Received: 29 April 2022 | Accepted: 17 March 2023

DOI: 10.1111/2041-210X.14104

RESEARCH ARTICLE

Methods in Ecology and Evolution

Classifying the unknown: Insect identification with deep hierarchical Bayesian learning

Sarkhan Badirli[1] | Christine Johanna Picard[2] | George Mohler[3] | Frannie Richert[3] | Zeynep Akata[4] | Murat Dundar[3]

[1]Computer Science Department, Purdue University, West Lafayette, Indiana, USA

[2]Department of Biology, Indiana University—Purdue University, Indianapolis, Indiana, USA

[3]Computer and Information Science Department, Indiana University—Purdue University, Indianapolis, Indiana, USA

[4]Cluster of Excellence Machine Learning, University of Tübingen and Max Planck Institute for Intelligent Systems, Tübingen, Baden-Württemberg, Germany

Correspondence
Murat Dundar
Email: mdundar@iupui.edu

Present address
Murat Dundar, , 723 W. Michigan StreetIndianapolis, Indiana 46202, USA

Funding information
National Science Foundation, Grant/Award Number: ATD-2124313 and IIS-1252648

Handling Editor: Andrew Mahon

Abstract

1. Classifying insect species involves a tedious process of identifying distinctive morphological insect characters by taxonomic experts. Machine learning can harness the power of computers to potentially create an accurate and efficient method for performing this task at scale, given that its analytical processing can be more sensitive to subtle physical differences in insects, which experts may not perceive. However, existing machine learning methods are designed to only classify insect samples into described species, thus failing to identify samples from undescribed species.
2. We propose a novel deep hierarchical Bayesian model for insect classification, given the taxonomic hierarchy inherent in insects. This model can classify samples of both described and undescribed species; described samples are assigned a species while undescribed samples are assigned a genus, which is a pivotal advancement over just identifying them as outliers. We demonstrated this proof of concept on a new database containing paired insect image and DNA barcode data from four insect orders, including 1040 species, which far exceeds the number of species used in existing work. A quarter of the species were excluded from the training set to simulate undescribed species.
3. With the proposed classification framework using combined image and DNA data in the model, species classification accuracy for described species was 96.66% and genus classification accuracy for undescribed species was 81.39%. Including both data sources in the model resulted in significant improvement over including image data only (39.11% accuracy for described species and 35.88% genus accuracy for undescribed species), and modest improvement over including DNA data only (73.39% genus accuracy for undescribed species).
4. Unlike current machine learning methods, the proposed deep hierarchical Bayesian learning approach can simultaneously classify samples of both described and undescribed species, a functionality that could become instrumental in biodiversity monitoring across the globe. This framework can be customized for any taxonomic classification problem for which image and

DNA data can be obtained, thus making it relevant for use across all biological kingdoms.

KEYWORDS
biodiversity, classification, computer vision, deep learning, machine learning, undescribed species

1 | INTRODUCTION

Understanding biodiversity for insects requires both discovery and identification. Insects are one of the largest and most diverse animal groups on the planet with an estimated 5.5 million species, yet only 20% are described (Stork, 2018) and many are disappearing faster than they can be identified (Costello et al., 2013), making it difficult to assess biodiversity. Once an insect is collected, a taxonomist will identify the insect to its lowest taxonomic level possible based on existing morphological character keys (Buck et al., 2009). Traditionally, taxonomists use identification keys describing physical characters to identify a given specimen. This presents a real-world challenge because undescribed species cannot be uniquely identified by existing characters, and only through the comprehensive analysis of characters could one distinguish undescribed from described species.

DNA-based technologies, such as barcoding (sequencing certain conservative yet sufficiently variable regions of the genome) (Hebert et al., 2003), have helped confirm new species in cases where the DNA sequence variation exceeds the established intraspecific variation, or in cases where species are not distinguishable by their phenotypic characters (cryptic species) (Burns et al., 2008). While such powerful DNA-based methods are able to provide an estimate of biodiversity, they do not alone contribute to the knowledge base. The DNA Barcode Database (BOLD) (Ratnasingham & Hebert, 2007, 2013), with a search of the order Diptera yields 2.4 million records (DNA sequences) and 126,000 BINs (barcode indexed numbers, a representative measure of sequence diversity based on thresholds). However, only 25,000 species have been identified out of the 126,000 BINs represented in BOLD Diptera record set. This indicates that while it is true DNA is facilitating the discovery of new species, identification is occurring at a much slower rate. Species identification is made difficult by a lack of taxonomists given the vast diversity of insects, and the fact that the art of traditional taxonomy is on the decline (Hopkins & Freckleton, 2002; Lee, 2000; Orr et al., 2020). Therefore, a novel way to efficiently scale both the discovery and identification of existing as well as new species is crucial for making the assessment of biodiversity feasible.

Machine learning methods can be leveraged to find intricate patterns and relationships in data, which have corresponding labels that indicate group membership, for example genus and species, for classification and outlier detection tasks. When combined with images, thus forming a computer vision task, machine learning can extricate subtle insect morphological characters, which are then used to classify described species and identify undescribed species. Classification models that use only images are enticing because images are significantly easier to obtain than DNA samples. While image-based classification models cannot yet compete with DNA-based methods, recent studies demonstrated that machine learning approaches for image-based taxonomic identification could eventually achieve human-expert level accuracy (Milošević et al., 2020; Raitoharju & Meissner, 2019; Valan et al., 2019). Advances in machine learning have led to a surge of interest in entomology, a domain for which there are many challenges machine learning methods could help overcome.

Specifically, deep learning approaches (a subset of machine learning) have been utilized in pest-detection (Ding & Taylor, 2016; Sun et al., 2018), digitization of museum collections (Hedrick et al., 2020; Meineke et al., 2020), measuring invertebrate biodiversity (Mayo & Watson, 2007; Wang et al., 2012), investigating the plant-insect interactions (Tran et al., 2018) and many more applications (Høye et al., 2021). Deep learning methods have also been employed for the more challenging task of automatically detecting species in video and time-lapse images (Pegoraro et al., 2020). The main drawback of these methods is that they are focused on specific insect groups and considered only a very small number of subgroups.

Traditional machine learning models are inherently limited by incomplete insect data repositories available for training the model; it is often impossible to create a training repository with a complete set of insect species represented for a given taxon. For example, some insect species are rare or not yet described, and thus well-characterized training images of insects from these species cannot be obtained. Moreover, insects pose an additional challenge due to their morphologically distinct life stages where in some cases, the insects' immature stages look significantly different than their adult morphologies.

At the heart of the issue is that insect identification requires a method that can *both* classify samples from described species and identify samples from undescribed species or species not included in the training data. Many existing methods assume that all possible species are represented in the training data set (Geng et al., 2020); such methods, therefore, could not identify samples from rare or unidentified species. Additionally, current classification methods were used with relatively small data sets and would not scale well to problems with a large number of classes (Geng et al., 2020). Furthermore, such approaches were restricted to detecting if an insect sample is an outlier and could not differentiate between different types of outliers (Bendale & Boult, 2016; Perera & Patel, 2019; Scheirer & Boult, 2016). This limits their usefulness in entomology as the

Insecta class contain a large number of similar species and the hierarchy in its taxonomy necessitates outlier differentiation.

In order to accomplish both tasks, we adapted the generalized zero-shot learning (ZSL) setting (Xian et al., 2018), with genus and species taxonomic levels as auxiliary information, to test whether this method could be used to facilitate the identification of new insect species. ZSL indicates that the model will be tested not only on samples from species seen during training (referred to as described species throughout) but also on samples from species not seen during training (referred to as undescribed species throughout); we analysed both the model's accuracy in assigning samples from described species to their correct species, and its accuracy in assigning samples from undescribed species to their most likely genus of origin.[1] In brief, we sought to answer whether recent advances in deep learning and computer vision can extract subtle yet potentially discernible morphological characters that when combined with DNA sequence data, facilitate more accurate identification of insects of described species and aid the discovery of insects of undescribed origin, further grouped by genera (see Figure 1).

2 | MATERIALS AND METHODS

In this section, we first describe the DNA and image data used in this study. Next, the use of deep learning models for extracting information-rich feature vectors from insect images and DNA barcodes is described followed by a description of splitting the feature data training, validation and testing, and for simulating undescribed species. We then lay out the details for our novel hierarchical Bayesian classifier, including a description of transductive and inductive machine learning approaches, which join the image and DNA feature vectors to boost model accuracy over using either image-only or DNA-only feature vectors as input to the classifier. Finally, we detail a bioinformatics baseline model that was included for comparison.

2.1 | Data collection

Our study used paired insect image and DNA sequence data obtained from the Barcode of Life Data System (BOLD) (Ratnasingham & Hebert, 2007, 2013), from four major Insecta orders: Diptera, Coleoptera, Lepidoptera and Hymenoptera. Tables 1 and 5 and Figure 2 detail the breakdown of the dataset by order. Each insect data point contained a 658 bp DNA barcode sequence (*cytochrome oxidase subunit I*—COI); an image; and additional information such as country of origin, life-stage, order, family, subfamily, genus and species names.

The raw images were in full colour (3 colour channels—red, green and blue) and generally had a width of 640 pixels by a height of 300–1000 pixels. Only images that had matching DNA barcodes were included, and each image was manually inspected so that low quality images, duplicate images, images containing incomplete insect bodies or immature pigmentation, and missing images (e.g. just a label was present) were removed. Only species with a minimum of 10 images within a single barcode index number (BIN) were included.

