Why many Batesian mimics are inaccurate: evidence from hoverfly colour patterns

Christopher H Taylor¹, Tom Reader¹ and Francis Gilbert¹

¹ School of Life Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Mimicry is considered a classic example of the elaborate adaptations that natural selection can produce, yet often similarity between Batesian (harmless) mimics and their unpalatable models is far from perfect. Variation in mimetic accuracy is a puzzle, since natural selection should favour those that are most easily misidentified. A number of hypotheses exist to explain the persistence of inaccurate mimics, but testing them has been hampered by the difficulty in measuring pattern similarity, a key aspect of mimicry. Here, we use a recently developed method, based on the distance transform of binary images, to quantify pattern similarity both within and among species for a group of hoverflies and their hymenopteran models. This allowed us to test several key hypotheses regarding inaccurate mimicry. We find that levels of phenotypic variation are similar across most hoverfly species, providing evidence against the idea that selection is more relaxed in less accurate mimics. We also show that mimics do not have to compromise between accuracy to multiple model species. However, we find evidence for a trade-off between mimetic accuracy and thermoregulation which could account for many examples of inaccurate mimicry in temperate regions. Our results shed light on a classic problem concerning the limitations of natural selection.

Batesian mimicry | Imperfect mimicry | Syrphidae | distance transform

INTRODUCTION

Charles Darwin regarded mimicry as a beautiful example of the extreme results of natural selection (1, p.392), and the topic has since been well studied as a powerful and conspicuous demonstration of the evolution of phenotypes (2). Batesian mimics are harmless organisms that resemble a more dangerous "model" in order to deceive potential predators (3), and while some show an astonishing level of similarity to their models, others bear only a passing resemblance. Experiments show that these less accurate mimics do gain some protection from predators, but at lower levels than more accurate ones (4, 5). We would therefore expect the most accurate mimics in a population to have the highest fitness, and that natural selection should drive ever increasing perfection in resemblance to the model. Contrary to this prediction, there are many examples, including some snakes (6), spiders (7) and hoverflies (8), that seem far from accurate in their mimicry. By exploring this discrepancy between expectation and observation, the study of inaccurate Batesian mimicry provides an excellent opportunity to develop a better understanding of the ecological forces which determine the evolution of phenotypes.

There is no shortage of hypotheses proposed to address the existence of inaccurate mimicry, and these have been well reviewed elsewhere (2, 9-11). Here, we focus on one particular group of study organisms, the hoverflies (Diptera: Syrphidae) and some of the key hypotheses which have been proposed to apply to this group, although all could equally be applied to other systems. Hoverflies have been a major focus for studies of inaccurate mimicry, as the taxon comprises a large number of species, many of which are abundant and widespread, ranging from non-mimetic to highly accurate mimics of various hymenopteran models, with a wide range of inaccurate mimics in between (8, 11).

Theoretical explanations for inaccurate Batesian mimicry in hoverflies have produced a number of testable predictions about variation within and among mimic species. In particular, the evolution of mimicry can be predicted using Signal Detection Theory (12), which assumes that predators receive information from signals subject to noise, and therefore uncertainty. Signal Detection Theory suggests that, past a certain minimum level of similarity, further improvements in mimetic accuracy provide very little decrease in predation risk (13). Mimics that have reached this critical level of similarity will therefore experience relaxed selection. From this, Holloway et al. (14) make the prediction that more accurate mimics should show greater phenotypic variation. They suggest that less accurate mimics are under strong selection but lack the genetic variation to evolve closer similarity to the model, and hence have low phenotypic variation.

However, alternative predictions arise if we consider that mimic species may not all be equally attractive to predators. The threshold similarity level described above, beyond which selection is relaxed (13) depends on what has been described as the "incentive to attack" (15). A predator is less likely to risk an attack with an uncertain outcome if the cost of attacking a model is high relative to the benefit of consuming a mimic, or if the abundance of models is high relative to the mimics. Penney et al. (16) argue that smaller mimics have a lower calorific value to the predator, resulting in a low incentive to attack, and hence favouring relatively imperfect mimicry in smaller species. Furthermore, if small mimics offer a low incentive to attack, they are predicted to be under relatively relaxed selection by predators compared with larger species, and may therefore show greater phenotypic variability.

