Artificial intelligence versus natural selection: using computer vision techniques to classify bees and bee mimics

Tanvir Bhuiyan[a], Ryan M. Carney[b], Sriram Chellappan[a]

{*bhuiyan,ryancarney,sriramc*}*@usf.edu*

[a]*Computer Science and Engineering, University of South Florida*
[b]*Integrative Biology, University of South Florida*

Abstract

Bees play a very positive role in nature, and bumble bees in particular are critical pollinators. Unfortunately, bee numbers are dramatically decreasing worldwide. A popular platform geared towards piquing public interest in bees and other organisms is iNaturalist, where citizen scientists upload their observations. In this paper, we design computer vision techniques (deep neural networks) for two problems related to bees that leverage such crowd-sourced imagery. When presented with an image of an insect, our proposed techniques are geared towards a) classifying whether the image contains a bee or not, and if yes, b) whether or not it is that of a bumble bee. Our neural network models were trained, validated, and tested on a dataset of 6,332 images on color separately, grayscale separately, and both combined. For the first classification problem (detecting bees from non-bees), our model trained with grayscale images achieved an accuracy of 91.71%. For the second problem (detecting bumble bees from non-bumble bees), our model trained on a combination of color and grayscale images achieved an accuracy of 88.86%. Using state-of-the-art explainable AI methods such as class activation maps (CAMs), we also validated whether or not our models learn from appropriate components within the image, which in turn provided anatomical insights. Additionally, a number of insects have independently evolved mimicry of bees and wasps. Therefore, we tested our techniques as a proxy for a natural predator, by evaluating bee mimics and non-bee mimics across twelve disparate families and three orders. Results reveal that classification accuracy is inversely related to the level of aggressive mimicry. For example, the bee killers (Asilidae: *Mallophora* spp.) exhibited the best mimicry in both models, yielding the lowest accuracies of 18.33% and 47.22%. Furthermore, the between-group clustering of the t-SNE results (a technique to visualize high-dimensional data) replicates the phylogeny of these convergently evolved bee mimics, and with perfect within-group clustering. Ultimately, we believe that techniques in this paper can enhance global citizen science efforts across multiple domains. Our methodologies also enable novel approaches to better investigate complex problems related to the mimicry and morphology of bees and insects in general, through the transdisciplinary synthesis of artificial intelligence and natural selection paradigms.

Keywords: artificial intelligence, bee, bumble bee, citizen science, computer vision, deep learning, natural selection, insects, machine learning, mimic

Introduction

Bees have played a critical role as pollinators ever since the Cretaceous [1]. As a source of honey, widespread exploitation of honey bees dates back to at least the early Neolithic farmers (∼9 kya, [2]), and the earliest known apiculture was practiced by the Ancient Egyptians [3]. Darwin's experiments with "humble bees", as bumble bees were once called, exemplifies how these insects have furthered our understanding of natural selection with respect to co-evolution, ecosystem webs, and social behavior [4]. Such a rich history makes it all the more tragic that today, despite their ecological and economic importance, bees are facing an unprecedented anthropogenic decline in diversity and abundance [5] – making their identification and conservation urgent concerns.

The mimicry of bees also has a storied history, albeit inadvertently so. As recounted by poets from Virgil to Shakespeare, the ancient superstitious ritual of *bugonia* held that bees spontaneously generated from the decaying carcasses of animals such as oxen [6]. The primary culprit of this deception was likely the honey bee-mimicking *Eristalis tenax* (Syrphidae), a cosmopolitan hoverfly that lays its eggs upon carrion (Ibid.). Its common name, the common drone fly, stems from the resemblance of members of this genus to the drones of honey bees (Figure 1).

Indeed, Hymenoptera, consisting of bees, wasps, and ants, is the most mimicked insect order [7], which is not surprising given their formidable sting. A resemblance to such well-equipped insect models protects harmless

mimics, by fooling would-be predators with a false harmful signal. This type of defensive mimicry is known as "Batesian mimicry" [8].

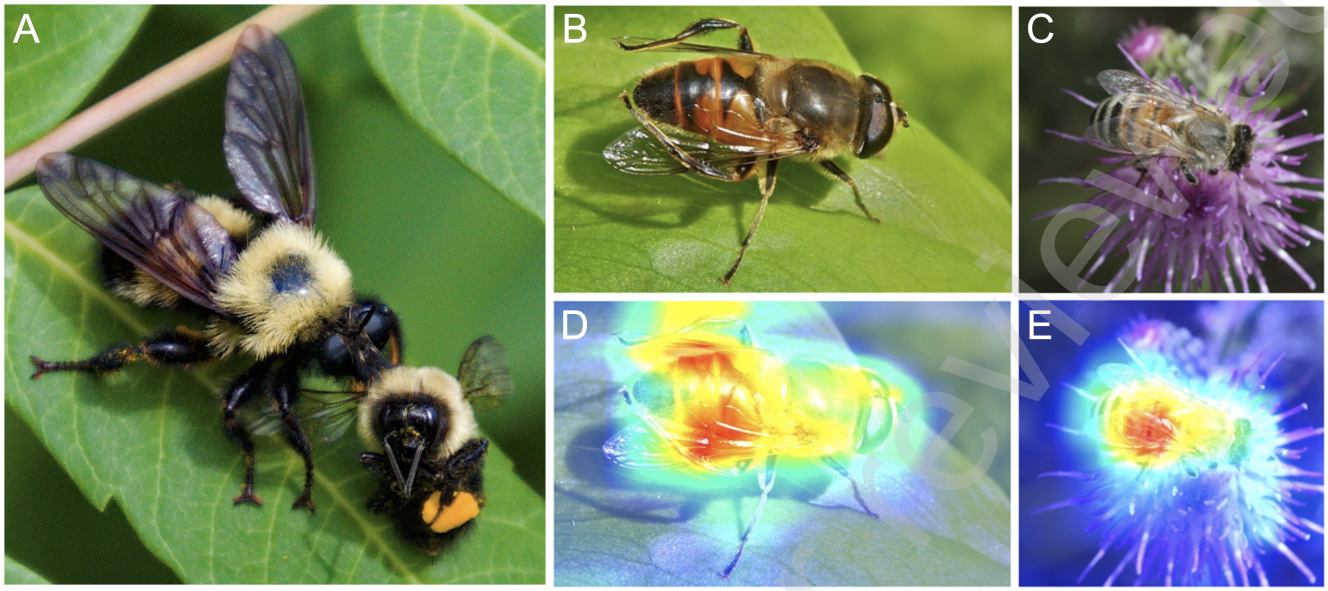

Figure 1: Citizen scientist photos of bees and bee mimic flies. A. Bumble bee mimic (Asilidae: *Laphria thoracica*) preying on a bumble bee (Hymenoptera: *Bombus* spp.), an example of aggressive mimicry. Honey bee mimic (B, Syrphidae: *Eristalis tenax*) and honey bee (C, Hymenoptera: *Apis mellifera*), an example of defensive mimicry. Red areas in the class activation maps (D,E) denote the importance of the wings and abdominal markings in these two correct classifications (mimic as a non-bee, honey bee as a bee) by our AI algorithms, elaborated subsequently in this paper. Original images from iNaturalist.