BOLD differs from other genetic databases (e.g. Agarwala et al., 2018) in that it accepts data for unidentified or unknown organisms. BOLD's DNA-based grouping algorithms will first assign a BIN to the unidentified sample: the BINs are closely aligned (but not perfectly) with species groupings. The BOLD database then translates the sample's DNA sequence to its protein sequence and searches its database for a species or genus match. BOLD will assign the sample to a species if its sequence contains less than 1% divergence from a

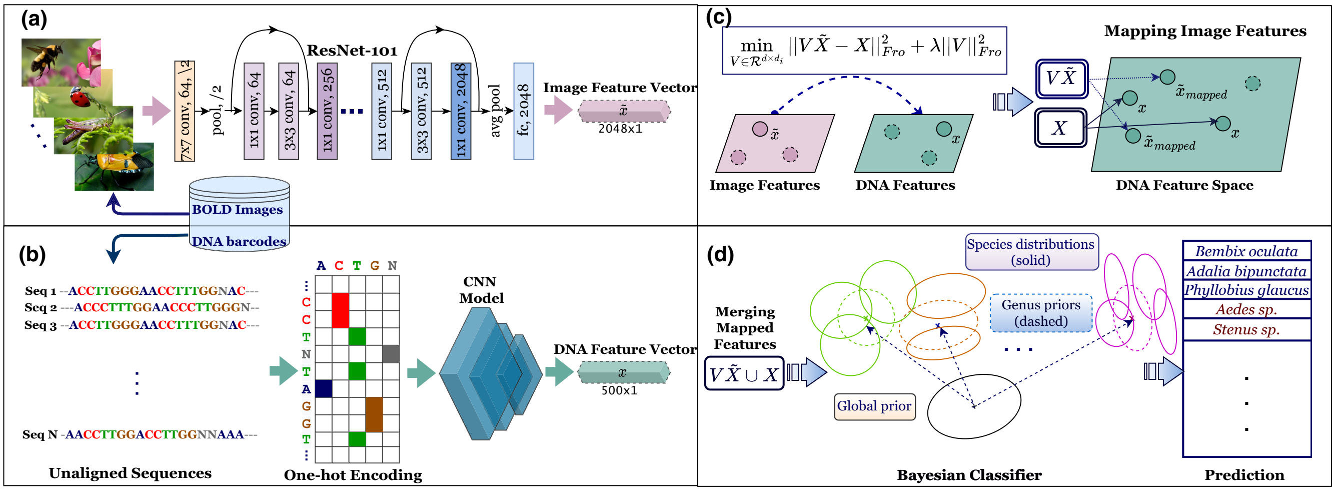


FIGURE 1 Deep hierarchical Bayesian classification with described and undescribed species. (a) Image feature vectors (2048-dimensional) were obtained from a pretrained ResNet-101 (He et al., 2016) model. (b) DNA feature vectors (500-dimensional) were obtained from a custom CNN model. (c) The optimal way to merge the image and DNA features was to first map image features to DNA feature space as learned by transductive ridge regression. (d) Hierarchical Bayesian model was trained on the merged training set $V\tilde{X}_{\text{train}} \cup X_{\text{train}}$ and used for classification. A test sample then was either assigned to one of the described species or identified as a new species belonging to one of the described genera (indicated with genus name followed by sp.).

TABLE 1 A breakdown of the data set by order.

Order	# Genera	# Species	# Samples
Diptera (true flies)	63	108	2270
Coleoptera (beetles)	164	329	4764
Hymenoptera (sawflies, wasps, bees and ants)	59	189	3173
Lepidoptera (butterflies and moths)	82	414	22,641
Totals	368	1040	32,848

reference sequence, and will assign a genus if the divergence is between 1% and 3%; otherwise, the sample is left unidentified. As of 6 June 2022, the BOLD Insecta database had 7,192,313 records with DNA sequences, while only 2,774,112 of these records had species names, indicating that the majority of records remain unidentified. While the BOLD database is essential for the discovery of new species, it has a consequential limitation: it does not facilitate the identification of such new species beyond the measures described above.

2.2 | Feature extraction

Prior to implementing the hierarchical Bayesian method, we employed deep learning models to extract meaningful features from the raw insect image and DNA barcode data. Humans can look at an insect and identify many distinguishing morphological features such as "lime-green scales" or "setose antenna"; computers are given an insect image file, which is a $\text{width} \times \text{height} \times 3$ data matrix representing pixels filled with RGB values between 0 and 255, and are asked to do the same. Humans come up with a list of text describing the insect features, whereas computers produce feature vectors, a set of numeric values that have been learned to best distinguish one class from another when used in classification models. In addition to being discriminative, representations learned from deep learning are often significantly smaller in size than the original raw data representation (especially for images), which leads to better scalability when used in downstream machine learning models.

In particular, we used a pretrained ResNet-101 (He et al., 2016) model, a 101-layered residual convolutional neural network (CNN), in the Python PyTorch library, to extract image feature vectors (2048-dimensional) from the insect images, following the standard guidelines (Xian et al., 2018) (see Section 4 for more detail regarding how ResNet-101 was chosen among possible other ResNet models). The power of the ResNet suite of models is that they have been pretrained on over one million ImageNet (Russakovsky et al., 2015) images representing 1000 different classes, the majority of which are not insects; pre-training on such a significant quantity of images allows the ResNet models to learn to identify general image characteristics such as edges, blobs, colours and texture patterns, much like a human would if asked to describe an image. The feature vectors the ResNet-101 model returns are condensed representations of the raw images in terms of these learned general image characteristics and are used in place of the raw images due to their discriminative capacity in the downstream hierarchical Bayesian classifier. The raw insect images were transformed prior to use in the ResNet-101 model as is standard and necessary to ensure their size and value ranges were compatible with those in the pretraining ImageNet images. The transformations included resizing the images to $256 \times 256 \times 3$, centre-cropping the resized images to $224 \times 224 \times 3$ and finally normalizing the resized images by colour channel using the ImageNet image pixel means $[0.485, 0.456, 0.406]$ and standard deviations $[0.229, 0.224, 0.225]$. We did not fine-tune the ResNet-101 model on our dataset, after concluding that fine-tuning did not have a significant impact on model accuracy; for further discussion of the experiments we conducted that led us to this conclusion, see Supporting Information, Fine-Tuning the ResNet-101 Model section.

DNA barcode sequence feature vectors (500-dimensional) were extracted from a convolutional neural network (CNN) (LeCun et al., 1989; Russakovsky et al., 2015) architecture developed specifically for this problem in Python using the Tensorflow-Keras API. For use as input to this CNN model, the DNA barcodes were each transformed into 658×5 arrays which indicated the presence (=1) or absence (=0) of 4 DNA tokens (adenine, guanine, cytosine and thymine) and other (which indicated a missing or ambiguous token) for each position in the 658 length sequence. In the CNN, we used three blocks of convolutional layers, each followed by batch normalization and 2D max-pooling. The output of the third convolutional layer was flattened and batch normalized before feeding the data into the fully-connected layer with 500 units, the output of which were the learned features. A softmax layer completed the CNN architecture. See Supporting Information for further detail.

2.3 | Handcrafted versus deep features

Handcrafted features are predominantly data-agnostic and manually designed by experts to overcome specific challenges, like occlusion and variations in scale and illumination (Nanni et al., 2017) or to characterize a priori known characteristics (shape, colour etc.), whereas deep features are more generic and data-driven, given that they are learned directly from input images (Bora et al., 2016; LeCun et al., 2015). Human experts evaluate features qualitatively whereas computers require quantitative features. In the case of a large-scale insect classification task, handcrafting features to capture subtle characteristics of insect species represented by dozens of dichotomous keys may not be very practical.

On the other hand, a deep network is by default trained to learn quantitative features that will maximize classification accuracy of the network for a specific task. Deep features can be extracted at multiple levels of abstraction: the initial layers of the neural network (NN) resemble Gabor filters and tend to learn low level image features such as edges and blobs (Figure 3a), that is transferable to many different categories of objects and tasks (Yosinski et al., 2014), while deeper layers learn more complex relationships that can represent