Significance

What are the limitations of natural selection? Many otherwise undefended animals gain protection from predators by resembling more dangerous prey, in a phenomenon known as Batesian mimicry. Inaccurate mimics are less effective in their deception and therefore ought to fall behind in the struggle for survival, yet many examples persist in nature, such as wasp-mimicking hoverflies. We test a number of suggested explanations by examining the similarity among a large number of hoverflies and the wasps they resemble. Our study reveals that hoverflies may have to compromise between wasp-like patterns to protect against predation and dark patterns in order to warm up quickly, and that this trade-off could account for the existence of inaccurate mimics.

Reserved for Publication Footnotes
Table 1. GLS models of within-species variability. The contribution of each predictor to the model was assessed using a likelihood ratio test. All tests had Δdf = 1. Sample size was 32 for females and 34 for males.

<table>
<thead>
<tr>
<th>sex</th>
<th>predictor</th>
<th>likelihood ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>accuracy:size</td>
<td>0.15</td>
<td>0.697</td>
</tr>
<tr>
<td></td>
<td>accuracy</td>
<td>0.021</td>
<td>0.8843</td>
</tr>
<tr>
<td></td>
<td>size</td>
<td>1.09</td>
<td>0.296</td>
</tr>
<tr>
<td>Male</td>
<td>accuracy:size</td>
<td>1.08</td>
<td>0.2977</td>
</tr>
<tr>
<td></td>
<td>accuracy</td>
<td>0.95</td>
<td>0.3293</td>
</tr>
<tr>
<td></td>
<td>size</td>
<td>0.74</td>
<td>0.3896</td>
</tr>
</tbody>
</table>

Table 2. MCMCglmm model of mimetic accuracy, which was logit transformed for normality. SSU was included as a random effect, with a variance structure that accounts for phylogenetic relatedness. Backwards model selection was used on the basis of the p values. Posterior means are quoted for coefficients of all predictors present in the minimum adequate model. All factors have df = 1. N = 638.

<table>
<thead>
<tr>
<th>predictor</th>
<th>posterior mean</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>proportion black</td>
<td>-0.191</td>
<td>0.564</td>
</tr>
<tr>
<td>thorax width</td>
<td>0.043</td>
<td>0.352</td>
</tr>
<tr>
<td>sex (F)</td>
<td>0.818</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>proportion black: thorax width</td>
<td>-0.231</td>
<td>0.018</td>
</tr>
<tr>
<td>proportion black: sex (F)</td>
<td>0.384</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>thorax width: sex (F)</td>
<td>-0.18</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>proportion black: thorax width: sex (F)</td>
<td>0.732</td>
<td></td>
</tr>
</tbody>
</table>

If selective pressures other than those imposed by predators influence the mimic's appearance, then inaccurate mimics could represent a trade-off between such opposing pressures. For example, increasing similarity to the model may come with a physiological cost, such as reduced ability to regulate temperature. In temperate climates, darker coloured insects are able to warm up more quickly (17, 18), and thus improve performance in areas such as flight activity (19). We might therefore expect to see a conflict in cooler regions between the bright colours required for mimicry and dark colours that allow effective temperature regulation.

Finally, a trade-off could exist even if predator behaviour were the sole cause of selection on mimetic patterns. Computer models predict that mimics with an intermediate similarity to several model species can be better protected than an accurate mimic of a single model species (10, 13), and thus increasing similarity to one model might come at the cost of lower accuracy to another. It is highly likely that predators will encounter more than one model species in their foraging, but the extent to which this influences inaccurate mimicry is not known (10, 11).

Despite the wealth of theories which seek to explain inaccurate mimicry, only recently has attention turned to testing the various hypotheses against each other in real systems. Penney et al. (16) carried out a comparative study of 38 hoverfly species, along with 10 putative models, using both morphological data and human judgement to measure degree of similarity. They found evidence that inaccurate mimics are not just artefacts of human perception, and suggested that no species are intermediate between several models. However, they found a positive relationship between size and mimetic accuracy, which they interpret as evidence for the relaxed selection theory, suggesting that larger hoverflies are more valuable prey and therefore under stronger selective pressure.

Another comparative study by Holloway et al. (14) investigated the levels of phenotypic variation in a number of hoverfly and wasp species. They used rankings of mimetic accuracy as calculated from behavioural responses of pigeons recorded in Dittrich et al. (4), and were consequently limited to the few species used in the pigeon study. Holloway et al. (14) found high levels of variation in many species, giving no indication that a lack of genetic variation was limiting the evolution of accuracy.
They did not find a clear trend between mimetic accuracy and phenotypic variation, although particularly high variation in the model species and one accurate mimic, *Temnostoma vespiforme*, led them to conclude that relaxed selection may be acting in those cases.