A less common, and opposite – although not mutually exclusive – phenomenon, is aggressive mimicry (Figure 1). This "wolf in sheep's clothing" strategy evolved to fool prey into thinking that the mimic is harmless. A classic example of this is the anglerfish. However, the term aggressive mimicry was actually first ascribed to bumble bee mimics, also from the hoverfly family Syrphidae (genus *Volucella*), by Poulton in 1890 [7] following similar observations of this genus by Kirby & Spence 1817 [9] and Wallace 1871 [10] (see also [11]). Poulton noted: "*In some cases the Mimicry enables the aggressive form to lay eggs in the nest of that which it resembles, so that its larvæ live upon the food stored up by the latter or even upon the larvæ themselves. The boldness of these enemies sometimes depends upon the perfection of their disguise.*"

Our motivation for this study is two-fold: first, to evaluate the performance of machine learning techniques in accurately identifying convergently evolved bee mimics. In other words, does the visual resemblance to bees gained through natural selection fool artificial intelligence algorithms that were trained on bees? Our second motivation is geared towards conservation of bumble bees, especially among citizen scientists. In particular, the iNaturalist [12] platform is a crowning example of a global citizen science effort towards conservation. It was launched in 2008, and as of January 2022, citizens have contributed more than 89 million observations of animals, plants, and other organisms worldwide. At the time of writing this paper, the platform has $1,206,783$ image observations of bees from $192,822$ observers. For each observation contributed, multiple classifications are made (since any non-expert can also classify) in iNaturalist. Among these, "Research Grade" classifications are most reliable, for which two conditions must be met. The observation a) must contain a valid date, location, photo or sound, and not be of a captive/cultivated organism; and b) at least two knowledgeable experts in the field must agree on the identification for that observation [13]. For bees, we found that among the $1,206,783$ observations, $669,155$ of them are Research Grade (i.e., around 55%), which means that close to half among uploaded observations are not reliably identified.

It is easy to infer that in order to enhance the scale, utility and efficiency of citizen science efforts, and also to pique interest among citizen scientists, automated, accurate and rapid classification of their observations can go a long way. With this motivation, in this paper, we present AI (computer vision) techniques to classify bee and bee mimics using images contributed by citizen scientists in the iNaturalist platform. Our specific contributions are:

- **An AI model for classifying bees from other insects.** We design a model based on VGG16 [14] to classify bees from insects. This model was trained, validated and tested on $3,029$ bee and $2,943$ non-bee insect images.

- **An AI model for classifying bumble bees from other bees.** We design a model based on ResNet-101 [15] to classify bumble bees from non bumble bees. This model was trained, validated and tested on $1,554$ bumble bee and $1,475$ non-bumble bee images.

- **Evaluating both AI models against independently evolved bee mimics:** Our taxa comprise 19 bee mimic species across six diverse insect families (Asilidae, Bombyliidae, Scarabaeidae, Sphingidae, Syrphidae, Tachinidae) in three orders – Coleoptera (beetles), Diptera (flies), and Lepidoptera (moths) – as well as related outgroups of nine species of wasp mimics and 13 species of non-mimics to serve as controls (see Supplemental Information). We hypothesize that 1) within each clade, bee mimics will exhibit lower model classification accuracy compared to their non-mimic counterparts, and 2) wasp mimics will exhibit intermediate accuracy; furthermore, 3) bee mimics that also exploit aggressive mimicry toward bees will exhibit better mimicry – as defined by lower AI classification accuracy – compared to bee mimics that employ only defensive mimicry. Finally, we visualize such classifications of artificial intelligence vis-a-vis natural selection through a novel integration of t-Distributed Stochastic Neighbor Embedding technique [1] (phenotype) and evolutionary relationships (phylogeny).

- **AI-driven insights on the fidelity of our techniques.** To evaluate the fidelity and explainability of our AI models, as well as to explore the role of color in aposematic mimicry, we adopt the technique of Class Activation Maps (CAM) [17] to pinpoint which pixels in an image are most used to classify (Figure 1D,E). We also conduct studies to convert images to grayscale, and evaluate the performance of the grays-scale dataset, and make observations on the roles of color and texture within an image that AI models leverage for insects classification.

Results

The metric we used to evaluate our models is classification accuracy, defined as the percentage of correct predictions out of the total number of specimens.

$$accuracy = \frac{100 \times (True\ positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)}. \quad (1)$$

We have used $6,332$ distinct images in total to train, validate and test our AI models. These are presented in Tables 1, 2, 3 and 4. Table 1 lists all orders of insects and corresponding counts of images used in this paper, starting with Hymenoptera (encompassing bees). Table 2 lists species of bumble bees and counts of images used. Table 3 lists the genera of non bumble bees used, and finally, Table 4 lists the family names of all mimics used along with corresponding image counts.

a. Classifying bees from other insects: For this problem, we chose images presented in Tables 1, 2 and 3 for training and validation. Broadly, 80% of images in each row in the Tables was used for training and validation, and the other 20% is un-seen and purely used for testing. Note that, images of bumble bee species in the last five rows in Table 2 were used only for testing, and not for training or validation.

Results of our VGG-based AI algorithm [14] [2] for the classification of bees from other insects are presented in Table 5. All testing results reported are for un-seen images only. We trained three AI models separately trained on color only, grayscale only, and an equal combination of color and grayscale images. Each model was tested separately on an equal number of color and grayscale images and results are presented in the appropriate column in Table 5. Boldface percentages represent top two classification accuracies for color and gray images among the three models. Row 3 are the accuracies (in percentage) for classifying bees from insects. For this, a total of 683 bee and 539 non-bee (insect) images equally split between both classes across all species, and between color and grayscale was used. The following rows denote accuracies for the same AI model in classifying mimics in Table 4. For testing mimics 30 color and 30 corresponding grayscale images for each of the 12 groups were used. Note that none of mimic images was used for training/ validation of the AI models. As such, the mimics are purely un-seen by the AI. We consider a classification as correct if the AI model identifies a mimic as a non-bee. Unfortunately though, the accuracy of classifying mimics is not very good with the VGG model. Furthermore, we see that grayscale image based models overall perform better that color based models for the classification of bees from insects.