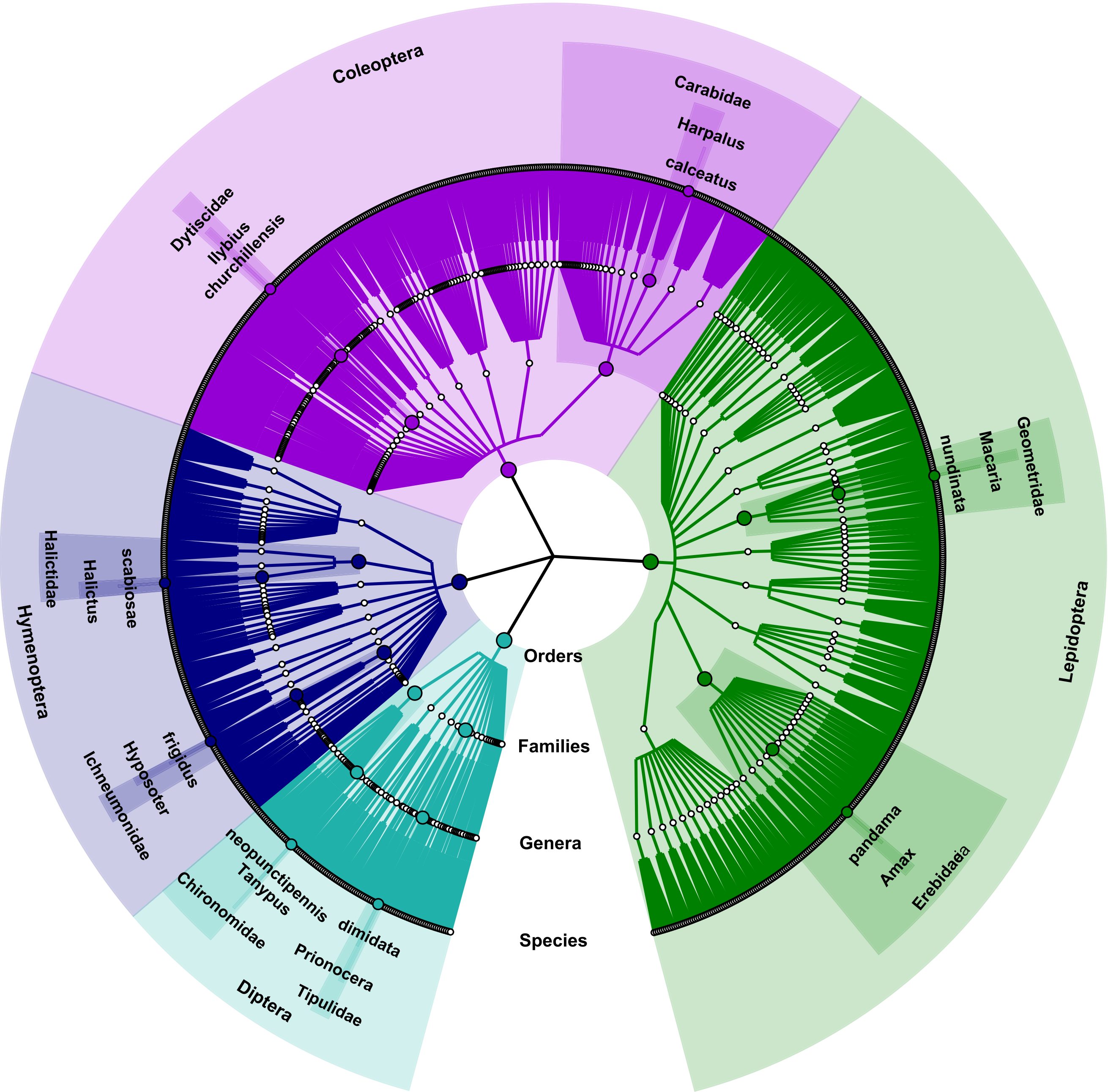
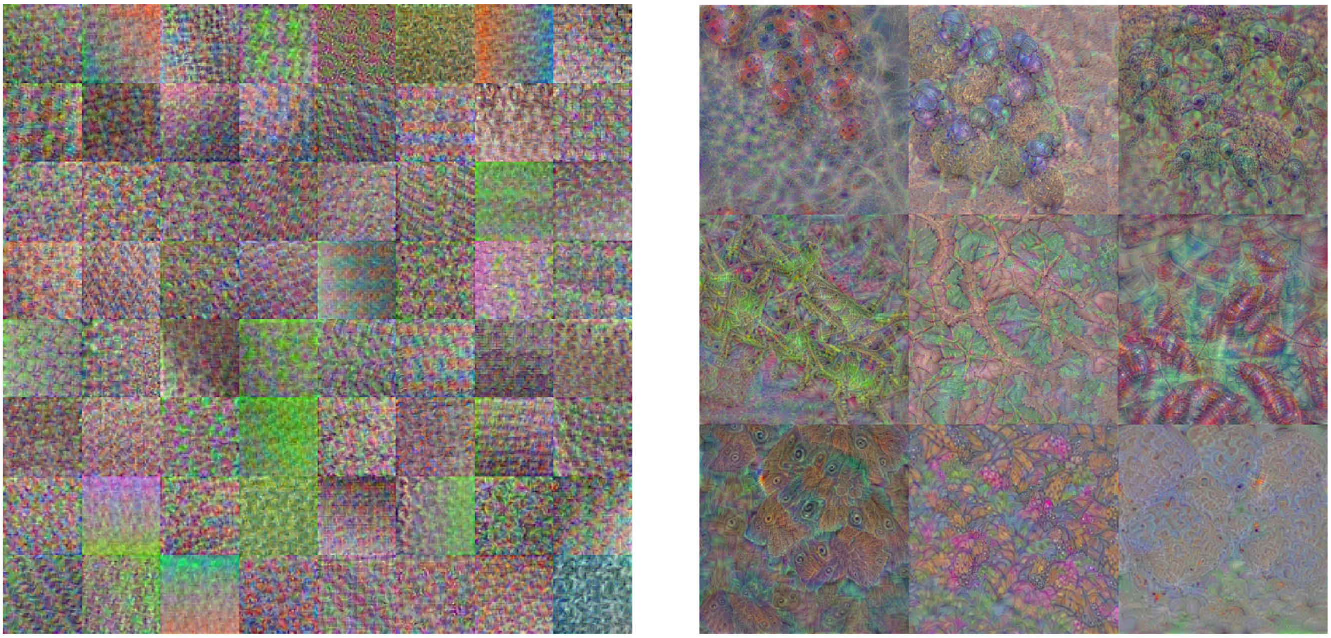


FIGURE 2 Phylogenetic tree of the four orders from the dataset. Two species were randomly chosen from each order, with their complete taxonomic hierarchy illustrated.

high level semantics (Figure 3b). Capturing semantics shared across instances of the same objects in the feature space mitigates intra-class variability in the image space (Chan et al., 2015), which is vitally important for fine-grained image classification tasks involving thousands of classes and limited number of samples per class.

From a taxonomist perspective, the ultimate goal would be to have an algorithm that both identifies samples from undescribed species and lists their distinguishing morphological character(s) to aid in the process of new species discovery and an identification key created. However, efforts to match semantic descriptors (e.g. dichotomous keys) with quantitative features (e.g. deep features) often introduce significant feature redundancy and noise, which may negatively impact classifier performance. Although the lack of interpretability of these learned features still remains as a big hurdle confronting deep learning models, recent advances in self-supervised representation learning (Caron et al., 2021), attention and saliency maps (Simonyan et al., 2013) are expected to gradually close this interpretability gap between deep-learned and hand-crafted features. Such methods can learn to identify distinguishing areas in images, some without the help of image labels, and it is plausible that in the future, deep learning algorithms will be able to identify such fine-grained discriminative areas in images, possibly with the integration



(a) Low level ResNet-101 features

(b) High-level features corresponding to classes (from top-left to bottom-right) ladybug, dung beetle, weevil, grasshopper, walking stick, cockroach, ringlet, monarch, lycaenid.

FIGURE 3 Visualizing features of ResNet-101 pretrained on the ImageNet 1K dataset. Images are generated to maximize activations at corresponding nodes. (a) Rectified linear unit nodes of the intermediate residual block. (b) Output nodes corresponding to nine insect related classes in the ImageNet 1K dataset.

of known morphological characters (see Section 5), and they would automatically output a new identification key per taxonomic group defined by the user (i.e. genus-level key) that a human could use in the future.

2.4 | Merging image and DNA data

Combining different data modalities within a single deep network (Nanni et al., 2017; Yang et al., 2019) can potentially learn representations that can better correlate different data types and improve predictive performance. However, this type of fusion network increases the complexity of the network, because the network requires additional layers to reconcile different data modalities in the representation space. Given the very large yet non-exhaustively defined number of training classes and limited number of samples per class, we avoided training an end-to-end network for combining image and DNA data and instead chose to use features independently extracted from two different networks, one for each of image and DNA data, and let the hierarchical Bayesian model harness these features through creating dependencies across classes via local and global priors.

The DNA and image features were considered in the model separately and together to show the model's enhanced performance when both were included. We experimented with both inductive and transductive methods for merging the DNA and image features. Note that any methods for combining DNA and image data have a potentially consequential flaw: they require test samples to also have an image with an accompanying DNA barcode, which may not be feasible in all applications. Inductive learning methods learn rules from labelled training data (e.g. data for which the species label is known) that are then applied to unlabelled test data, and are traditionally referred to as supervised learning methods as the learning process is directed by the data labels. Transductive learning methods, on the other hand, learn from the combined set of labelled training data and unlabeled test data without utilizing the data labels in the learning process, and produce rules that are applied to the test data; because transductive methods learn patterns from both the training and test data, transductive methods are optimized for certain tasks (such as clustering or grouping), especially when the test data are significantly different from the training data.

We considered both a standard inductive approach to merging these data sets (concatenating their features) and a Bayesian approach (which considers the sum of normalized likelihood vectors generated from Bayesian classifiers of the image and DNA features, respectively). More detail regarding the implementation of these two inductive methods can be found in Supporting Informations 3.1.2 and 3.1.3.

The transductive approach learns a linear mapping from image feature space to DNA feature space by using Ridge regression with the training and test data sets as input, without using any class labels (i.e. without knowing a sample's genus or species). The value of using both the training and test data sets for learning the optimal mapping is that the test data set alone contains samples from undescribed species, which gives transductive methods the theoretical and as detailed later, empirical, advantage over inductive methods in the task of identifying undescribed species.

Prior to using either the merged or individual data sets in the model, we reduced their dimension to 500 using principal component analysis (PCA) to ensure the input data dimensions were low enough to be feasible for use in the hierarchical Bayesian model. The transductive approach is outlined in Figure 1c and further detail regarding these processes is in Supporting Information.