The few empirical studies which have attempted to test theoretical predictions about variation in mimetic accuracy have been constrained by the difficulties of characterising mimetic phenotypes systematically and objectively in order to measure similarity to relevant model taxa. For example, Holloway et al. (14) characterised mimetic phenotype simply using the proportion of yellow versus black on two tergites of the abdomen. The descriptors that Penney et al. (16) used to create a multivariate measure of mimetic accuracy include morphometric data (e.g. antenna length, thorax width, wing length) as well as some summary variables relating to the abdominal pattern (e.g. mean RGB values, number of stripes), but very little about the pattern itself.

Recently, we have developed a new objective measure of mimetic accuracy by comparing entire abdominal patterns using the distance transform method (20). Our method provides a measure of mimetic accuracy much closer to human and avian estimates than previous empirical measures, even without the inclusion of any morphometric data (20). In the current study, we use this new methodology to characterise the mimetic patterns of hoverflies in detail, and to test some of the predictions which have emerged from theoretical work. We plot a large number of model species and mimic individuals in 'similarity space', giving a picture not only of how species compare with one another in appearance, but also of the variation within species. We then test four predictions associated with three theoretical explanations for the existence of inaccurate mimicry:

1. Relaxed selection
   a. Lack of genetic variation: Less accurate mimics are under strong selection but lack the genetic variation to evolve increased accuracy; more accurate mimic species experience relaxed selection and thus have higher levels of phenotypic variation.
   b. Incentive to attack: Less accurate mimic species have higher levels of phenotypic variation since they provide a lower incentive to attack and are under more relaxed selection.

2. Multiple models: Increasing accuracy to one model decreases accuracy to others; inaccurate mimics represent a compromise between two or more model phenotypes.

3. Thermoregulation: Less accurate mimics have more black in their pattern and hence will be better able to regulate their temperature; there is a trade-off between accurate mimicry and effective thermoregulation.

RESULTS

Sexual dimorphism

We found evidence for size dimorphism in seven of the mimic species in our dataset, and for pattern dimorphism in a further eleven (Table S3). Males and females of these species were treated separately in subsequent analysis, while for all other species, data were pooled between the sexes. Of the 72 resulting SSUs, 51 were classified as mimics of *Vespula vulgaris*, 11 of *Apis mellifera*, seven of *Vespula crabro*, and three of *Vespula germanica*. The level of mimetic accuracy to the assigned model varied from 0.55 to 0.87 (Table 1).

1. Relaxed selection

   For both females and males, the OU evolutionary model was found to be a significantly better fit to the data than BM (females: LR = 11.71, p = 0.0006; males: LR = 6.10, p = 0.014; both df = 1). In neither case was size or mimetic accuracy significantly associated with phenotypic variability of sampled species (Table 2 and Figure 2; see also Table S4). When the analysis was repeated with accuracy calculated on the basis of all 12 model species, accuracy had a significant negative effect on variability, but this appeared to be artefactual - see Supplementary Results.

2. Multiple models

   Mimetic accuracy values of SSUs to the four main models are strongly positively correlated, particularly for accuracy to the two *Vespula* species, and to a slightly lesser extent *Vespula crabro* (see Table S5).

   Considered separately, almost all SSUs show either a significant positive correlation or no significant correlation among similarity values to the four main model species (Tables S6 and S7). There was only one negative correlation with p < 0.05: in males of *Srphus ribesii*, accuracy to *Apis mellifera* was negatively correlated with accuracy to *Vespula crabro* (r = -0.56, p = 0.009, N = 21). Under the null hypothesis, if all tests were independent, we would expect 10 negative correlations through type I error on average.

3. Thermoregulation

   Mimes with a greater proportion of black on their abdomen tend to be less accurate to their model (slope = -0.191, p = 0.564). This trend is more pronounced in larger hoverflies, as shown by a significant interaction between proportion black and thorax width (p = 0.018). The effect of size alone is not significant (slope = 0.043, p = 0.352). There is a significant effect of sex, with females in general being more accurate (p < 0.001). In addition, both proportion black (p < 0.001) and thorax width (p < 0.001) interact with sex, with females showing a weaker version of the trend described above (Table 3 and Figure 3; see also Table S8 and Figure S3).

