Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/8/eaaw4967/DC1

Supplementary Materials for

Deep learning on butterfly phenotypes tests evolution's oldest mathematical model

Jennifer F. Hoyal Cuthill*, Nicholas Guttenberg*, Sophie Ledger, Robyn Crowther, Blanca Huertas

*Corresponding author. Email: j.hoyal.cuthill@elsi.jp (J.F.H.C.); ngutten@gmail.com (N.G.)

Published 14 August 2019, *Sci. Adv.* **5**, eaaw4967 (2019)
DOI: 10.1126/sciadv.aaw4967

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/8/eaaw4967/DC1)

Table S5 (Microsoft Excel format). Mean pairwise squared Euclidean phenotypic distances between subspecies from *H. erato* and *H. melpomene*.
Table S6 (Microsoft Excel format). Mean pairwise Euclidean geographic distances between subspecies from *H. erato* and *H. melpomene*.
Table S7 (Microsoft Excel format). Number of sampled butterfly individuals for each subspecies.
Table S8 (Microsoft Excel format). Broad pattern class of the type specimen of each subspecies.
Table S9 (Microsoft Excel format). Traditionally hypothesized mimicry complexes of *H. erato* and *H. melpomene* subspecies.
Table S10 (Microsoft Excel format). Statistical comparisons of pairwise Robinson-Foulds distances between sets of phylogenetic trees.
Table S11 (Microsoft Excel format). Statistical comparisons with hybrids excluded of pairwise Robinson-Foulds distances between sets of phylogenetic trees.
Supplementary Computer Code in ipynb Format
Supplementary Computer Code in PDF Format

Supplementary Methods

Image acquisition and pre-processing

1269 butterfly specimens from the species *Heliconius erato* and *H. melpomene* were photographed at the Natural History Museum London (NHM) by two full-time research assistants over a two month period in 2016 (specimen numbers and image filenames, Table S2). Data collection costs were £4207 ($5807) for the complete dataset, averaging £3.31 ($4.57) per specimen and £1.66 ($2.29) per photograph. A consistent photographic setup was used throughout with a Nikon digital camera, ring light, light-grey foam background, cm to 0.5 mm rulers, and colour analysis chart constructed from a Spyder Checkr 24 colour card. Photographs were then screened for poor image quality or specimen wing pattern damage, giving a final dataset of 1234 butterflies and 2468 photographs, including a dorsal and ventral photograph of each butterfly. Photographs were cropped to include only the butterfly specimen and a 10 pixel border by image segmentation using MatLab scripts. For deep learning, images were then re-sized to a consistent, low-resolution height of 64 pixels (maintaining the original image aspect ratio and padded to 140 pixels wide). Moderate image compression improves the performance of deep learning by reducing the number of pixels that must be compared between images. The photographic dataset used for deep learning is provided in the Dryad Data Repository with filenames corresponding to joined data in Table S2: doi:10.5061/dryad.2hp1978.

Taxonomy, locality data and specimen selection

Taxonomic and locality data were recorded from NHM *Heliconius* butterfly specimen labels (Table S2). Capture localities were then matched to approximate latitudes and longitudes (Table S2) based on NHM records. Subspecies taxonomy follows reference (*30*). The complete photographic dataset covers thirty-seven named subspecies and one labelled cross: 21 subspecies from *H. erato* and 17 from *H. melpomene*. Specimens of these subspecies were sampled exhaustively from the NHM