2.5 | Training, validation and test data

Machine learning classification models are generally built by an iterative process of tuning the model with training data and measuring the trained model accuracy on validation data, until the most optimal final model has been found. This final model is then tested by measuring its accuracy on a test data set, which has not been seen previously by the model, to gauge the model's generalizability to future data sets. To prove our model was a viable method for identifying undescribed species, our model was validated and tested with data sets that contained samples for species the model had not seen before. Since the BOLD data we collected, by design, contained no true undescribed species data, test undescribed species data had to be simulated as described in Figure 4. For validation, the training undescribed species were split in a similar manner into described and undescribed species. Some insect species had multiple images, each capturing a different view of the insect (e.g. ventral and dorsal views). All insect species with multiple images were restricted to the training set, leaving 27 of the described species with no representatives during testing. In the test data set, there were a total of 4965 samples from 770 described species and 8463 samples from 243 undescribed species (see Table 2).

2.6 | Hierarchical Bayesian model

Insect species have a predefined taxonomic hierarchy (order < family < subfamily < genus < species). Despite the shared morphological characters at each level of the hierarchy, which carry valuable information for classification tasks, the taxonomy is often overlooked in machine learning methods. A hierarchical Bayesian model was recently introduced in computer vision for zero-shot classification of object classes (Badirli, Akata, et al., 2021; Badirli et al., 2020) using visual attributes (Badirli et al., 2020) or DNA (Badirli, Akata, et al., 2021) as auxiliary information. Here, a similar model was developed to identify both described and undescribed species by replacing the visual attributes with the inherent insect taxonomy (species < genus). The hierarchical Bayesian model (HBM) is in essence just a collection of class probability distributions, one for each described species and one for each genus (collectors for undescribed species), which are parameterized by training data statistics and hyperparameters that have been tuned to optimize model performance. The HBM takes as input a test sample's feature vector, and returns a likelihood value from each probability distribution; the test sample is assigned to the class with the highest likelihood, either a described species or genus, where assignment to a genus indicates the sample is most likely from an undescribed species.

The taxonomic hierarchy for our HBM is encapsulated in a three-layer generative model, in contrast to standard Bayesian approaches which have a two-layer generative model. In standard Bayesian approaches, the top layer uses a prior to generate class distributions and the bottom layer uses class distributions to generate individual data instances of each class. In our three-layer hierarchical generative model, classes were defined by species. To distinguish between prior distributions at different layers, we introduced the notion of global and genus priors as follows. At the top of the hierarchy is a global prior distribution that generates genus prior distributions in the middle layer. In the middle layer are genus prior distributions that generate distributions of species at the bottom layer. At the bottom are species distributions that generate feature vectors of individual insect samples. The intuition behind this generative model (see Figure S2) is that species sharing similar haplotypes would group in the same phenotypic space characterized by feature vectors. The generative model design is given below (see Figure S3 for a graphical depiction)

$$\begin{aligned}&\boldsymbol{x}_{jik} \sim N(\boldsymbol{\mu}_{ji}, \Sigma_j), \quad \boldsymbol{\mu}_{ji} \sim N(\boldsymbol{\mu}_j, \Sigma_j \kappa_1^{-1}), \quad \boldsymbol{\mu}_j \sim N(\boldsymbol{\mu}_0, \Sigma_j \kappa_0^{-1}),\\ &\Sigma_j \sim W^{-1}(\Sigma_0, m),\end{aligned} \tag{1}$$

where j, i, k represent indices for genus priors, described species and data instances, respectively.

We assume that the data instance $\boldsymbol{x}_{jik}$ comes from a Gaussian distribution defined by mean $\boldsymbol{\mu}_{ji}$ and covariance matrix Σ_j; note species from the same genus share the same covariance matrix Σ_j to

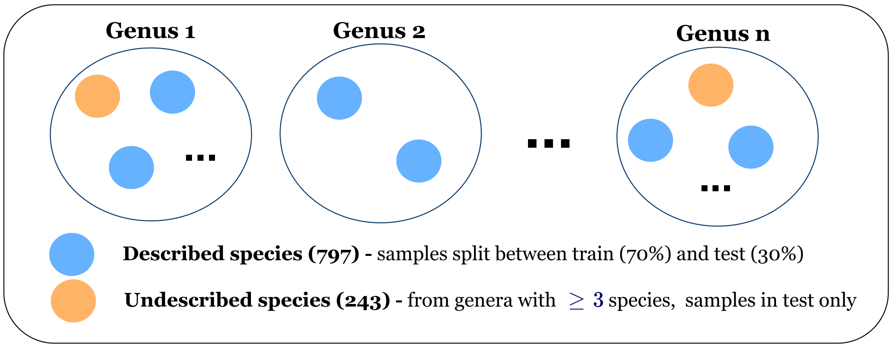

Genus 1
Genus 2
Genus n
...
...
...
Described species (797) - samples split between train (70%) and test (30%)
Undescribed species (243) - from genera with ≥ 3 species, samples in test only

FIGURE 4 For genera with ≥3 species, one-third were randomly assigned undescribed while the rest were assigned described; only the test set contained undescribed species to test the model's ability to identify them.

preserve conjugacy. The data instances are generated independently and conditioned on the hyperparameters of both global and genus priors. The hyperparameter κ_1 is a scaling constant that adjusts the dispersion of the described species means ($\boldsymbol{\mu}_{ij}$) around the center of their corresponding genus prior. A larger κ_1 leads to smaller variations in species means from the mean of their corresponding genus prior, suggesting a fine-grained (harder to distinguish) relationship among species sharing the same genus. Conversely, a smaller κ_1 dictates coarse-grained (easier to distinguish) relationships among species sharing the same genus.

Each genus prior is Gaussian and characterized by the parameters $\boldsymbol{\mu}_j$ and Σ_j. The mean vectors of the genus priors are in turn distributed according to a Gaussian prior and κ_0 is a scaling constant that adjusts the dispersion of these mean vectors around the mean vector μ_0 of the global prior. A smaller value for κ_0 suggests that genus centres are expected to be farther apart from each other whereas a larger value suggests they are expected to be closer to each other. On the other hand, Σ_0 and m dictate the expected shape of the described species distributions, as under the inverse Wishart distribution assumption, the expected covariance is $E(\Sigma|\Sigma_0, m) = \frac{\Sigma_0}{m-D-1}$, where D is the dimension of the data. The minimum feasible value of m is equal to $D+2$; the larger m is, the less individual covariance matrices will deviate from the expected shape.

TABLE 2 Train-test split.

	Train	Test (described)	Test (undescribed)
# of samples	19,420	4965	8463
# of species	797	770	243

2.7 | Hyperparameters and statistics

The described species and genus prior posterior predictive distributions (PPDs), derived in Supporting Information, are a function of both hyperparameters and sufficient statistics. The hyperparameter μ_0 (mean of the global prior) is the mean of the described species means, while the hyperparameter Σ_0 is the mean of the described species covariance matrices scaled by s (also referred to as the pooled covariance); each were calculated with training data. The species-specific sufficient statistics, also calculated with training data, are: $\bar{x}_{jc}$ (mean vector), n_{jc} (number of samples), and $\bar{\Sigma}_{jc}$ (covariance matrix), where c represents the current described species. The algorithm with pseudo-code for deriving these values is in Supporting Information.

It is worthwhile to note here that the genus prior PPD formulations relied upon the quantity of described species data available to the model during training, which did not encompass all possible species, both as a result of simulating undescribed species by removing some species from the training set and as a result of having an incomplete data set. Therefore, the genus prior, while described as fully as possible given the training data, was not a complete representation of all its member species.

2.8 | Classification

Bayesian models classify a sample by assigning it the label of the class whose distribution maximizes its likelihood. The set of class labels included both the described species and genera, so classification involves the simultaneous comparison of likelihoods across all