[1] An AI technique for dimensionality reduction, particularly well suited for the visualization of high-dimensional datasets [16].
[2] The AI algorithm details are elaborated in Methods section.

Insects Order	**Count**
Hymenoptera	330
Blattodea	329
Coleoptera	337
Diptera	328
Lepidoptera	630
Odonata	329
Orthoptera	660

Table 1: Insect orders and image counts

Bumble Bee Species	**Count**
affinis	247
griseocollis	250
impatiens	250
melanopygus	150
pascuorum	157
pensylvanicus	200
terrestris	250
bimaculatus	10
flavifrons	10
lucorum	10
terricola	10
vosnesenskii	10

Table 2: Bumble bee species and image counts. Images in last five rows are not used for training/ validation, but used only for testing

b. Classifying bumble bees from non-bumble bees: For this problem we employ a similar approach as above. The AI model used was ResNet-101 [15] [3]. For this problem also, we selected 80% of images from the first seven species from Table 2 and the first seven genera from Table 3 for model training and validation. The rest were used for testing. Here also, we trained three AI models separately trained on color only, grayscale only, and an equal combination of color and grayscale images. Each model was tested separately on an equal number of color and grayscale images and results are presented in the appropriate column in Table 6. All testing results are again for un-seen images only. We present two types of results here. The third row in Table 6 presents the accuracy of detecting bumble bees from non bumble bees in a dataset comprising of un-seen images from those species of bumble and non bumble bees used in training/ validation. The fourth row presents the accuracy of detecting bumble bees from non bumble bees in a dataset comprising of un-seen images from those species of bumble and non bumble bees that were not used at all in training/ validation. In other words, these species were also unseen by the AI model. The number of testing images of seen species used for testing 585 color and correspondingly 585 grayscale images distributed evenly among all species. The number of testing images of un-seen species used for testing was 98 color and correspondingly 98 grayscale images distributed evenly among all species. More details of species used in testing are in Supplementary Documentation.

The following rows in Table 6 denote accuracies for the same AI model in classifying mimics in Table 4. For testing mimics 30 color and 30 corresponding grayscale images were used. Note that here also, none of mimic images was used for training/ validation of the AI models. As such, the mimics are purely un-seen by the AI. We consider a classification as correct if the AI model identifies a mimic as a non bumble-bee. The ResNet-101 model does much better for classifying mimics as compared to the previous VGG model. Also, while the grayscale image based models perform better that color based overall for the problem of classifying bumble bees from non-bumble bees, both are overall similar in performance when it comes to classifying mimics

c. Evaluating model fidelity: In order to evaluate the fidelity and explainability of our AI models, we adapt the technique of Class Activation Maps (CAM) [17] to pin point which areas (pixels) in any image are most used to make a classification by the AI. If the pixels highlighted in the CAM appear on anatomical components of the insect, that means the model is learning to classify correctly, while ignoring the background. The warmer a pixel is in the CAM (i.e., redder), the higher the weight of that pixel used for classification. From all figures - 2, 3, 4, 5, and 6 for all our AI models and image classes, we see that our model focuses primarily on the anatomical components of the insect (for both color and grayscale images), and has learned well enough to ignore the background. Results are indeed generalizable, and increase confidence in our AI models. Furthermore, these techniques (i.e., highlighting pixels) can also better pique excitement among citizen scientists, when they see the power of AI algorithms, and explainability also.

d. Training, Hardware, and Inference time: Our training and validation was done on a GPU cluster of four Nvidia GeForce GTX TITAN X cards each having $3,072$ CUDA cores and 12 GB memory each [18]. It took around 28 hours to train and validate the VGG16 based bee vs nonbee model, and took 46 hours to train and

[3] The AI algorithm details are elaborated in Methods section.

Non-Bumble Bee Genus	**Count**
Andrena	179
Anthidium	180
Apis	518
Centris	239
Megachile	184
Melissodes	174
Osmia	1

Table 3: Non bumble bee species and image counts

Bee mimics and *non-bee mimics* Family	**Count**
Scarabaeidae (bee mimic)	30
Scarabaeidae (non-mimic)	30
Asilidae (bee mimic)	30
Bombyliidae (bee mimic)	30
Syrphidae (bee mimic)	30
Syrphidae (wasp mimic)	30
Tachinidae (bee mimic)	30
Tachinidae (wasp mimic)	30
Tachinidae (non-mimic)	30
Sphingidae (bee mimic)	30
Sesiidae (wasp mimic)	30
various (non-mimic)	30
Total	360

Table 4: Mimics and Image counts. Mimics are categorized into three orders: Coleoptera, Diptera and Lepidoptera. See Supplementary Information for species-level designations.

training image type	**color**		**gray**		**color+gray**		average
testing image type	color	gray	color	gray	color	gray	
bees vs. non-bees	84.63	79.79	**91.36**	**91.71**	90.85	91.62	88.33
testing accuracy							
Scarabaeidae (bee mimic)	23.33	43.33	**60**	**53.33**	43.33	43.33	44.44
Scarabaeidae (non-mimic)	76.67	80.00	83.33	83.33	**86.67**	**86.67**	82.78
Asilidae (bee mimic)	10.00	**36.67**	16.67	13.33	**20.00**	13.33	18.33
Bombyliidae (bee mimic)	**46.67**	**53.33**	20.00	16.67	33.33	26.67	32.78
Syrphidae (bee mimic)	**23.33**	**36.67**	**23.33**	16.67	13.33	13.33	21.11
Syrphidae (wasp mimic)	36.67	30.00	46.67	**36.67**	**50.00**	33.33	38.89
Tachinidae (bee mimic)	23.33	**30.00**	**30.00**	20.00	16.67	13.33	22.22
Tachinidae (wasp mimic)	60.00	60.00	**83.33**	**86.67**	63.33	70.00	71.67
Tachinidae (non-mimic)	66.67	60.00	**80.00**	**83.33**	60.00	63.33	68.89
Lepidoptera (bee mimic)	30.00	50.00	56.67	**60.00**	**70.00**	**60.00**	54.45
Lepidoptera (wasp mimic)	66.67	**83.33**	**73.33**	70.00	70.00	70.00	72.22
Lepidoptera (non-mimic)	86.67	90.00	96.67	**100.00**	**100.00**	**100.00**	95.56
average	45.83	**54.44**	**55.83**	53.33	52.22	49.44	51.95

Table 5: Comparison of accuracy (in %) between three VGG16-based models[5] that classify bees from non-bees, evaluated against testing images of bee mimics and non-bee mimics (italicized). A correct classification is when the model accurately identifies these insects as a non-bee. Boldface percentages represent the top classification accuracy for color and gray images among the three models for each row.