collection. All available specimens of *H. erato* and *H. melpomene* were selected, within the limits of the data collection period, moving systematically through collection drawers. The complete photographic dataset (fig. S4, Dryad Data Repository: doi:10.5061/dryad.2hp1978) covers both specimens closely representative of subspecies descriptions (*30*) (*31*) (including available holotypes, syntypes, and paratypes, Table S2) as well as other, naturally varying, individuals. These variants include some likely hybrid specimens showing varying levels of phenotypic admixture from other subspecies (see additional taxonomic information, Table S2). Inclusion of all available specimens in machine learning covers a very broad range of the phenotypic diversity within these species, providing the deep learning network with all available information from which to learn phenotypic features correlated with subspecies identification (see deep learning methods, below). The extent to which the named subspecies (*30*) are objectively distinguishable was then explicitly tested based on classification accuracy during network testing (Deep learning methods, below). Locations for all sampled specimens in a phenotypic spatial embedding (Table S3) were calculated (Deep learning methods, below) permitting further analysis of phenotypic distances between any subset of specimens. Two sets of statistical analyses were then conducted, one set including all 1269 photographed butterfly specimens and the second set excluding potential hybrid specimens to give a reduced dataset of 815 specimens and 1630 photographs (Table S2, fig. S4).

The extent of phenotypic similarity between subspecies, including Müllerian co-mimics, was explicitly tested using statistical analyses of phenotypic distances generated by the deep learning network (see statistical analyses of phenotypic distance, below). All sampled specimens were included in the main statistical analyses. Specimen selection was therefore independent of hypotheses of mimicry (e.g. as referring only to standard phenotypes of taxonomic type specimens (*31*)). This facilitated conservative tests of the extent of mimicry among wild-caught specimens, without any potential bias from specimen exclusion on our part. Supplementary statistical analyses on the reduced dataset (with hybrids excluded) then enabled further testing, independent of any

effect from a preponderance of hybrids, which may potentially be overrepresented in museum collections relative to the wild.

Numbers of butterfly individuals sampled from each subspecies (Table S7) were variable, reflecting differences in abundance within the Natural History Museum collection. The average number of sampled butterfly individuals per subspecies was 32, the maximum number was 130 (for *H. erato petiverana*) and the minimum number was 1 (for *H. melpomene penelope*). Of 1234 total specimens in the dataset, 60% were from *H. erato* and 40% were from *H. melpomene*. The average number of sampled butterfly individuals per subspecies for *H. erato* was 35, for *H. melpomene* this was 29. Twenty-seven of the thirty-eight included subspecies were in one of twelve traditionally hypothesised (*12*) mimicry complexes (Tables S1, S9).

Deep learning

Image classification and spatial embedding were performed using a 15 layer deep learning network (Supplementary Computer Code), which we name ButterflyNet (figure 1). This makes use of a triplet embedding loss function (*21*) (*22*) to train a network to organise its inputs (images) in a space such that proximity in that space is highly correlated with identity (in this case subspecies). The network is trained on triplets of butterfly images, with each triplet containing two images sampled from the same subspecies and one image sampled from a different subspecies. The specific meaning of proximity is given by the distance function used in the loss function (the optimisation objective). In this study, the distance function was Euclidean distance. The learned embedding was then passed through an additional small network to perform direct categorical subspecies classification. Overall, the total network optimises the sum of the triplet loss and the categorical cross entropy

$$triplet\ loss = E[\|z_A - z_{A'}\|_2^2 - \|z_A - z_B\|_2^2] \quad (1)$$

Where E is the expectation over the dataset, $\|z_A - z_{A'}\|_2$ is the Euclidean distance, z_A is the spatial embedding of butterfly A, butterflies A and A' are sampled from the same subspecies and butterfly B is sampled from a different subspecies

$$categorical\ cross\ entropy = E[-\log(p(y))] \quad (2)$$

Where $p(y)$ is the probability assigned by the network of a given butterfly belonging to its named subspecies y.

The computer code used for machine learning is provided as a Python script (Supplementary Computer Code) in ipynb format which can be viewed using a text editor and run using the Jupyter Notebook App (http://ipython.org). The script makes use of the PyTorch, Scikit-learn and Adam packages.