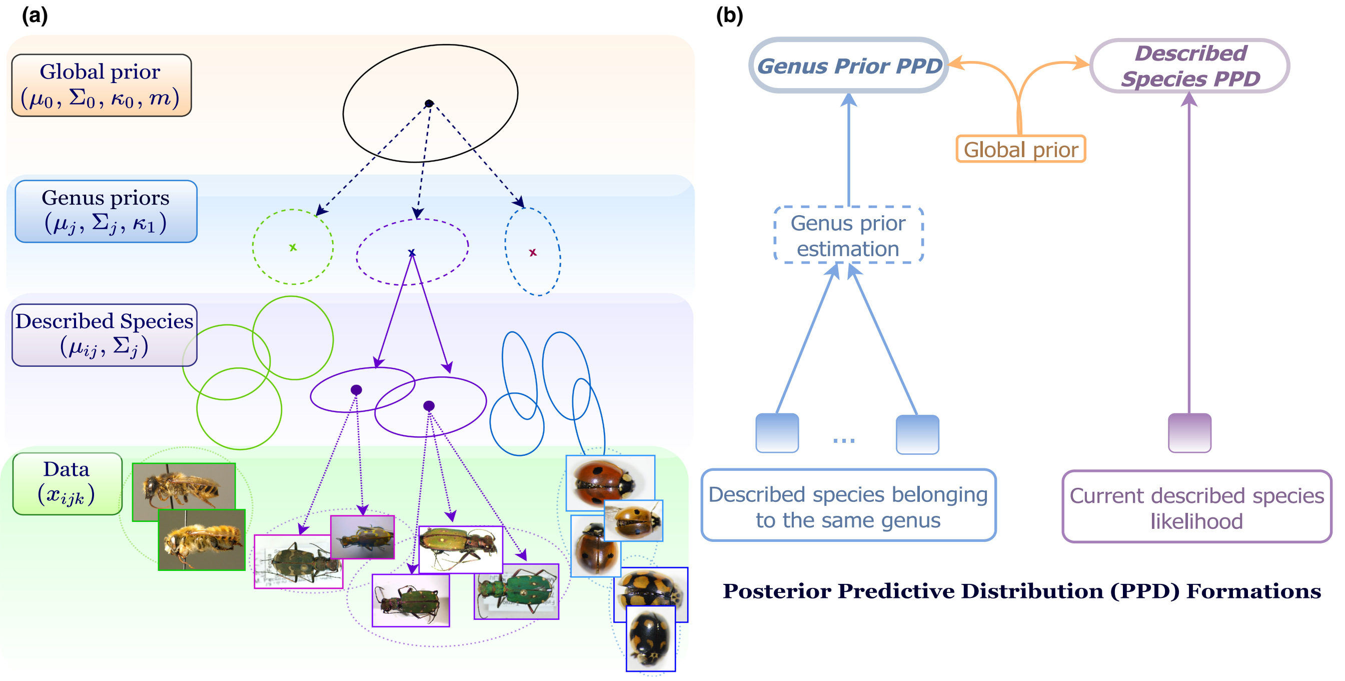

(a)
Global prior
$(\mu_0, \Sigma_0, \kappa_0, m)$
Genus priors
$(\mu_j, \Sigma_j, \kappa_1)$
Described Species
(μ_{ij}, Σ_j)
Data
(x_{ijk})
(b)
Genus Prior PPD
Described Species PPD
Global prior
Genus prior estimation
...
Described species belonging to the same genus
Current described species likelihood
Posterior Predictive Distribution (PPD) Formations

FIGURE 5 Hierarchical Bayesian model. (a) Generative model. Hyperparameters are defined in the Methods section. (b) Posterior predictive distribution formation for described species and genera.

described species and all genus prior likelihoods. The likelihoods of the samples' merged feature vectors were evaluated with the class-conditional posterior predictive distributions (PPDs). Genus priors comprise the middle layer of the 3-layer hierarchical Bayesian model while described species priors comprise the bottom layer, so their PPD derivations differed slightly (as outlined in Figure S3a and depicted in Figure 5). If the sample was assigned to a genus by the model, we predicted that the sample originated from an undescribed species; the genus labels were used to assess how accurate the model was in assigning undescribed species to their correct genus. If the sample was assigned a species by the model, we predicted that the sample originated from that described species.

2.9 | Optimization

The goal of our HBM was to classify test samples from described species to their respective species and test samples from undescribed species to their respective genus. The classification performance was assessed by the average described species accuracy (referred to throughout as just described species accuracy) and average undescribed species genus accuracy (referred to throughout as just undescribed species accuracy), as well as their harmonic mean, as shown in the following equation:

$$\begin{aligned}\text{Test Described Species Accuracy (S)} &= \frac{1}{770}\sum_{i=1}^{770}\frac{y_j}{n_j},\\ \text{Test Undescribed Species Accuracy, by Genus (US)} &= \frac{1}{134}\sum_{i=1}^{134}\frac{y_j}{n_j},\\ \text{Test Harmonic Mean} &= \frac{2\times S\times US}{S+US},\end{aligned}\tag{2}$$

where for class j, y_j is the the number of correctly classified samples and n_j is the number of total samples.

The hyperparameters κ_0, κ_1, m and s were tuned through cross-validation on the validation set to produce the maximum harmonic mean of the described and undescribed species validation accuracies which ensures the model would be capable of doing both tasks well when given test data (see Supporting Information for evaluated hyperparameter values). The harmonic mean, a standard form (Xian et al., 2018) for evaluating model performance, is an overall accuracy measure between the described and undescribed species accuracies, and is more representative than the usual average measure would be due to the fact that these two accuracy measures do not share a common denominator; since there are more described species (770) than genera (134), the described species accuracies would have been more dominant in the overall accuracy measure if the usual average would have been used, skewing model performance.

2.10 | Baseline bioinformatics approach

To show our HBM's performance against a more traditional distance-based method, we included results in Table 3 for the following Bioinformatics baseline approach which uses DNA barcode data only (no images), and was developed with Matlab's bioinformatics tool (The MathWorks, Inc.). For each described species, DNA nucleotide sequences were aligned using training samples available for that species. Aligned sequences were then used to compute a consensus nucleotide sequence. For each test sample, the Jukes–Cantor distance (Jukes & Cantor, 1969) between the test sample's sequence and the consensus sequences from each described species was found. Test samples were assigned to the described species with the minimum distance, only if the minimum distance was smaller than a designated threshold. If the minimum distance was larger than this threshold, the test sample was predicted to be from an undescribed species and assigned to the genus of the species with the minimum distance. The distance threshold was chosen by cross-validation.

2.11 | Experimental design

Several models were developed and tested for their classification accuracy. First, the Bayesian model was trained and tested on just the DNA feature vectors (HBM-DNA) and then just the image feature

TABLE 3 Hierarchical Bayesian model (HBM) classification results. For described species test samples, the mean of the species classification accuracy is reported. For undescribed test samples, the mean of the genus classification accuracy is reported. Best results are displayed in bold and the second-best results are underlined. Tr, Ts_d and Ts_{ud} represent train, test described and test undescribed data, respectively. Even when only 25% of the undescribed species test data is available for learning the mapping between image and DNA feature spaces, the HBM-DIT ($Tr + 25\% Ts_{ud}$) model still outperforms the image-only and DNA-only models.

Model	Data used	Undescribed	Described	Harmonic mean
Bioinformatics	DNA only	71.85	**98.65**	83.16
HBM	Image only	35.88	39.11	37.42
HBM	DNA only	73.39	96.15	83.24
HBM	DNA & Image—Inductive Combination (DIC)	77.26	97.26	86.25
HBM	DNA & Image—Inductive Likelihood (DIL)	**81.95**	98.21	**89.35**
HBM	DNA & Image—Transductive (DIT)—($Tr + Ts_d + Ts_{ud}$)	81.39	96.66	88.37
HBM	DIT—($Tr + 50\% Ts_{ud}$)	79.94	96.66	87.53
HBM	DIT—($Tr + 25\% Ts_{ud}$)	77.48	96.63	86.01

vectors (HBM-IMG), to prove that using both datasets together produced better accuracy than either dataset alone. Accuracies from the Bioinformatics-DNA baseline model were also included to show our method could improve on a more traditional method that relies on distance thresholds. Next, the model was trained and tested with the HBM model using both DNA and image data (DI) under inductive as well as transductive settings. There were two inductive settings considered: the standard approach, which combined their features (HBM-DIC), and the likelihood approach (HBM-DIL) (for more details on inductive approaches see Supporting Informations 3.1.2 and 3.1.3). Finally, there was the transductive setting, which found a mapping from image to DNA feature space to bring the two data sets together (HBM-DIT), which was tested under three different percentage inclusions of undescribed species data in the test data set (25%, 50%, and 100%) to show if having at least some test data in the transductive mapping was better than having none in comparison to methods which did not use a combination of DNA and image features.

3 | RESULTS

Classification accuracies for all model variations for both undescribed and described accuracies, and their harmonic means, are reported in Table 3.

It is clear from the results that classifiers, which used image data alone achieved minimal accuracy (39.11% described accuracy). As expected, DNA data proved to be informative for species classification. The bioinformatics baseline method had the highest described classification accuracy at nearly 99% while achieving an undescribed accuracy of 72%, a 10% reduction compared to the transductive model (HBM-DIT). In comparison to the Bioinformatics method, the HBM-DNA yielded a better undescribed accuracy, but slightly lower described accuracy.

Combining image and DNA data in all five HBM scenarios increased accuracy over image-only and DNA-only models, particularly for undescribed species test samples. Transductive (HBM-DIT) and heuristic inductive likelihood (HBM-DIL) methods performed best, with >88% harmonic means and 81% undescribed accuracies. While both inductive methods (HBM-DIC and HBM-DIL) performed reasonably well, they present a real-world challenge in that they require test samples to have both image and DNA data present, which may be difficult to obtain.

Under the transductive setting, the quantity of undescribed species available in the test data for learning the mapping between image and DNA feature spaces impacted accuracy. The last two rows of Table 3 display model performance utilizing decreasing portions of undescribed species data in the test data set. Note that the model was not tuned for these configurations and used the same optimal model parameters as the other HBM-DIT models. These findings show how the transductive method can outperform non-combined-data models even when only 25% of the available undescribed species data was included in the test set, and model performance increased as more undescribed testing data were included.

The transductive model, which included all of the available undescribed testing data, yielded 96.66% overall described classification accuracy with 4827/4965 correct classifications (Table 5). Unsurprisingly, the accuracy declined for undescribed samples, with three of the four orders having >81% undescribed sample accuracy which was still remarkably good. HMB-DIT misclassified the genera of many undescribed Diptera samples (Table 5). When examining the different family groups and their classification accuracies (Table 5), the Culicidae (the mosquitoes), Syrphidae (the hover flies) and Tipulidae (the crane flies) had the most misclassifications.