training image type	**color**		**gray**		**color+gray**		average
testing image type	color	gray	color	gray	color	gray	
species used in training, validation	73.48	77.82	82.33	81.99	**85.78**	**86.87**	81.38
species not used in training, validation	86.73	86.73	**94.90**	**93.88**	92.86	92.86	91.33
average for testing	77.90	80.79	86.52	85.95	**88.14**	**88.86**	86.36
testing accuracy							
Scarabaeidae (bee mimic)	**93.33**	**93.33**	76.67	60.00	**93.33**	90.00	84.44
Scarabaeidae (non-mimic)	**90.00**	**86.67**	80.00	**86.67**	80.00	83.33	84.45
Asilidae (bee mimic)	**66.67**	**56.67**	33.33	26.67	56.67	43.33	47.22
Bombyliidae (bee mimic)	**96.67**	66.67	56.67	63.33	90.00	**83.33**	76.11
Syrphidae (bee mimic)	**100.00**	96.67	96.67	96.67	**100.00**	**100.00**	98.34
Syrphidae (wasp mimic)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	100.00
Tachinidae (bee mimic)	**100.00**	80.00	70.00	66.67	93.33	**86.67**	82.78
Tachinidae (wasp mimic)	**100.00**	**100.00**	96.67	**100.00**	**100.00**	**100.00**	99.45
Tachinidae (non-mimic)	**100.00**	96.67	93.33	86.67	**100.00**	**100.00**	96.11
Lepidoptera (bee mimic)	**90.00**	70.00	70.00	66.67	86.67	**73.33**	76.11
Lepidoptera (wasp mimic)	96.67	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.45
Lepidoptera (non-mimic)	66.67	56.67	**90.00**	66.67	86.67	**70.00**	72.78
average	**91.67**	83.61	76.67	76.67	90.56	**85.83**	84.77

Table 6: Comparison of accuracy (in %) between three ResNet-101-based models[7] that classify bumble from non-bumble bees, evaluated against testing images of bee mimics and non-bee mimics (italicized). A correct classification is when the model accurately identifies these insects as a non-bumble bee. Boldface percentages represent the top classification accuracy for color and gray images among the three models for each row.

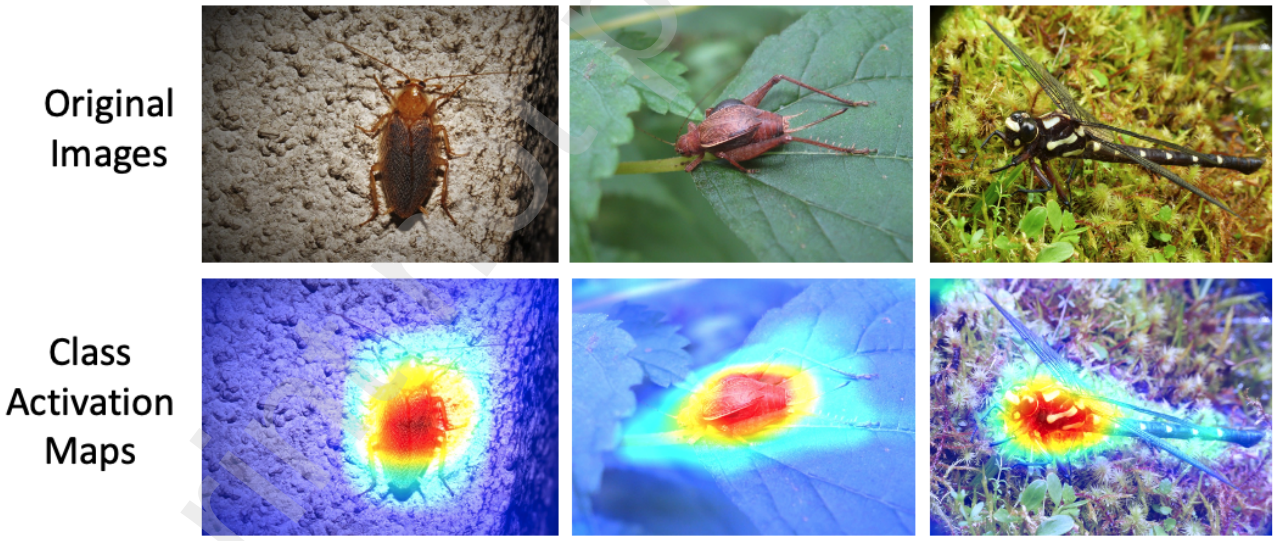

Figure 2: Class activation maps on non-bee images, using the bee vs. non-bee classifier (VGG16-based model, trained with color images). From left: *Ectobius vittiventris* (correct), *Hapithus agitator* (correct), *Uropetala carovei* (correct)).

validate the ResNet-101 based bumble bee vs non-bumble bee model. Inference time for a single image was less than a second for both models.

Figure 3: Class activation map on non-bumble bee images using bee-vs-nonbee classifier (VGG16 based model, trained with color-images). From left - *Andrena cineraria* (correct), *Apis dorsata* (correct), *Megachile mendica* (correct)).

Figure 4: Class activation map on bumble bee images using bee-vs-nonbee classifier (VGG16 based model, trained with color-images). From left - *Bombus impatiens* (correct), *Bombus impatiens* (correct), *Bombus pensylvanicus* (correct)).

Discussion

The ability to classify bees – and specifically bumble bees – from other insects has important applications for conservation and education. Primarily, the adoption of robust and rapid AI techniques will significantly enhance the scale and permeance of citizen science platforms. Immediate feedback to non-expert citizen scientists will go a long way to piquing interest among the general public. Our accuracies for detecting bees from insects is more than 90%, while the same for detecting bumble bees from non-bumble bees is more than 80% when trained with a combination of color and grayscale images. These numbers will only improve with more datasets and training. Furthermore, quick and accurate feedback to citizen scientists for critically endangered species (like for example the spotting of *Bombus affinis*) may have direct impact to sustaining efforts for the conservation. Deploying our algorithms in real-time is part of our on-going efforts.