Network training

Network training used the Adam optimizer with a learning rate of 10^{-4}. The method of training is as follows. Each batch is composed of sets of image triplets sampled from the training data. To generate a triplet, first the majority label is chosen, and a pair of images sharing that label are randomly selected. Then a third image is sampled under the constraint that its label does not match the majority label (but is otherwise distributed according to the distribution of labels in the training data). The batch composition is chosen such that each majority label is selected an equal number of times to compose the batch. Images are subjected to augmentation by a uniform random translation in the range of [-3,3] pixels on x and y. We trained for a total of 30000 batches, each composed of 99 image triplets. For the first 1000 batches, the loss function was constructed as an equally weighted sum of the classification and triplet losses, with L2 regularization applied with a coefficient 6×10^{-5}. After 1000 batches, the weighting applied to the classification loss was reduced to 0.1 relative to the triplet loss. This helps for training during the initial transient phase in

which the global structure of the embedding is first emerging by providing a stronger constraint on the embedding space (that it must contain sufficient information to predict the butterfly classes). These choices were determined as part of a hyperparameter search to optimize test classification accuracy as part of the process of development of the model (covering comparisons of regularization strategies, choice of activation function between ReLU and ELU, and general architecture dimensions and training process), but that search was not exhaustive and does not include some recent innovations such as ResNet-type architectures, nor does it include more extensive data augmentation strategies. Refinements of the model based on these avenues of exploration are left as possibilities for future work, and would likely be able to increase the classification accuracy further. In terms of structural considerations, in order for the embedding space to respect global topological relationships between the butterfly subspecies, it is necessary to have some relative ambiguity in classification so that positioning of clusters with respect to each-other has measurable consequences to the classifier loss. As such, if the network architecture is improved, it may be necessary at that time to either add additional data or refine the problem to maintain a fixed level of difficulty if the utility of the embedding is to be preserved.

Network testing

After the network was trained on 1500 images randomly sampled from the 2468 images in the dataset, network testing was performed on the remainder (968 images). Testing presents the trained network with new images, which it has not encountered before. The network then classifies the new images by subspecies, image classifications are compared to the known subspecies identities and the overall accuracy of test classifications is reported. The accuracy of classification expected simply by chance for this dataset was 5%. Additional testing was performed using a support vector classifier (SVC) trained on the embeddings from the main network (ButterflyNet). This tested the accuracy of classification of specimens to subspecies based on their locations in the phenotypic spatial embedding. This therefore tests the extent to which the spatial embedding locations generated by ButterflyNet are predictive of subspecies identity (e.g. relative to the original image data).