With Culicidae, 45/58 of the misclassifications were *Aedes vexans* records classified to the *Culex* genus. As a semi-independent test, the DNA sequence of a random *Aedes vexans* record from our dataset was passed through BLASTn (Altschul et al., 2021) in Genbank; the resulting BOLD records that populate were from the Culicinae subfamily, indicating the overlap in DNA sequences for species under the Culicinae subfamily may be obstructing classification (data not shown). For Syrphidae, 18 *Platycheirus neoperpallidus* records were misclassified to *Platycheirus clypeatus*. When random *P. neoperpallidus* records were aligned to other *Platycheirus* species, it was noted that there was a great deal of similarity with *P. quadratus*, a species not present in the training set (data not shown). Again, this demonstrated the need for a more representative training dataset to ensure accuracy within specific genera.

Misclassifications will also increase if there is a large amount of sequence divergence between species within a given family. For example, all 14 Tipulidae records which were misclassified belonged to a single species: *Tipula coloradensis*. The majority of the misclassifications were to the same subfamily (Tipulidae) but misclassified to the *Nephrotoma* genus; four of the 14 were misclassified to Syrphidae. What was of note with this dataset was that the training data contained three species of *Tipula*: (*T. caliginosa*, *T. salicetorum*, *T. shirakii*). Sequence similarities were calculated between the three species in the training set and *T. coloradensis*; what became apparent is that *T. salicetorum* and *T. caliginosa* were closely related (interspecific sequence similarity of 97%). In contrast, the sequence similarity of *T. coloradensis* with either *T. salicetorum* or *T. caliginosa* was 88%. Further, *T. shirakii* was perhaps the most different, with 85% sequence similarities from the remainder of the *Tipula* species included in this analysis (data not shown).

4 | DISCUSSION

While the most successful method in this study employed both image and DNA barcode data, the use of image-only methods or DNA-only methods each had varying levels of success. DNA lends strong support for new species identification if the sequence variation falls outside of the normal bounds of intraspecific variation. BOLD uses a cut-off of <1% sequence divergence to identify species to a reference specimen (which is itself identified as having

<2% sequence divergence for three or more records) and <3% to assign to a genus (Ratnasingham & Hebert, 2007). In some cases, DNA barcodes have been integral in differentiating between morphologically indistinguishable species, confirmed through additional nuclear DNA sequencing (Janzen et al., 2017). DNA has proven to be a powerful method, yet it does not allow for the development of morphologically based identifications for any future work, and depends on DNA-based approaches, which can be expensive.

Image-only analysis has shown promise in real-time insect species monitoring, but commonly has suffered when image background extraction is necessary. Furthermore, in the application of these image-only methods, only described species were monitored; two examples are pest management (Van Horn et al., 2018; Wu et al., 2019) and biodiversity surveys (Schneider et al., 2022). When deep learning methods were used with images to identify described classes of insects, accuracy gains reached 90% or greater (Milošević et al., 2020; Raitoharju & Meissner, 2019; Valan et al., 2019; Visalli et al., 2021), and in some cases, approached or surpassed taxonomist accuracies (He et al., 2015). However, these methods were tested either on coarse-grained datasets (easier to distinguish between classes) or with a limited number of species (generally <15).

Furthermore, the lingering issue of identifying rare or undescribed species and the inherent data imbalance continue to plague the ability of more efficient means of identifying new species. This is especially true within the Insecta class, where the majority of the species continue to be unidentified and would present a significant advancement to the field of entomology if identified. More broadly, identifying undescribed species helps us to better understand ecosystems and their processes, of which insects likely play a significant role (Yang & Gratton, 2014).

4.1 | ResNet model complexity

We experimented with extracting image features from different ResNet architectures with increasing complexity (ResNet-34, 50, 101 and 152), to evaluate if a more complex model would be optimal (see Table 4). We found that ResNet-34 produced a lower harmonic mean (*H*) (*H*=0.342) than the deeper ResNet-50, ResNet-101 and ResNet-152 architectures. ResNet-50 (*H*=0.376), and ResNet-101 (*H*=0.375) were on par. ResNet-152 was slightly better (*H*=0.391); however, we did not think this slight improvement justifies the use of the more complex ResNet-152 (~60M parameters) model over ResNet-101 (~40M parameters). ResNet-50 might have been preferred, but we stuck with ResNet-101 to be consistent with the architecture used in other related work.

TABLE 4 Harmonic means evaluated with different ResNet architectures, from least complex (ResNet-34) to most complex (ResNet-152).

Model	Harmonic mean
ResNet-34	0.342
ResNet-50	0.376
ResNet-101	0.375
ResNet-152	0.391

4.2 | HBM-DIT improvement over HBM-DNA

The hierarchical Bayesian model trained on DNA data alone (HBM-DNA) achieved a compelling 96.15% described species accuracy, where 670 out of 770 described species classes had entire test sample sets correctly classified to their true species. For the more challenging task of identifying undescribed species and assigning them to their true genera, this model's performance dropped to 73.39%. HBM-DIT, which leveraged auxiliary image data in addition to DNA data as input to the HBM, significantly boosted the undescribed performance to 81.39% (an 11% increase) with only a minimal increase in described accuracy. HBM-DIT classified 677 out of 770 described species classes with 100% accuracy. Further, there were 11 different genera for which HBM-DNA was not able to accurately predict any test samples from undescribed species belonging to these genera, as is shown in Figure 6, while HBM-DIT was able to predict seven out of these same 11 genera with an average undescribed accuracy of 87%, a marked improvement.

4.3 | Striking morphological similarity between species belonging to the same genus

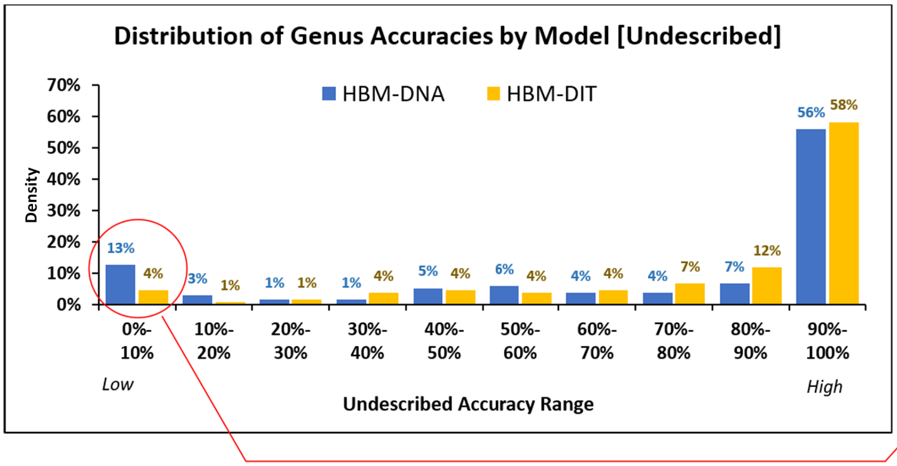
Physical variation in some insects is nearly invisible to the human eye, especially if one is not a specialized expert and species are closely related. Nevertheless, machine learning models can extract these subtle differences from images and when combined with DNA data, can classify these difficult cases correctly. To illustrate, we present a simple challenge in Figure 7 where one sample from four different species belonging to genus *Agabus* is displayed. The HBM-DNA model correctly classified all test samples from three described species classes, but did not perfectly assign samples from undescribed species to the true genus, *Agabus*. The HBM-DIT model, on the other hand, correctly classified all described and undescribed test samples. This observation revealed that 658bp DNA sequences (*cytochrome oxidase subunit I—COI*) alone lacked the level of differentiation needed to distinguish these species but was be more successful when combined with image representations that filled in the information gaps.

4.4 | Effect of background noise on model performance

High-quality images are an integral part of any successful machine learning approach and heavily impact the model performance

TABLE 5 Described and undescribed accuracy by insect family for five or more species per family. The reported results are from our transductive model (HBM-DIT). Note that the 'Summary' row reports the summary results from all families belonging to that order, including families having less than five species in our dataset.

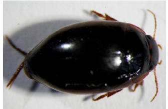
		Described species			Undescribed species	
Order	Family	# training	# test samples	Accuracy	# test samples	Accuracy
Coleoptera	Brentidae	94	18	100.00%		
	Cantharindae	226	43	93.02%	77	94.81%
	Carabidae	1660	346	95.66%	128	95.31%
	Cerambycinae	210	43	100.00%		
	Chrysomelidae	564	114	99.12%	37	89.19%
	Coccinellidae	226	46	100.00%		
	Curculionidae	348	68	94.12%	55	96.36%
	Dytisicidae	146	30	100.00%	18	88.89%
	Elateridae	242	47	100.00%	12	100.00%
	Scarabaeidae	106	23	91.30%		
	Staphylinidae	714	150	92.67%	47	100.00%
	Tenebrionidae	186	24	100.00%		
Summary (C)	37	5680	1143	95.80%	751	85.22%
Diptera	Calliphoridae	190	35	100.00%	13	92.31%
	Chironomidae	464	96	97.92%	24	100.00%
	Culicidae	496	107	89.72%	58	22.41%
	Drosophilidae	392	85	84.71%	80	81.25%
	Muscidae	104	22	90.91%		
	Sciaridae	150	33	100.00%		
	Syrphidae	342	71	97.18%	45	60.00%
	Tipulidae	122	26	96.15%	14	0.00%
Summary (D)	20	2744	570	93.68%	273	61.17%
Hymenoptera	Andrenidae	192	39	100.00%	53	79.25%
	Colletidae	190	32	100.00%	56	100.00%
	Crabronidae	312	66	100.00%	60	96.67%
	Eulophidae	226	47	100.00%	183	100.00%
	Halictidae	344	70	98.57%	113	80.53%
	Ichneumonidae	306	67	100.00%	12	100.00%
	Megachilidae	296	55	100.00%	28	53.57%
	Tenthredinidae	864	169	91.72%	261	66.28%
	Vespidae	106	22	100.00%	22	77.27%
Summary (H)	19	3282	660	97.27%	872	82.22%
Lepidoptera	Coleophoridae	994	206	99.51%	170	82.35%
	Crambidae	1054	176	99.43%	482	87.14%
	Depressariidae	1836	269	100.00%	380	67.63%
	Erebidae	4288	464	97.20%	694	74.78%
	Gelechidae	268	59	96.61%	41	82.93%
	Geometridae	1170	230	96.96%	328	89.63%
	Hesperiidae	2294	14	85.71%	566	47.00%
	Noctuidae	3246	570	98.95%	525	82.10%
	Notodontidae	4068	257	100.00%	959	94.89%
	Nymphalidae	554	37	100.00%	166	84.94%
	Saturniidae	890	31	100.00%	111	99.10%
	Tortricidae	968	170	100.00%	144	96.53%
Summary (L)	18	22,564	2592	98.61%	6567	81.01%