For the mimics studied, the bee-vs-nonbee classifying (VGG16 based) model yielded lower classification accu-

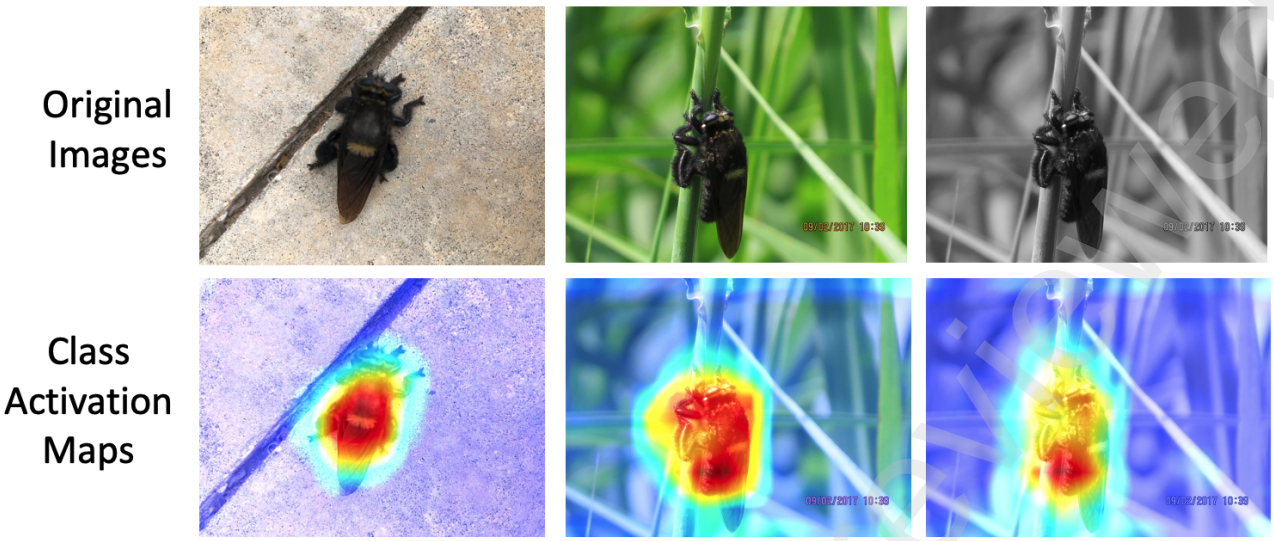

Figure 5: Class activation map on images from the family Asilidae using bee-vs-nonbee classifier (VGG16 based model, trained with color-images). From left - *Mallophora leschenaulti* (incorrect), *Mallophora leschenaulti* (incorrect), *Mallophora leschenaulti* (gray) (correct)).

Figure 6: Class activation map on images of Scarabaeidae bee mimics, using bee vs. non-bee classifier (VGG16-based model, trained with color images). From left: *Trichius fasciatus* (incorrect), *Trichius gallicus* (incorrect), *Trichius sexualis* (correct).

racies compared to those of the bumble bee-vs-non-bumble bee (ResNet-101 based) model (Tables 5 and 6). This difference between both models could be due to the different neural network architectures, and/or the phylogenetic specificity of both the bumble bee and non-bumble bee classes. Note that the ResNet-101 model is much heavier with much more layers and is better suited at learning finer grained discriminators compared to the lighter weight VGG models. Second, the majority of the bee mimics were bumble bee mimics (as opposed to honey bee or carpenter bee mimics). Second, the non-bumble bee class was trained using non-bumble bee bees, as opposed to the more general non-bee insects used to train the bee model.

Results supported our first hypothesis, as the bee mimics within all four relevant clades – Scarabaeidae, Syrphidae, Tachinidae, and Lepidoptera – had the lowest classification accuracy compared to their wasp mimic and non-bee mimic counterparts, across both bee and bumble bee models (Tables 5, 6). The only exception to this

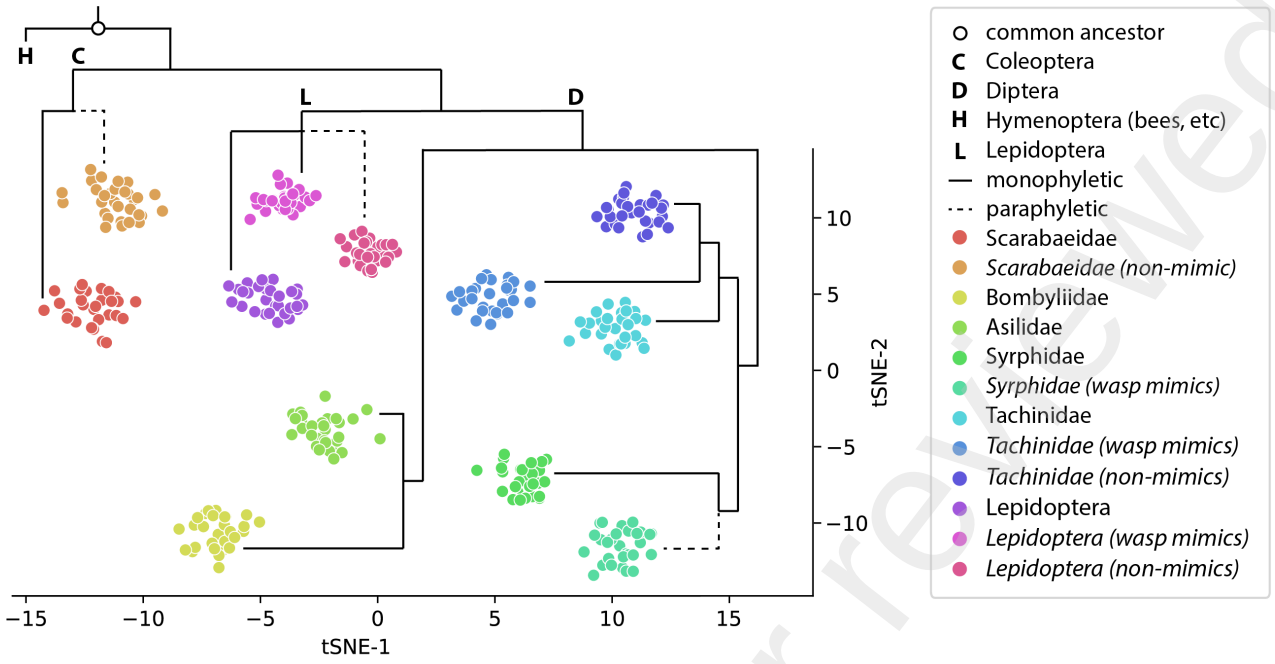

Figure 7: t-SNE plot of bee mimics (n=180), *wasp mimics* (n=90), and *non-mimics* (n=90) from Table 2, with the phylogenetic tree overlaid. Note the perfect clustering within each group, as well as the clustering between groups that grossly corresponds to the phylogeny. Evolutionary relationships follow [19], [20], [21], [22], and [23].

was the bumble bee classifying model yielding a lower accuracy for the Lepidoptera non-mimics compared to bee mimics (72.78% compared to 76.11%; also, the difference was negligible for Scarabaeidae).