DISCUSSION

By comparing colour patterns using the distance transform method (20) we can get an accurate picture of the mimetic relationships in a community of insects. With the inclusion of multiple individuals per species, all from the same geographic area, we have been able to study variation in mimetic accuracy both within and among species. The lack of a trend between accuracy and phenotypic variation suggests that inaccurate mimics are not accounted for by the fact that they have not been able to evolve to the point of maximum protection (Prediction 1a) or by relaxed selection caused by a reduced incentive of predators to
attack (Prediction 1b). Rather, the data suggest that inaccurate phenotypes represent the result of a trade-off between opposing selective pressures. A trade-off caused by selection for similarity to multiple models (Prediction 2) is not supported, but the results suggest a role for thermoregulation in limiting the adaptive value of increased accuracy (Prediction 3).

The absence of a trend in phenotypic variation with mimetic accuracy and the relatively high levels of phenotypic variation are broadly in line with the results from Holloway et al. (14). It therefore seems unlikely that inaccurate mimics are limited by a lack of genetic variation. We cannot tell from these data how much of the variation is heritable; at least some will be attributable to measurement error, and some to phenotypic plasticity, as adult patterns are known to change with the temperature experienced by the puparium (21). However, the few studies of the genetic component of pattern variation in hoverfly species have found a high level of heritability in those cases (22, 23).

The relaxed selection hypothesis predicts that, above a certain level of similarity, any further improvements in mimetic accuracy are selectively neutral (13). Penney et al. (16) found a correlation between size and morphomimetic similarity to the model, and argue that smaller prey items are less valuable, and so relaxed selection allows the persistence of inaccurate mimicry in smaller hoverflies. However, a predator’s optimal diet depends not only on the caloric value of the prey but also on search and handling times (24), and it is not clear whether large hoverflies provide the best trade-off in that regard. Furthermore, although Penney et al. (16) found that larger hoverflies tend to be more similar to their models in terms of morphology, our results reveal a more complicated relationship between pattern similarity and size. There is no direct effect of size on accuracy (Table 3) although there is an interaction with the colour proportions of the abdomen (see below), and in the case of males, the smallest are indeed the least accurate (Figure 3). Most importantly though, our data show no association between phenotypic variation and either size or mimetic accuracy. If hoverflies are indeed experiencing relaxed selection, then it appears to be present across the board, and thus we are no closer to explaining the observed variation in mimetic accuracy.

We are left with the likely explanation that there is some kind of opposing selective pressure that is balanced against the advantage of increased mimetic accuracy. The multiple models hypothesis provides one possibility. With our observational dataset, it would be very challenging to test whether multiple models are influencing the evolution of mimetic colour patterns. The formulation of a null model, in which each mimic is influenced by only a single model species, would require numerous assumptions regarding which areas of phenotypic space are potentially reachable, in order to test whether intermediate phenotypes are disproportionately common. What we can do, however, is to ascertain whether there is a potential trade-off mediated by multiple models, assuming that the level of protection experienced by an individual is indeed influenced by more than one model species.

In terms of shape, hoverflies are clearly distinct from Hymenoptera, and there are no hoverflies that occupy a phenotype intermediate to two or more model species (16). In terms of pattern, we find the distinction is less clear. When we ordinate our data in 2D space, there are a large number of hoverfly individuals that, for example, occupy the space in between Apis mellifera and Vespa spp. (Figure 1). However, we cannot necessarily assign an adaptive explanation for these intermediate phenotypes. For each species of mimic, there is either no correlation or a positive correlation among similarity values to each potential model species. This implies that, at least in terms of pattern, there is no multi-model trade-off: assuming the observed variation has an underlying genetic component, it would be possible for each mimic to improve its similarity to one or more models without compromising similarity to others.

In contrast, a trade-off between mimicry and thermoregulation is supported by the data. Hoverflies maintain a temperature excess (a body temperature above that of the surrounding air) though a combination of basking and shivering (25). Darker coloured insects absorb more solar radiation, and therefore can heat up more rapidly (17, 18), so we might expect darker hoverflies to be at a fitness advantage in cooler conditions. More rapid temperature gain during basking will reduce the opportunity cost of thermoregulation as well as possibly reducing predation risk.

In support of this, a number of hoverfly species have been found to show seasonal variation in their colour patterns, with darker morphs being more common outside the summer months (21), which is thought to have an adaptive function in relation to temperature regulation (26).