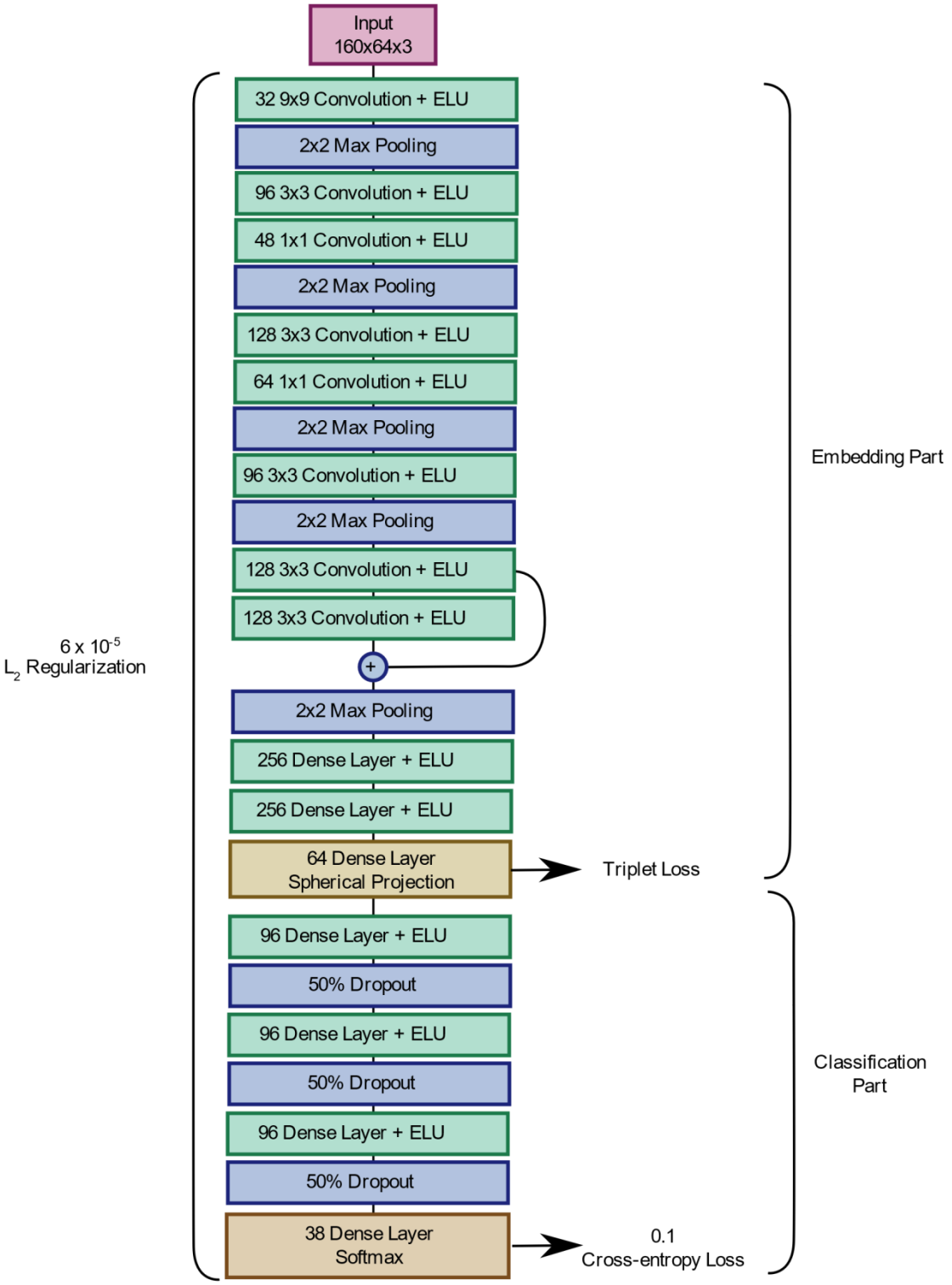


Fig. S1. Diagram of the architecture of the deep learning network ButterflyNet used in this study. Green boxes indicate intermediate network layers that perform matrix multiplications (image convolutions or dense, all-to-all, mappings). Blue layers perform operations on the output from previous layers (e.g. pooling or dropout operations). Brown layers indicate dense layers which produce a main output of the analysis (spatial embedding or image classification).

Statistical analyses of phenotypic distance

Pairwise Euclidean phenotypic distances between all images were calculated from the coordinates of all 2468 images within a spatial embedding with 64 dimensions, generated using the network (Table S3). These distances were then used to calculate the average pairwise Euclidean phenotypic distances between all subspecies (Table S4). Squared Euclidean distances were also calculated for heat-map visualisation (fig. S3, Table S5). Principal component analysis of the image coordinates was additionally used to visualise the principal component scores in the space of principal component axes 1 and 2 as well as 3 and 4 (Fig. 2). The Euclidean phenotypic distances between all subspecies (Table S4) were used in comparisons between different sets of unique unordered subspecies pairs (Fig. 3). The first comparisons (Fig. 3a) included identity pairs (diagonal of Table S4; average distances between images within each subspecies), pairs of subspecies traditionally hypothesised (*12*) to mimic each other (Table S1) and all other subspecies pairs (neither co-mimics nor identity pairs). The second comparisons (Fig. 3b) separated the two species to compare identity pairs from *H. erato* versus *H. melpomene*, other pairs (neither co-mimics nor identity pairs) where both members were of *H. erato* versus *H. melpomene*, and pairs of co-mimics in which one subspecies was of *H. erato* and the other of *H. melpomene*. Average phenotypic distances for subspecies sets and principal component scores were calculated using MatLab scripts. Nonparametric statistical analyses (robust to different sample sizes and non-normally distributed data) were conducted using the program Past 3, after Shapiro-Wilk's tests indicated that distances for some subspecies sets were non-normally distributed (using an alpha value of 0.05). These analyses included Kruskal-Wallis tests for equal medians and, where this overall test was significant, subsequent Mann-Whitney pairwise comparisons of statistical distributions between groups.