However, our second hypothesis was not supported, as the accuracy of the wasp mimics was found to be intermediate between bee mimics and non-mimics in only one of the four comparisons, including the same near perfect classification (99.45%) for wasp mimics within Tachnidae and Lepidoptera by the bumble bee/ResNet-101 model. Thus, while the bee mimics are better at fooling both algorithms, the phenotypic divergence of the wasp mimics may be more easily detected and not confused with the appearance of bees. Phylogeny may play a non-mutually exclusive role within Tachinidae as well, as the wasp mimics and non-mimics are sister groups (Figure 7), and have more similar accuracies compared to that of the bee mimics (based on both models; Table 5, 6). In the bumble bee model, note the perfect or near-perfect classification of the wasp mimics (Syrphidae, Tachinidae, Lepidoptera) across all three training image types (color, gray, color+gray) and both testing image types (color, gray).

All of the bee and wasp mimicry herein is at least in part defensive (Batesian), conferring aposematic warning to would-be predators by falsely appearing to be a stinging taxon. The worst bee mimicry, based on the highest classification accuracy, was exhibited by the bee hawkmoths (*Hemaris*, Sphingidae, Lepidoptera) in the bee model (54.45%) and the bee beetles (*Trichius*, Scarabaeidae) in the bumble bee model (84.44%) (Tables 5, 6).

Conversely, the best bee mimicry, as defined by the lowest AI classification accuracy, was exhibited by Asilidae, and in both models (18.33%, 47.22%). Herein, this robber fly family Asilidae is represented by the genus *Mallophora*, known as the bee killers. These flies mimic the appearance of bumble bees and carpenter bees, as well as the bee-like buzz while flying [24]. Intriguingly, these flies mimic bees not just to escape predation (defensive), but also to enhance their own predation given that they feed upon bees (aggressive). Such bidirectional trophic selective pressure may be responsible for the evolution of more accurate mimicry within *Mallophora*. Furthermore, asilids are unique in that their aggressive mimicry is actually two-fold, across multiple life stages – exploiting Kirbyan mimicry to prey on bee larvae, and Batesian-Wallacian mimicry to prey on bee adults. These results confirm our third hypothesis, given that the superlative bee mimicry of these bee killers is not only defensive, but also doubly aggressive. In other words, such extreme selective pressures across the life history of these particular mimics is presumably driving their superior mimicry.

Findings from the remaining groups are also consistent with this third hypothesis. The related sister group Bombyliidae – a large clade known as the bee flies (15 subfamilies, 5K described species; [25]) – are also aggressive mimics, but only in the Kirbyan sense. While their adult forms do not feed upon bees like those of *Mallophora* do,

Bombylius (Bombyliidae) larvae are similarly ectoparasitoids in the nests of bees [26]. Larvae of some Syrphidae and the bee mimic genus *Tachina* also parasitize bees [27]. Interestingly, these three groups exhibit accuracy values intermediate between the bee mimics in Asilidae and in Lepidoptera/Scarabaeidae, in the bee model. (Less so for the syrphids and tachinids in the bumble bee model, which may be due to their mimicking of honey bees rather than bumble bees as in the bombylids).

Among the "incorrect" classifications of the Scarabaeidae *Trichius* – those which "evaded" the algorithm's detection – the CAMs were often located on the bee-like hairy thorax, or the elytron/a (Figure 6, left). Elytra are the modified, hardened forewings of the beetle order Coleoptera, which is interesting in that this differentiates the order from both the bee mimics in Diptera and the bees in Hymenoptera. In *Trichius*, the elytra mimic the gold and black banding pattern of bee abdomens (Figure 6). We can quantify the role of elytra color (4%) vs. pattern (44%) based on the percentage of images in which the CAMs selected the elytron/a in the color "incorrect" images compared to the grayscale "incorrect" images: 48% (10/21) vs. 44% (7/16), respectively. Interestingly, the VGG16-based model trained only on color images performed much better on grayscale versions of the testing images (43.33% vs. 23.33%; Table 5).

When comparing training with color images and grayscale images, we notice a trend of grayscale images giving consistently similar performance when compared to training with color images (except for Bombyliidae in Table 5, and Scarabaeidae/ Bombyliidae - bee mimic and Asilidae in Table 6). This means that the AI models learn significantly from the lighter-weight grayscale images also. This phenomenon can be explained. Researchers have discovered that in some cases, when color may not be relevant to a classification problem, AI models work as good or sometimes better with grayscale images [28] [29] [30] [31]. In the case of insect morphology, textures can play a very critical role in classification - for example - a very unique T-shaped area of hairs on the thorax of *Bombus affinis*. For detecting such markers, grayscale images may be enough, and also AI models will not be confused by colors appearing in any images, which are likely to happen in citizen-generated images due to diverse sources of background, inconsistencies in light, camera capabilities and more. grayscale images overcome these issues, which aid in robust AI models for problems related to insect morphology analysis.

With respect to the t-SNE results (Figure 7), it is interesting that there was perfect clustering within each group, with no mismatched data points. Such clustering is surprising within the paraphyletic groups (Scarabaeidae non-mimics, Syrphidae wasp mimics) and especially the Lepidoptera non-mimics, which comprise five distinct families within four superfamilies. Furthermore, given the imperfect mimicry within Syrphidae [22], it is notable that there was such clear separation of the bee mimic and wasp mimic clusters. This finding is inconsistent with the multimodel hypothesis [32], which would predict overlap due to imperfect mimicry of multiple models. This separation also corroborates the findings of Penney et al 2012 [22], which found no syrphids intermediate in appearance between hymenopteran models.

Distance within the plot cannot be used as an exact proxy for evolutionary distance, due to the probabilistic nature of the t-SNE approach ([33], [34]). However, it is important to note that the pattern of clustering between groups grossly corresponds to the evolutionary relationships, as denoted by the phylogenetic tree (Figure 7). Specifically, all within-clade clusters group together for Scarabaeidae, Lepidoptera, Syrphidae, Tachinidae – and even the Asilidae+Bombylidae clade, which represents the superfamily Asiloidea.

Conclusion

Our methods in this study can enhance the operational efficiency of citizen scientist-driven identification of bees, while also piquing interest among the public towards conservation. Furthermore, by applying machine learning techniques within an evolutionary framework – from our novel integration of t-SNE and phylogeny, to CAMs and grayscale images – we can approach not just "explainable AI" but "explainable mimicry". These methods provide a new solution to the long-standing challenge of "quantifying the extent of mimetic fidelity between mimics and models" [22], such as the role of color decoupled from pattern. By leveraging citizen science imagery from around the world, these methods yield a scalable and useful tool for bee conservation and ecology, as well as future studies of Batesian, imperfect, and aggressive mimicry.

Methods

Herein we designed a separate convolutional neural network (CNN) for two problems: a) classifying bees vs. other insects (including mimics) and b) classifying bumble bees vs. non-bumble bees. We present technical details now.

