated identification and retrieval of moth images based on wing attributes. *Pattern Recognition*, 51:225–241. https://doi.org/10.1016/j.patcog.2015.09.012.

Hartop, E., Srivathsan, A., Ronquist, F., and Meier, R. (2021). Large-scale integrative taxonomy (lit): resolving the data conundrum for dark taxa. *bioRxiv*. https://doi.org/10.1101/2021.04.13.439467.

Hebert, P. D., Braukmann, T. W., Prosser, S. W., Ratnasingham, S., DeWaard, J. R., Ivanova, N. V., Janzen, D. H., Hallwachs,

W., Naik, S., Sones, J. E., et al. (2018). A sequel to sanger: amplicon sequencing that scales. *BMC genomics*, 19(1):1–14. https://doi.org/10.1186/s12864-018-4611-3.

Karlsson, D., Hartop, E., Forshage, M., Jaschhof, M., and Ronquist, F. (2020). The swedish malaise trap project: a 15 year retrospective on a countrywide insect inventory. *Biodiversity Data Journal*, 8. https://doi.org/10.3897/BDJ.8.e47255.

Knyshov, A., Hoang, S., and Weirauch, C. (2021). Pretrained convolutional neural networks perform well in a challenging test case: Identification of plant bugs (hemiptera: Miridae) using a small number of training images. *Insect Systematics and Diversity*, 5(2):3. https://doi.org/10.1093/isd/ixab004.

Lepek, H., Nave, T., Fleischmann, Y., Eisenberg, R., Karlin, B. E., and Tirosh, I. (2020). Method for sex sorting of mosquitoes and apparatus therefor. US Patent App. 16/479,648.

May, R. M. (2011). Why worry about how many species and their loss? *PLoS Biol*, 9(8):e1001130. https://doi.org/10.1371/journal.pbio.1001130.

Perre, P., Faria, F. A., Jorge, L., Rocha, A., Torres, R. d. S., Souza-Filho, M., Lewinsohn, T., and Zucchi, R. A. (2016). Toward an automated identification of anastrepha fruit flies in the fraterculus group (diptera, tephritidae). *Neotropical entomology*, 45(5):554–558. https://doi.org/10.1007/s13744-016-0403-0.

Pfriem, A., Pylatiuk, C., Alshut, R., Ziegener, B., Schulz, S., and Bretthauer, G. (2012). A modular, low-cost robot for zebrafish handling. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 980–983. https://doi.org/10.1109/EMBC.2012.6346097.

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., Barrio-Amorós, C. L., Salazar-Valenzuela, D., and Prost, S. (2018). Real-time dna barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience*, 7(4):giy033. 10.1093/gigascience/giy033.

Sarpola, M., Paasch, R., Mortensen, E., Dietterich, T., Lytle, D., Moldenke, A., and Shapiro, L. (2008). An aquatic insect imaging system to automate insect classification. *Transactions of the ASABE*, 51(6):2217–2225. https://doi.org/10.13031/2013.25375.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivathsan, A., Hartop, E., Puniamoorthy, J., Lee, W. T., Kutty, S. N., Kurina, O., and Meier, R. (2019a). Rapid, large-scale species discovery in hyperdiverse taxa using 1d minion sequencing. *BMC biology*, 17(1):1–20. https://doi.org/10.1186/s12915-019-0706-9.

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S. N., Wong, J., Yeo, D., and Meier, R. (2021). Minion barcodes: biodiversity discovery and identification by everyone, for everyone. *bioRxiv*. https://doi.org/10.1101/2021.03.09.434692.

Srivathsan, A., Nagarajan, N., and Meier, R. (2019b). Boosting natural history research via metagenomic clean-up of crowdsourced feces. *PLoS biology*, 17(11):e3000517. https://doi.org/10.1371/journal.pbio.3000517.

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., and Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380. https://doi.org/10.1111/2041-210X.13103.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., et al. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590. https://doi.org/10.1111/2041-210X.13120.

Wang, W. Y., Srivathsan, A., Foo, M., Yamane, S. K., and Meier, R. (2018). Sorting specimen-rich invertebrate samples with cost-effective ngs barcodes: Validating a reverse workflow for specimen processing. *Molecular ecology resources*, 18(3):490–501. https://doi.org/10.1111/1755-0998.12751.

Watsa, M., Erkenswick, G. A., Pomerantz, A., and Prost, S. (2020). Portable sequencing as a teaching tool in conservation and biodiversity research. *PLoS biology*, 18(4):e3000667. https://doi.org/10.1371/journal.pbio.3000667.

World Economic Forum's Global Risk Initiative, T. (2020). The global risks report 2020. `http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf`.

Yeo, D., Puniamoorthy, J., Ngiam, R. W. J., and Meier, R. (2018). Towards holomorphology in entomology: rapid and cost-effective adult-larva matching using ngs barcodes. *Systematic entomology*, 43(4):678–691. https://doi.org/10.1111/syen.12296.

Author Contributions Conceptualization: R.M., T.v.R., L.W. and C.P.; writing original draft preparation: L.W., R.M. and M.G.; writing review and editing: C.P., R.M., S.S., P.C., M.B. and T.v.R.; visualization: L.W. and M.G. and S.S.; supervision: C.P., R.M. and T.v.R.; funding acquisition: C.P., T.v.R. and R.M.; L.W. and C.P. contributed equally. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement All image data that were used for training and testing are accessible at the media repository of the Museum für Naturkunde Berlin.

All files for printing the robot parts and the software code are accessible at the repository of the Open Science Framework.

Supplementary Materials Video