Testing evolutionary convergence

Relatively few examples exist of quantitatively demonstrated evolutionary convergence (*24*) (*25*). When considering categorical (discrete) traits, convergence can be defined as the repeated derivation of the same trait (e.g. phylogenetic character state) in two or more lineages (e.g. phylogenetic clades). For quantitative (continuous) traits, evolutionary convergence has been defined as an increase in similarity between two lineages (considering some specified axis or axes of variation) relative to their ancestral states (*25*). For consistency with categorical definitions, the broad phenomenon of evolutionary convergence may be considered to include some types of 'parallel' evolution (*24*) (*25*), such as parallel vectors of change to the same (or similar) trait values. Usually, ancestral states are unknown *a priori* (e.g. they must be estimated from contemporaneous taxa or non-contemporaries of uncertain ancestor-descent status). Consequently, tests of convergence in continuous traits have previously undertaken quantitative analyses of taxa that are qualitatively or functionally similar, and so potentially convergent, relative to respective sister-groups (immediate relatives) (*40*) (*25*) or to broader sets of close relatives (*24*). Where the putatively convergent taxa are quantitatively more similar to each other than are their relatives, this has been taken as support for convergence (*24*) (*25*). However, without additional information, especially on the temporal direction (polarity) of evolutionary change, it can be difficult to distinguish putative convergence from alternative patterns such as divergence by dissimilar sister taxa.

The studied case of *H. erato* and *H. melpomene* overcomes some such difficulties due to the sheer number of polymorphic mimicry types. For two compared clades with *n* cross-clade co-mimic pairs, each of which has a distinct pattern feature (or feature set), at most one of these distinct feature sets could potentially represent a shared ancestral state. All *n*-1, other co-mimic features must have been independently derived within each clade (since the minimum number of evolutionary derivations for a phylogenetic character on a given tree is the number of distinct character states minus one

(*41*)). A test of evolutionary convergence can then be applied in which the operational definition of convergence is essentially that for discrete traits (independent derivation of the same state e.g. in two clades) and quantitative analysis is employed to test the relative similarity of the traits in question and the number of quantitatively distinct trait states (e.g. clusters).

Comparative analyses of phenotypic convergence

To further explore the extent of reciprocal convergence in mimicry between *H. erato* and *H. melpomene*, comparative analyses (*40*) were conducted using twelve selected subspecies (fig. S7). First, two sets of subspecies were identified (each set including four subspecies), with each set consisting of two pairs of interspecies co-mimics in which conspecifics are nearest neighbours in the phenotypic spatial embedding, permitting phenotypic sister-group comparisons (fig. S7a-b). The ancestral pattern types and order of pattern evolution within *H. erato* and *H. melpomene* are not known with certainty. However, focal-co-mimics, with pattern features that are potentially derived, rather than ancestral, were identified for each comparative analysis based on all available independent information from gene phylogenies (*26*) (*27*), biogeographic distribution (fig. S2) and phylogeographic reconstruction (*26*). This additional information aids assessment of the most likely polarity of pattern evolution (e.g. directed towards the focal taxa). From a cladistic perspective, this process is equivalent to assessing the most likely phenotypic states at the hypothetical ancestral node for two considered taxa. For comparison, the analyses were then repeated with reversed polarity. Two focal subspecies and their nearest conspecifics (fig. S7c) were selected based on previous discussion of the influence of *H. melpomene* on *H. erato* (e.g. *H. erato petiverana* (*23*)), which has been historically controversial (*13*) (*23*). The position in phenotypic space of each of the focal subspecies was then compared to that of their nearest conspecific (a type of sister-group comparison (*40*) (*25*)). Compared distances were the squared distance from the mean location of the focal co-mimic, summed across all 64 spatial embedding axes, calculated using a Python script (Supplementary Computer Code). Expressing the locations in phenotypic space in terms of distance