Layer	Input size	Output size
VGG16 5 Conv blocks	224, 224, 3	7, 7, 512
global_average_pooling2d	7, 7, 512	512
dense_1 (Dense)	512	256
dropout_1 (Dropout)	256	256
dense_2 (Dense)	256	128
dropout_2 (Dropout)	128	128
dense_3 (Dense)	128	64
dropout_3 (Dropout)	64	64
dense_4 (Dense)	64	2

Table 7: VGG architecture details

Hyperparameter	Value
Loss	Binary Cross entropy
Optimizer	Adam Optimizer
Momentum	0.9
Early training epochs	50
Whole model training epochs	100
Learning rate for all epochs	0.0005

Table 8: VGG Hyperparameters

Data pre-processing: The image data source was the iNaturalist platform. Images there are typically large in size, and can go upto $2,048$ pixels in the longest dimension. Processing images of these sizes can be very complex and time-consuming. To speed up learning without compromising accuracy, we reduced the size of each image to $1,024$ pixels in the longest dimension. To evaluate the effect of color on classification accuracy, we trained our models on color images, gray images, and combination of color and gray images. The first dataset retained the original colors of the images from iNaturalist. For the gray versions, we converted all the images into grayscale using OpenCV [35]. In the third dataset, we combined all the color and gray versions of the images, thereby doubling the training dataset. We did the same for testing dataset also. The procedure was executed for both problems.

VGG16-based CNN for classifying bees from other insects: We designed a VGG16-based CNN for our first problem - classifying bees from other insects. The VGG16 architecture consists of five blocks of convolutional layers, followed by three fully connected layers [14]. It is an architecture that is simpler in size and complexity than most other standard CNN architectures. For our problem, after the five blocks of convolutional layers, we added our own four fully connected layers. Details of the architecture are shown in Table 7, where the first row represents the last layer of the base VGG16 architecture. Upto this layer we kept the VGG16 base model architecture as it is. On top of that layer, we have added one global pooling layer and four connected dense layers. Those layers are described from second to ninth row in table 7. For training, we followed the standard procedure of freezing the weights of the base VGG16 architecture for the first 50 epochs so that already trained weights are still retained, and weights for only the newly added layers are trained. Then, weights in the entire architecture were un-frozen and re-trained again for 200 epochs. Table 8 presents the critical hyperparameters in our architecture during training and validation. The loss function is the binary cross entropy loss function, which is given by $-log(p)$, where p is the model estimated probability of the ground truth class, which we want to minimize during training and validation.

ResNet-101-based CNN for classifying bumble bees from other bees: Classifying bumble bees from non-bumble bees is a more complex problem, since there are subtler difference between these two classes as compared to classifying bees from other insects. For this problem, the CNN architecture that worked best in our study was the more complex ResNet-101. ResNet-101 is a CNN with residual connections, wherein each layer, instead of feeding only into the next layer, also directly feeds into layers several hops away [15]. This was done to specifically improve learning at later layers. Thus, this is a more complex architecture with 101 blocks of convolutional layers. This architecture provided optimal results in its current form; therefore, we did not change the architecture, but changed only the weights via training and validation. Table 9 presents the critical hyperparameters in this architecture for training and validation. The loss function is once again the binary cross entropy loss function.

CAM generation: To get a better understanding how our models interpret pixels in an image for classification, we adopted the technique of Class Activation Maps (CAM). The CAM technique gives each pixel a weight which indicates the significance of that pixel towards classification. To execute the technique, we compute the output

Hyperparameter	Value
Loss	Binary Cross entropy
Optimizer	Adam Optimizer
Momentum	0.9
Epochs	100000
Learning rate (upto 50000 epochs)	0.0003
Learning rate (upto 80000 epochs)	0.00003
Learning rate (upto 100000 epochs)	0.000003

Table 9: ResNet-101 Hyperparameters

features generated at the final convolution layer of the CNN. Then, we traverse back in the architecture (at the conclusion of the last convolutional layer) to determine the weight (probability) of each pixel in the image that was used for classification. A higher weight for a pixel indicates a redder color in CAM, meaning that the particular pixel was a more significant factor in classification. Pixels with lower weight would appear comparatively bluer in the CAM technique, and these are pixels that were not dominant in classification. Our CAM model was based on VGG16 and trained using color images.

Phylogeny with t-SNE: t-SNE stands for t-Distributed Stochastic Neighbour Embedding [34], an algorithm developed as an improvement over Stochastic Neighbor Embedding [16]. t-SNE is an unsupervised, non-linear technique for dimensionality reduction, primarily used for visualising high-dimensional data (in our case, RGB values from image pixels). In other words, t-SNE gives an intuition of how high-dimensional data points are related in low-dimensional space. Compared to many other non-parametric visualisation techniques (e.g., Sammon mapping, principal components analysis, isomap, locally linear embedding), t-SNE proved more robust and significantly more effective for high-dimensional data visualisation [36]. t-SNE can be used for data-visualisation in a wide range of applications including biomedical signal processing, genomics, computer security research, bioinformatics, cancer research, and music analysis [36].

To generate the t-SNE, a few steps are followed which we present briefly below. We first start with a base neural network architecture, and from this t-SNE algorithm runs a combination of two sequential phases. In the first phase, t-SNE builds up a probability distribution matrix for data points, which consists of the RGB values from image pixels. Each pair of distinct data points are considered. For each pair, a probability value is generated. If there is a high level of similarity between the two objects in that pair, a large probability value is assigned, otherwise the probability value is small. In the second phase, t-SNE considers those data points in a lower dimensional space and generates another probability distribution following the similar procedure it did in the first phase. The algorithm then tries to minimize the loss or difference between the two probability distributions with respect to the locations on the map. To accomplish that, the algorithm calculates the Kullback-Leibler divergence (KL divergence) [37] value and minimises it over several iterations.

To visualize phenotype vis-a-vis phylogeny for the mimic images in this paper, we also plotted the data points from images of bee mimics and outgroups using the t-SNE algorithm on a two-dimensional graph, and overlaid the phylogenetic tree that illustrates the evolutionary relationships. To build that graph, at first we trained a twelve-class VGG16-based classifier deep-learning model (similar architecture to that in Table 7) with 360 bee mimic images. We then extracted features from the final convolution layer for all 360 images. These features are a matrix of size $14 \times 14 \times 512$. We flattened the data to a 100352 sized array for each image. Then we ran the t-SNE algorithm (steps of which were discussed above) over the 360 flattened feature data, resulting in two-dimensional coordinates for each image. When we plotted those coordinates on a two-dimensional graph and marked each of those data points with twelve different colors depending on their family type, we observed that those data points created eight individual clusters, and data points within each family are situated in the same cluster.