However, the results of our study show that the thermoregulatory benefits of darker patterns will also likely be associated with a reduction in mimetic accuracy. To be a perfect wasp mimic would require the amount of black on the abdomen to be limited to 51% (Table S2), but almost all hoverflies surveyed were above this value (Table 1). Aposematic signals are known to constrain temperature regulation, as observed in the moth Parasemia plantaginis (27). Moths with more black on their body were able to warm up more quickly, but suffered increased predation due to a less effective warning signal. Thus it is highly plausible that...
hoversfly colour patterns are constrained by their thermoregulation function. By contrast, wasp abdominal patterns are likely to be less constrained, since they do not rely much on basking for thermoregulation; social wasps achieve a high temperature excess through endothermy before they even leave their nest (28).

Interestingly, we find that the constraint on the colour pattern seems to be stronger in larger individuals, as revealed by the significant interaction between the proportion black and size. Larger insects are able to maintain a higher temperature relative to the ambient, but have slower heating rates (29). Thus any differences in rates of warming caused by colour are likely to have a greater effect on fitness in larger than in smaller insects, the latter being unable to depart far from ambient temperature and so rates of warming are less likely to be a relevant factor. Indeed, both theoretical predictions (30) and physical models (31) have shown that colour should have a greater effect on temperature in larger organisms.

For models, our study focused on four common species of Hymenoptera which are often regarded as the targets of mimicry in European hoversfly communities (11), but we caught a number of other Hymenopteran species in small numbers, which could potentially also serve as models. There are other species of bee and wasp that we did not encounter at all, but which may also have an influence on the evolution of mimicry in hoversflies. On the other hand, their lower abundance and/or visibility during our collection suggests that predators too will encounter them at a low rate, and therefore their significance as models is likely lower than those species that are widespread and conspicuous. Nonetheless, conclusions are similar when we incorporate these rarer model species into the analysis (see Supplementary Results).

By studying not just the most extreme and striking examples of adaptation, but the full variety that we see in nature, we can better understand how the same selective processes can lead in very different directions. Our work highlights the conflict between mimicry and thermoregulation in temperate regions, and more generally underlines the importance of trade-offs between ecological and physiological functions in shaping observed phenotypes. To resolve the puzzle of imperfect mimicry fully, further work is now needed to compare communities from different climates, and to gain more detailed insight into thermoregulatory mechanisms in mimetic species.

**MATERIALS AND METHODS**

Image processing and dissimilarity calculations were carried out in MATLAB (32). Statistical analyses were carried out in R version 3.0.3 (33).

**Specimens**

Insects were collected using a hand net from wild communities in Nottinghamshire, UK (particularly the Attenborough Nature Reserve) and surrounding areas, during May to October in the years 2012-2014. See Table S1 for full details of sampling sites. Target insects were any hoverflies or stingless Hymenoptera bearing a two-colour pattern (usually black and yellow), but excluding bumblebees and their putative mimics, which are notably excluded from the analysis as, not having a sting, their status as models is (under current conditions) debatable (they may still be unpalatable to predators due to other factors; 34). A total of 968 individuals were identified to species level and sexed using relevant keys (35-37).

They were euthanised by freezing and then their abdomens photographed, with legs and wings pinned out to the sides when necessary to give a clear view of the abdomen. They were placed inside a homemade photo studio – a white 18 x 10 cm open-topped box. A 5 mm scale bar was placed near to the insect. The specimen was photographed from above with a Canon 600D DSLR camera and Tamron 90mm macro lens under natural outdoor light conditions, in the shade. This method resulted in an image that was evenly lit and free from strong reflections or glare. While natural weather variation did lead to some changes in brightness from image to image, this did not affect the analysis since images were binarised before comparison (see “Image processing”).

**Model species**

Four species of Hymenoptera were treated as the main models of the sampled community: Vespula vulgaris (common wasp), Vespula germanica (German wasp), Vespa crabro (hornet) and Apis mellifera (honeybee). These were the only potential models that were caught in any significant number from the many hundreds of other wasp species that were also encountered. We know from both theory (38) and experiments (39) that a model’s importance in shaping predator behaviour increases with its abundance, and therefore we have excluded these low-abundance models from our main analysis. However, we ran a supplementary analysis including these rarer model species (see Supplementary Results).

**Image processing**

Images were selected, cropped, and rescaled to a standard alignment, and an algorithm was applied to remove noise and sharpen edges. An edge detection algorithm was used to find the outline of the abdomen. In some cases, a rough outline was drawn manually and passed to the algorithm as a starting point to find cases where the outline was difficult to detect against the background.