from two focal taxa (fig. 5 g-j) enables two-dimensional visualisation of the distances among compared taxa across any number of phenotypic axes. This also facilitates tests of mutual convergence in which convergence is characterised by decreasing distance between one focal taxon and another, relative to a conspecific, and divergence is conversely characterised by increasing distance. In each comparative analysis, Mann Whitney tests for equal medians tested whether conspecifics differed significantly in their distance from the focal co-mimic of the other species (after Shapiro Wilk's tests indicated that some subspecies values were non-normally distributed).

Phylogenetic analyses

Neighbour joining trees for subspecies were constructed based on phenotypic distances (e.g. Table S4) using MatLab scripts. In order to visualise the phylogenetic agreement versus conflict among different axes of the phenotypic spatial embedding (Table S3), neighbour joining trees were constructed based on repeated sub-sampling of the axes (sampling either all 64 axes with 1 replicate, or subsamples of 8 or 32 axes with 100 replicates). Consensus networks were constructed to visualise all splits (taxon partitions) implied among sets of trees using the program SplitsTree 4. To test the phylogenetic informativeness of the phenotypic distances against independent data sources, sets of neighbour joining phenotypic trees (of either all subspecies, *H. erato* only, or *H. melpomene* only) were compared against phylogenies reconstructed from published gene sequences (*27*) as well as random tree topologies. The subspecies coverage and individual samples sizes of our analysis exceed those typically used in current gene sequencing studies. However, published phylogenies (*26*) based on multi-locus gene sequences (*27*) were available that included 25 of the 38 studied subspecies (13 *H. erato*, 12 *H. melpomene*), from gene loci (sampled from a different, smaller set of 127 butterfly individuals) which were either associated with *Heliconius* wing colour pattern (*27*) (*optix*, *bves*, *kinesin*, *GPCR*, *VanGogh*) or were neutral markers (*mt COI-COII*, *SUMO*, *Suz12*, *2654* and *CAT*). For each gene set (pattern versus neutral loci), 100 trees were sampled from the output of previously published Bayesian phylogenetic analyses (sampling the MCMC chain

after burn-in) (*26*). One thousand equiprobable, random tree topologies were generated for each taxon set using the program Mesquite. Pairwise distances between trees from the different sets were calculated using the Robinson Foulds (symmetric distance) metric in the program PAUP. Robinson Foulds distances across different tree sets were statistically compared using nonparametric Mann-Whitney tests in the program PAST (after Shapiro-Wilk's tests indicated non-normal distributions). Tree-space visualisations were produced, based on the Robinson Foulds distance, using the TSV package in Mesquite.