Acknowledgements

We would like to acknowledge the contributions of citizen scientists on the iNaturalist platform, and in particular the photographs herein from psweet, savvaszafeiriou, derhennen, stevereekie1, johnwitton, bollyanna, matthew_wills, scibadger, suegregoire, greglasley, and dendzoscarab. This work was supported in part by the National Science Foundation under Grant No. IIS-2014547 to RMC and SC. Opinions, findings and conclusions are those of authors alone, and do not necessarily reflect the views of the funding agency.

Author contributions statement

Conceptualization, R.M.C.; funding acquisition, R.M.C. and S.C.; investigation, R.M.C., T.B. and S.C.; visualization, T.B. and R.M.C.; writing—original draft preparation, T.B., R.M.C. and S.C.; writing—review and editing, T.B., R.M.C. and S.C.; methodology, T.B, R.M.C. and S.C., Software, T.B. and S.C. All authors have read and agreed to the published version of the manuscript.

Additional information

We have no competing interests.

References

[1] Genise, J. F. *et al.* 100 ma sweat bee nests: Early and rapid co-diversification of crown bees and flowering plants. *PloS one* **15**, e0227789 (2020).

[2] Roffet-Salque, M. *et al.* Widespread exploitation of the honeybee by early neolithic farmers. *Nature* **527**, 226–230 (2015).

[3] Crane, E. *The world history of beekeeping and honey hunting* (Routledge, 1999).

[4] Darwin, C. *On the origin of species, 1859* (Routledge, 2004).

[5] Potts, S. G. *et al.* Global pollinator declines: trends, impacts and drivers. *Trends in ecology & evolution* **25**, 345–353 (2010).

[6] Osten-Sacken, C. R. *On the oxen-born bees of the ancients (bugonia) and their relation to Eristalis tenax, a two-winged insect* (Kessinger Publishing, 1894).

[7] Poulton, E. B. *The colours of animals: their meaning and use, especially considered in the case of insects* (D. Appleton, 1890).

[8] Pasteur, G. A classificatory review of mimicry systems. *Annual Review of Ecology and Systematics* **13**, 169–199 (1982).

[9] KIRBY, W. In w. kirby & w. spence. *An Introduction to Entomology* **2**, 529 (1817).

[10] Wallace, A. R. *Contributions to the theory of natural selection* (Macmillan and Company, 1871).

[11] Brower, L. P., Brower, J. V. Z. & Westcott, P. W. Experimental studies of mimicry. 5. the reactions of toads (bufo terrestris) to bumblebees (bombus americanorum) and their robberfly mimics (mallophora bomboides), with a discussion of aggressive mimicry. *The American Naturalist* **94**, 343–355 (1960).

[12] California Academy of Sciences, N. G. S. inaturalist. URL `https://www.inaturalist.org/`.

[13] California Academy of Sciences, N. G. S. inaturalist research grade. URL `https://www.inaturalist.org/posts/39072-research-grade`.

[14] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition (2015). URL `http://arxiv.org/abs/1409.1556`.

[15] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *CoRR* **abs/1512.03385** (2015). URL `http://arxiv.org/abs/1512.03385. 1512.03385`.

[16] Geoffrey, H. & Sam, R. Stochastic neighbor embedding (2002). URL `https://dl.acm.org/doi/10.5555/2968618.2968725`.

[17] Zhou, B., Khosla, A., Lapedriza, À., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. *CoRR* **abs/1512.04150** (2015). URL `http://arxiv.org/abs/1512.04150. 1512.04150`.

[18] Nvidia. Nvidia geforce titan x. URL `https://www.nvidia.com/en-us/geforce/graphics-cards/geforce-gtx-titan-x/s`

[19] Wiegmann, B. M. *et al.* Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC biology* **7**, 1–16 (2009).

[20] Gunter, N. L., Weir, T. A., Slipinksi, A., Bocak, L. & Cameron, S. L. If dung beetles (scarabaeidae: Scarabaeinae) arose in association with dinosaurs, did they also suffer a mass co-extinction at the k-pg boundary? *PLoS One* **11**, e0153570 (2016).

[21] Powell, J. A. Lepidoptera: moths, butterflies. In *Encyclopedia of insects*, 559–587 (Elsevier, 2009).

[22] Penney, H. D., Hassall, C., Skevington, J. H., Abbott, K. R. & Sherratt, T. N. A comparative analysis of the evolution of imperfect mimicry. *Nature* **483**, 461–464 (2012).

[23] Blaschke, J. D., STIREMAN III, J. O., O'hara, J. E., Cerretti, P. & Moulton, J. K. Molecular phylogenetics and piercer evolution in the bug-killing flies (diptera: Tachinidae: Phasiinae). *Systematic Entomology* **43**, 218–238 (2018).

[24] Linsley, E. G. Ethology of some bee-and wasp-killing robber flies of southeastern arizona and western new mexico (diptera: Asilidae). *University of California Publications in Entomology* **16**, 357 (1960).

[25] Evenhuis, N. & Greathead, D. World catalog of bee flies (diptera: Bombyliidae). revised september 2015 (2015).

[26] Yeates, D. K. & Greathead, D. The evolutionary pattern of host use in the bombyliidae (diptera): a diverse family of parasitoid flies. *Biological Journal of the Linnean Society* **60**, 149–185 (1997).

[27] Packard Jr, A. The parasites of the honey-bee. *The American Naturalist* **2**, 195–205 (1868).

[28] Ĉadík, M. Perceptual evaluation of color-to-grayscale image conversions. *Computer Graphics Forum* **27**, 1745–1754 (2008). URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2008.01319.x`. `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2008.01319.x`.

[29] Kanan, C. & Cottrell, G. W. Color-to-grayscale: Does the method matter in image recognition? *PLOS ONE* **7**, 1–7 (2012). URL `https://doi.org/10.1371/journal.pone.0029740`.

[30] Xie, Y. & Richmond, D. Pre-training on grayscale imagenet improves medical image classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018).

[31] Yohanandan, S., Song, A., Dyer, A. G. & Tao, D. Saliency preservation in low-resolution grayscale images. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[32] Edmunds, M. Why are there good and poor mimics? *Biological Journal of the Linnean Society* **70**, 459–466 (2000).

[33] Wattenberg, M., Viégas, F. & Johnson, I. How to use t-sne effectively. *Distill* **1**, e2 (2016).

[34] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).

[35] Culjak, I., Abram, D., Pribanic, T., Dzapo, H. & Cifrek, M. A brief introduction to opencv 1725–1730 (2012).

[36] Chatzimparmpas, A., Martins, R. M. & Kerren, A. t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Transactions on Visualization and Computer Graphics* **26**, 2696–2714 (2020).

[37] Perez-Cruz, F. Kullback-leibler divergence estimation of continuous distributions 1666–1670 (2008).