	Undescribed Accuracy	
Genus	**HBM-DNA**	**HBM-DIT**
Pristiphora [13]	0%	0%
Stenoma [100]	0%	0%
Tipula [14]	0%	0%
Eutomostethus [15]	0%	7%
Cryptocephalus [11]	0%	64%
Dichromodes [24]	0%	83%
Polydrusus [13]	0%	85%
Agriophara [35]	0%	89%
Ilybius [18]	0%	89%
Macaria [46]	0%	98%
Spilosoma [21]	0%	100%

Notation: [# test samples]

FIGURE 6 HBM-DIT improvement over HBM-DNA undescribed species accuracy. HBM-DNA did not accurately predict the genera for any undescribed species samples from the 11 genera listed in the table, while HBM-DIT, which combines both DNA and image features, was able to accurately predict seven out of the 11 with an average accuracy of 87%.

	Undescribed	*Described*		
	 Agabus *infuscatus* [0;15]	 **Agabus *sturmii* [9;4]**	 **Agabus *bipustulatus* [13;6]**	 **Agabus *uliginosus* [8;3]**
HBM-DNA	27% [4]	100%	100%	100%
HBM-DIT	100% [15]	100%	100%	100%
	Genus classif. acc.	Species classification accuracy		

Notation: [# training samples; # test samples] and [# correct]

FIGURE 7 Striking morphological similarity between four species from genus Agabus. The figure shows that deep learning is able to extract very subtle discriminative features from images, and when combined with DNA features, can improve performance over DNA-only models for the task of identifying samples from undescribed species. To underline the difficulty in classifying images from such similar species, imagine you are given just these four images (not labelled). You are told three images are from described species, which you are given the names of, and one is from is from an undescribed/unknown species. The figure above shows which species was randomly assigned as undescribed in our model. You look up known images from each of the described species to help you. How successful are you in assigning the described species images to the correct species and identifying which sample is from an undescribed species?

in computer vision tasks. It is well documented that due to the cross-entropy loss they have been trained with, many state-of-the-art pretrained CNN models were sensitive to the presence of subtle noise such as Gaussian, background noises, or blurriness in the image (Elsayed et al., 2018; Goodfellow et al., 2015; Khosla et al., 2020). Further, background information can sometimes dominate the relevant image features, as occurred for the only misclassified test sample from species *Bembidion minimum* (Figure 8), which also happened to be the only *Bembidion minimum* test sample with a "1 mm" scale line. This test sample was misclassified into genus *Drosophila* (from a different order entirely); samples from this genus also had a "1 mm" scale line, indicating that the model saw the scale line as the most discriminative feature in the test image, resulting in misclassification.

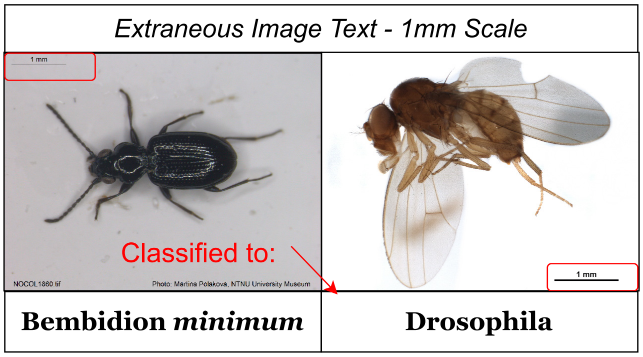

Extraneous Image Text - 1mm Scale
1 mm
Classified to:
NOCOL1860.tif
Photo: Martina Polakova, NTNU University Museum
1 mm
Bembidion *minimum*
Drosophila

FIGURE 8 Misclassification due to irrelevant background information (1 mm scale line) shared by images across different species.

4.5 | Effect of sample size on model performance

We ran additional experiments to investigate the number of samples per species on the overall performance. Our results suggest that our study could use as few as two samples per species and would still achieve a performance comparable to the classifier that uses all available samples from each species. More details are available in Supporting Information.

4.6 | Limitations

Immature specimens outside the scope of this work (eggs, larvae, pupae), especially for the holometabolous insects, will challenge the model, given drastic morphological differences across these life stages. However, the model could be trained to detect the stages hierarchically making the identification using computer vision possible. Also, our model assumed that the undescribed species belong to an existing genus, which may not be the case. A similar hierarchical Bayesian model, bumped up a taxonomic level from our model, can be implemented to consider undescribed genera under families for discovery and placement in the taxonomic hierarchy.

Cryptic species will undoubtedly be detected using this combined image and DNA model, thus leading to new biological discoveries. However, using image-only models with such a large number of species is not yet a viable classification method, despite its demonstrated robustness for differentiating cryptic species when there are only a small number of image classes. Conversely, phenotypically plastic species may contain different morphological characters; in theory, computer vision should detect these differences given enough training data. Future models would then reflect the detected plasticity with morphological keys, as demonstrated in plants (Divilov et al., 2017; Junior et al., 2013) and mosquitoes (Zacarés et al., 2018).

5 | CONCLUSIONS

This study developed a novel framework to facilitate discovery and identification of insect species, with much unknown biodiversity, at scale. The proposed model is the first in the literature to tackle this problem by leveraging image and DNA data together, is the first to be tested on more than a thousand species and is the first to have the ability to also classify undescribed species to genus. Our best performing hierarchical Bayesian classification model, trained with image and DNA feature data obtained from their respective deep learning models which were merged using the transductive linear mapping approach, classified described species with greater than 96% accuracy, and was 81% accurate in identifying the correct genus of undescribed species. Considering the transductive approach was built on a regularized linear mapping, there is great potential for achieving better performance by utilizing nonlinear mappings and/or more sophisticated approaches like generative adversarial networks (GAN) (Goodfellow et al., 2014) or variational autoencoders (VAE) (Kingma & Welling, 2014). Integrating GAN/VAE would create an end-to-end representation learning method that could potentially mitigate the shortcomings of supervised pretrained models such as ResNet-101. The HBM could also be extended to consider genera/species as subclasses and higher taxonomic levels, such as family, as superclasses. Such a classifier would readily deal with missing or unobserved genera. Given the large inter-species variation in DNA barcodes, a deep learning CNN model with a hierarchical loss function that considers information not just from species but genus, family and order could produce more robust DNA features, given a significantly larger dataset covering more genera and families. We are currently investigating what impact image resolution has on insect image classification accuracy, with the hypothesis that higher resolution images could yield stronger model performance. We are also investigating using images in conjunction with the dataset of morphological characters present in each image to see whether machine learning and deep learning models can learn to identify and name fine-grained characteristics in new image samples, which do not have accompanying morphological data as a starting point for translating deep features to semantic descriptors. Machine and deep learning methods are ever-evolving, and the hope is that the presentation of this methodology will highlight the promise of these methods for tackling insect identification and beyond.

AUTHOR CONTRIBUTIONS

Christine Johanna Picard, George Mohler and Murat Dundar conceived the original idea. Sarkhan Badirli, Zeynep Akata and Murat Dundar developed the theory. Sarkhan Badirli and Murat Dundar designed and conducted all experiments. Murat Dundar (machine learning) and Christine Johanna Picard (entomology) supervised the project. Sarkhan Badirli and Christine Johanna Picard analysed the results. Sarkhan Badirli, Murat Dundar, Frannie Richert and Christine Johanna Picard wrote the manuscript with input from all authors. All authors reviewed the manuscript.

ACKNOWLEDGEMENTS

We acknowledge the valuable reviews from our anonymous reviewers for significantly improving the final version of this manuscript. Murat Dundar and Sarkhan Badirli were sponsored by the National Science Foundation (NSF) grant IIS-1252648 (CAREER). George Mohler was sponsored by NSF grant ATD-2124313. Frannie Richert was supported in part from funding from the IUPUI Institute of Artificial Intelligence. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF.

CONFLICT OF INTEREST STATEMENT

None to report.

DATA AVAILABILITY STATEMENT

Data used in this study are available at IUPUI DataWorks https://doi.org/10.7912/D2/27 (Badirli, Picard, et al., 2021).

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14104.

ORCID

Christine Johanna Picard https://orcid.org/0000-0001-8589-8283
George Mohler https://orcid.org/0000-0003-4293-5106
Zeynep Akata https://orcid.org/0000-0002-1432-7747
Murat Dundar https://orcid.org/0000-0001-5752-468X

ENDNOTE

[1] Samples from undescribed species cannot be assigned to their respective species of origin because their species is not known and thus not represented during training.