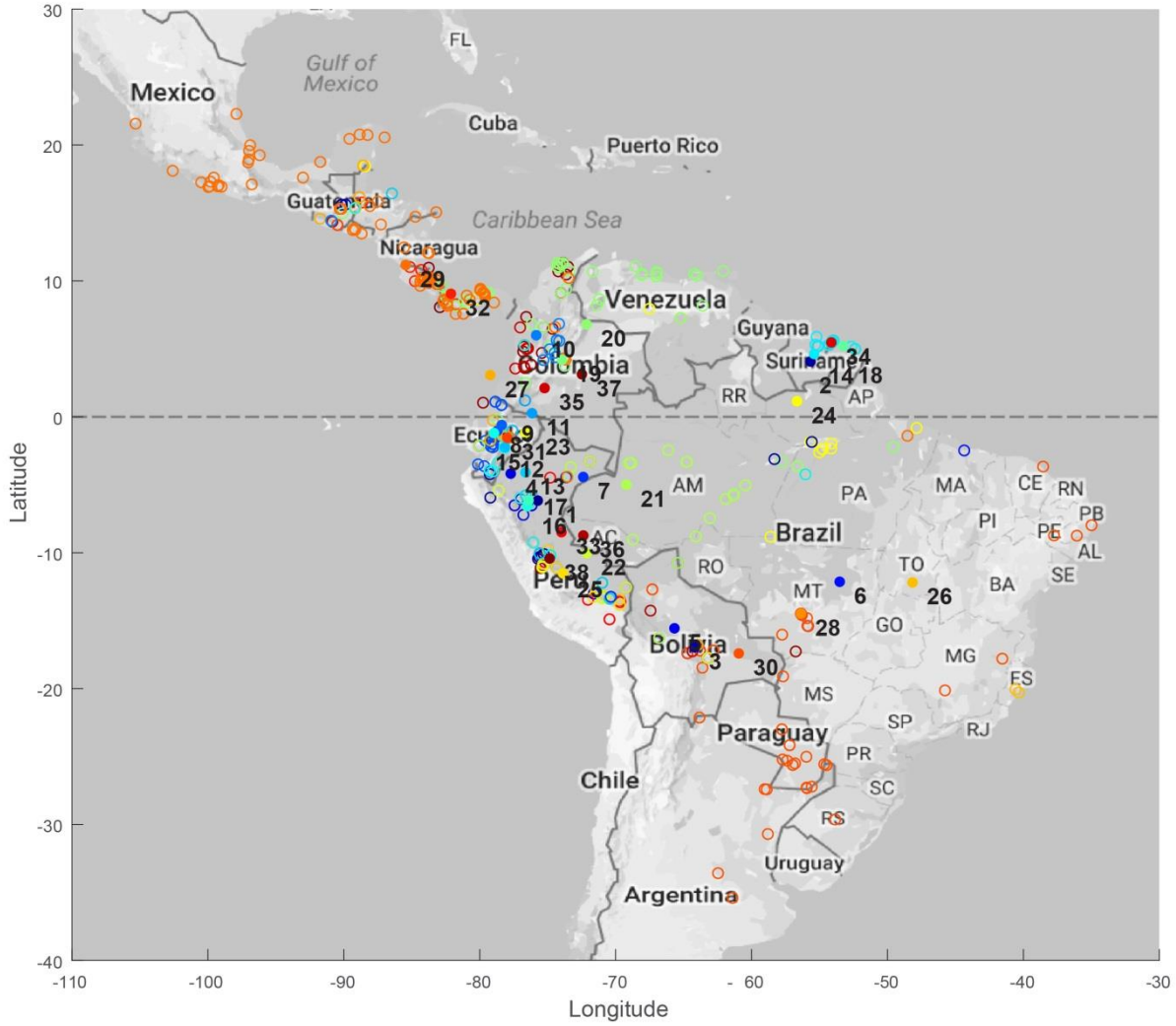


Fig. S2. Geographic localities for sampled butterfly specimens from the polymorphic mimicry complex of *H. erato* and *H. melpomene*. Open circles indicate approximate capture localities for historical butterfly specimens held in the Natural History Museum (NHM), London (Table S2). Based on specimen labels, capture localities were identifiable for 94% of 1234 butterflies in the dataset. Filled circles show the mean location for each subspecies. Numbers and circle colours indicate the subspecies number (Table S7).