The abdomen was automatically segmented into two colour regions (typical mimic species in our dataset for which we have >10 data points) to produce clear segmentations, often due to fading of the colours after death (C. Taylor, pers. obs.) and were discarded from further analyses. To quantify the colour proportions in the pattern, we calculated the proportion of pixels within the abdominal image that were classified as “black” (i.e. the darker of the two segments) after segmentation.

See Supplementary Methods and Figure S1 for more detail on the image processing.

**Mimetic accuracy**

We calculated dissimilarity values for all possible pairings of images within each model species (40). For each species, we calculated the centroid for each SSU, and then the distance, they obey the triangle inequality), which is not always the case when using non-metric multidimensional scaling, as we considered it important to use a method in which the resulting inter-point distances would be linearly related to the original distance matrix. We did this in order to preserve the magnitude of the variation in the dataset, despite the fact that CMDS assumes that distances between individuals are metric (that is, they are linearly related to the original distance matrix). We did this in order to preserve the magnitude of the variation in the dataset, despite the fact that CMDS assumes that distances between individuals are metric (that is, they obey the triangle inequality), which is not always the case when using non-metric multidimensional scaling.

On the basis of a scatter plot, we chose the first four dimensions of the CMDS as the best representation of the data. Using these four dimensions, we calculated the centroid for each SSU, and then the distance, z, of each individual to it via lineal regression. The mean of a group’s z values provides a measure of within-group variability (42).

To test for an influence of mimicry accuracy on the rate of warming, we used a generalised least squares model (GLS; 44) in the R package “ape” version 3.1-1 (45). GLS is equivalent to a general linear model, but with the inclusion of a correlation matrix derived from the species’ phylogeny to control for relatedness among species. We used mean z value for an SSU as the response and mean mimetic accuracy and mean thorax width (plus their interaction) as predictors. Thorax width was included in the model as a proxy for size (46), because Penney et al. (16) argued that larger hoverflies starting offer a larger “incentive to attack” due to their greater nutritional...
value. The width of the thorax at the base of the wings was measured in image using the unprocessed images, using the stem bar from each image to set the scale. Note that in the early stages of the project, photographs did not include a scale bar, and therefore in some cases (e.g. see Table 1) samples for size measurements are smaller than for other measures such as pattern.

We ran the model under two different evolutionary scenarios: Brownian motion evolution (BM), and Ornstein-Uhlenbeck evolution (OU; similar to Brownian motion, but traits are constrained towards an “optimum” value). These different scenarios were represented by two different correlation matrices passed to the GLS model, calculated from a composite phylogeny (Figure S2) based on information from Rotheray and Gilbert (47) and Stãhls et al. (48). We selected the evolutionary model that gave the better fit on the basis of a likelihood ratio test. We then used backwards stepwise model simplification with likelihood ratio tests to find the minimum adequate model. In order to allow for sexual dimorphism, we conducted two separate analyses, one with data from only female individuals, and the other with only males.

2. Multiple models
To test for a potential trade-off in similarity to multiple models, we tested within SU5 for correlation (using Pearson’s r) in mimetic accuracy towards the four main model species. A negative correlation would imply that, for a given SU5, increasing similarity to one model comes at the cost of decreased similarity to another. We tested all SU5 for which we had data on at least six individuals.

3. Thermoregulation
We tested for a trade-off between accuracy and the extent of black in the pattern (‘proportion black’) using a Markov Chain Monte Carlo general linear mixed model, implemented in the R package “MCMCglmm” (49). Again, this method allowed us to control for phylogenetic relatedness among species. Accuracy of individual mimics to their closest model was the response variable, logit transformed for normality of residuals. Fixed effects were the proportion black, thorax width (see above), which was included as size can have a major impact on thermoregulation (50), and sex. Species was included as a random effect, and we calculated a covariance structure for the random effect based on the phylogenetic tree (Figure S2; also see “relaxed selection” above). We then used backwards stepwise model simplification based on p values to find the minimum adequate model.

**Supporting Information**
-Supplementary Methods (image processing)
-Supplementary Results and Discussion (rare model species)
-Figures S1-3 and Tables S1-8

**Acknowledgements**
We would like to thank staff at the Attenborough Nature Reserve and Nottinghamshire Wildlife Trust for permission to collect specimens, and Katie Threadgill for assistance with fieldwork.
Please review all the figures in this paginated PDF and check if the figure size is appropriate to allow reading of the text in the figure.

If readability needs to be improved then resize the figure again in 'Figure sizing' interface of Article Sizing Tool.