REFERENCES

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bourexis, D., Brister, J. R., Bryant, S. H., Canese, K., Cavanaugh, M., Charowhas, C., Clark, K., Dondoshansky, I., Feolo, M., Fitzpatrick, L., Funk, K., Geer, L. Y., Gorelenkov, V., … Kimura, M. (2018). Database resources of the National Center for biotechnology information. *Nucleic Acids Research*, *46*, D8–D13.

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (2021). *BLASTn (from NCBI)*. https://blast.ncbi.nlm.nih.gov/Blast.cgi

Badirli, S., Akata, Z., & Dundar, M. (2020). Bayesian zero-shot learning. In *European conference on computer vision Workshops* (pp. 687–703). Springer.

Badirli, S., Akata, Z., Mohler, G., Picard, C. J., & Dundar, M. (2021). Fine-grained zero-shot learning with DNA as side information. In *NeurIPS*. Curran Associates, Inc.

Badirli, S., Picard, C. J., Mohler, G., Akata, Z., & Dundar, M. (2021). Data from Classifying the unknown: Insect identification by deep zero-shot Bayesian learning. https://doi.org/10.7912/D2/27

Bendale, A., & Boult, T. E. (2016). Towards open set deep networks. In *CVPR*. IEEE.

Bora, K., Chowdhury, M., Mahanta, L., Kundu, M. K., & Das, A. K. (2016). Pap smear image classification using convolutional neural network. In *Tenth Indian conference on computer vision, graphics and image Processing*. ACM.

Buck, M., Woodley, N. E., Borkent, A., Thomas Pape, J., Vockeroth, R., & Marshall, S. A. (2009). Key to Diptera families-adults. *Manual of Central American Diptera*, *1*, 95–156.

Burns, J. M., Janzen, D. H., Hajibabaei, M., Hallwachs, W., & Hebert, P. D. (2008). DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in area de Conservacion Guanacaste, Costa Rica. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6350–6355.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9650–9660). IEEE.

Chan, T., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). Pcanet: A simple deep learning baseline for image classification. *IEEE Transactions on Image Processing*, *24*, 5017–5032.

Costello, M. J., May, R. M., & Stork, N. E. (2013). Can we name Earth's species before they go extinct? *Science*, *339*(6118), 413–416.

Ding, W., & Taylor, G. (2016). Automatic moth detection from trap images for pest management. *Computers and Electronics in Agriculture*, *123*, 17–28.

Divilov, K., Wiesner-Hanks, T., Barba, P., Cadle-Davidson, L., & Reisch, B. I. (2017). Computer vision for high-throughput quantitative phenotyping: A case study of grapevine downy mildew sporulation and leaf trichomes. *Phytopathology*, *107*(12), 1549–1555.

Elsayed, G., Krishnan, D., Mobahi, H., Regan, K., & Bengio, S. (2018). Large margin deep networks for classification. In *NeurIPS*. Curran Associates, Inc.

Geng, C., Huang, S. J., & Chen, S. (2020). Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3614–3631.

Goodfellow, I., Abadie, J. P., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS*. Curran Associates, Inc.

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR*. OpenReview.net.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*. IEEE.

Hebert, P. D., Cywinska, A., Ball, S. L., & Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *270*, 313–321.

Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Grassa, C. J., Park, D. S., Kennedy, J., Clarke, J. A., Cook, J. A., Blackburn, D. C., Edwards, S. V., & Davis, C. C. (2020). Digitization and the future of natural history collections. *BioScience*, *70*, 243–251. https://doi.org/10.1093/biosci/biz163

Hopkins, G., & Freckleton, R. P. (2002). Declines in the numbers of amateur and professional taxonomists: Implications for conservation. *Animal Conservation*, *5*(3), 245–249.

Høye, T. T., Ärje, J., Bjerge, K., Hansen, O. L. P., Iosifidis, A., Leese, F., Mann, H. M. R., Meissner, K., Melvad, C., & Raitoharju, J. (2021). Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(2), 118.

Janzen, D. H., Burns, J. M., Cong, Q., Hallwachs, W., Dapkey, T., Manjunath, R., Hajibabaei, M., Hebert, P. D. N., & Grishin, N. V. (2017). Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(31), 8313–8318.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian protein metabolism* (Vol. *3*, pp. 21–132). Academic Press. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7

Junior, J. J. M. S., Rossatto, D. R., Kolb, R. M., & Bruno, O. M. (2013). A computer vision approach to quantify leaf anatomical plasticity: A case study on *Gochnatia polymorpha* (Less.) Cabrera. *Ecological Informatics*, *15*, 34–43.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. In *NeurIPS*. Curran Associates, Inc.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR*. Arxiv.org.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Lee, M. S. (2000). A worrying systematic decline. *Trends in Ecology & Evolution*, *15*(8), 346.

Mayo, M., & Watson, A. T. (2007). Automatic species identification of live moths. *Knowledge-Based Systems*, *20*, 195–202.

Meineke, E. K., Tomasi, C., Yuan, S., & Pryer, K. M. (2020). Applying machine learning to investigate long-term insect-plant interactions preserved on digitized herbarium specimens. *Applications in Plant Sciences*, *8*(6), e11369.

Milošević, D., Milosavljevic, A., Predic, B., Medeiros, A. S., Savić-Zdravković, D., Piperac, M. S., Kostić, T., Spasić, F., & Leese, F. (2020). Application of deep learning in aquatic bioassessment: Towards automated identification of non-biting midges. *The Science of the Total Environment*, *711*, 135160. https://doi.org/10.1016/j.scitotenv.2019.135160

Nanni, L., Ghidoni, S., & Brahnam, S. (2017). Handcrafted vs non-handcrafted features for computer vision classification. *Pattern Recognition*, *71*, 158–172.

Orr, M. C., Ascher, J. S., Bai, M., Chesters, D., & Zhu, C. D. (2020). Three questions: How can taxonomists survive and thrive worldwide? *Megataxa*, *1*(1), 19–27.

Pegoraro, L., Hidalgo, O., Leitch, I. J., Pellicer, J., & Barlow, S. E. (2020). Automated video monitoring of insect pollinators in the field. *Emerging Topics in Life Sciences*, *4*, 87–97.

Perera, P., & Patel, V. M. (2019). Deep transfer learning for multiple class novelty detection. In *CVPR*. IEEE.

Raitoharju, J., & Meissner, K. (2019). On confidences and their use in (semi-)automatic multi-image taxa identification. In *IEEE symposium series on computational intelligence (SSCI)*. IEEE.

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The barcode of life data system (http://www.Barcodinglife.Org). *Molecular Ecology Notes*, *7*(3), 355–364.

Ratnasingham, S., & Hebert, P. D. (2013). A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLoS ONE*, *8*(7), e66213.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *IJCV*, *115*(3), 211–252.

Scheirer, W. J., & Boult, T. E. (2016). Statistical methods for open set recognition. In *CVPR Tutorial*. IEEE.

Schneider, S., Taylor, G. W., Kremer, S. C., Burgess, P., McGroarty, J., Mitsui, K., Zhuang, A., deWaard, J. R., & Fryxell, J. M. (2022). Bulk arthropod abundance, biomass and diversity estimation using deep learning for computer vision. *Methods in Ecology and Evolution*, *13*(2), 346–357.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv Preprint arXiv:1312.6034*.

Stork, N. E. (2018). How many species of insects and other terrestrial arthropods are there on earth? *Annual Review of Entomology*, *63*, 32–45.

Sun, Y., Liu, X., Yuan, M., Ren, L., Wang, J., & Chen, Z. (2018). Automatic in-trap pest detection using deep learning for pheromone-based *Dendroctonus valens* monitoring. *Biosystems Engineering*, *176*, 140–150.

Tran, D. T., Høye, T. T., Gabbouj, M., & Iosifidis, A. (2018). Automatic flower and visitor detection system. In *European signal processing conference (Eusipco)*. IEEE.

Valan, M., Makónyi, K., Maki, A., Vondrácek, D., & Ronquist, F. (2019). Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Systematic Biology*, *68*, 876–895.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8769–8778). IEEE.

Visalli, F., Bonacci, T., & Borghese, N. A. (2021). Insects image classification through deep convolutional neural networks. In *Progresses in Artificial Intelligence and Neural Systems* (pp. 217–228). Springer.

Wang, J., Lin, C., Ji, L., & Liang, A. (2012). Automatic species identification of live moths. *Knowledge-Based Systems*, *33*, 102–110.

Wu, X., Zhan, C., Lai, Y. K., Cheng, M. M., & Yang, J. (2019). Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8787–8796). IEEE.

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(9), 2251–2265.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, *59*(8), 3370–3388.

Yang, L. H., & Gratton, C. (2014). Insects as drivers of ecosystem processes. *Current Opinion in Insect Science*, *2*, 26–32.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *arXiv Preprint arXiv:1411.1792*.

Zacarés, M., Salvador-Herranz, G., Almenar, D., Tur, C., Argilés, R., Bourtzis, K., Bossin, H., & Pla, I. (2018). Exploring the potential of computer vision analysis of pupae size dimorphism for adaptive sex sorting systems of various vector mosquito species. *Parasites & Vectors*, *11*(2), 151–164.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Data S1: Supplementary Material.

How to cite this article: Badirli, S., Picard, C. J., Mohler, G., Richert, F., Akata, Z., & Dundar, M. (2023). Classifying the unknown: Insect identification with deep hierarchical Bayesian learning. *Methods in Ecology and Evolution*, *00*, 1–16. https://doi.org/10.1111/2041-210X.14104