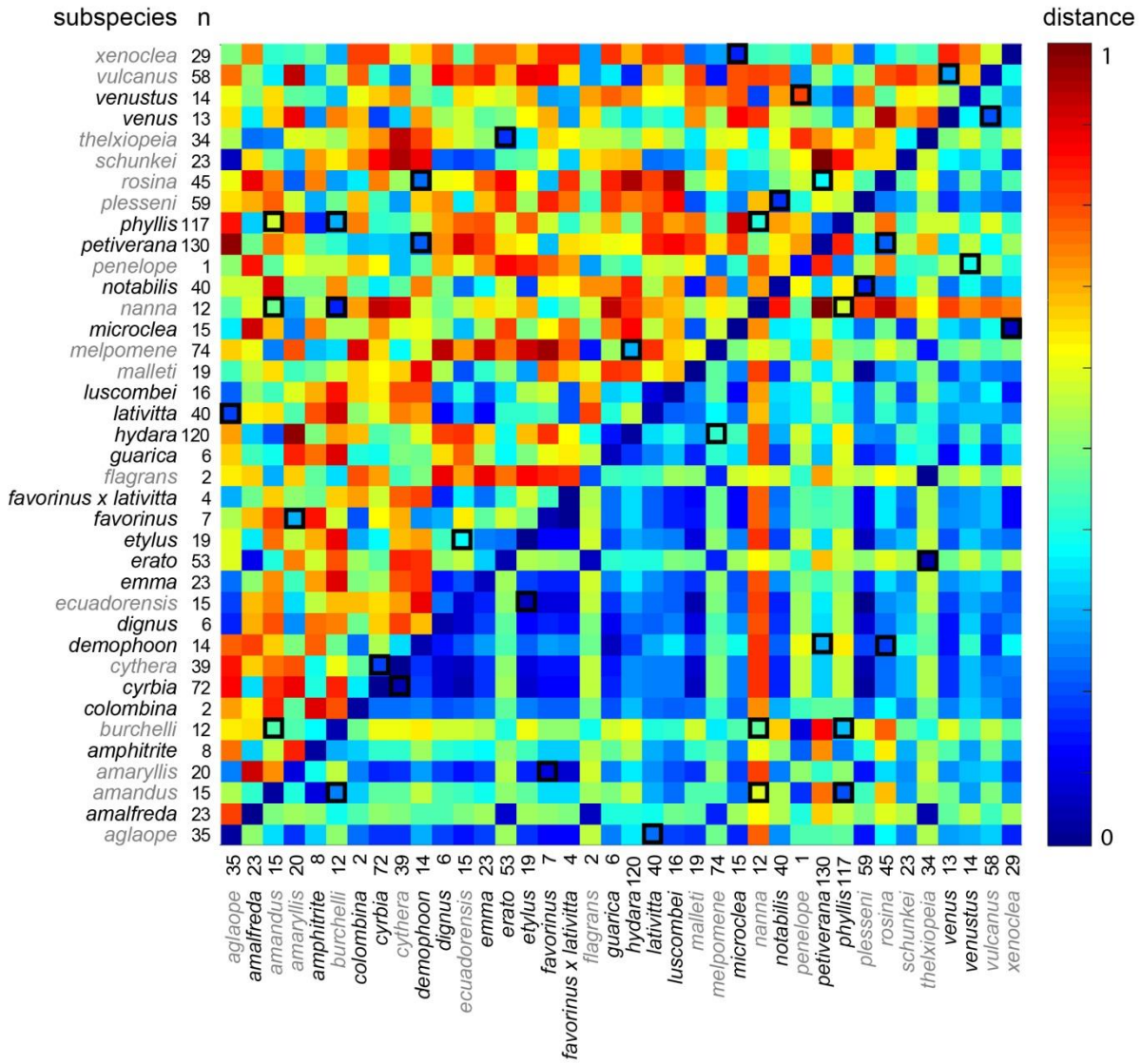
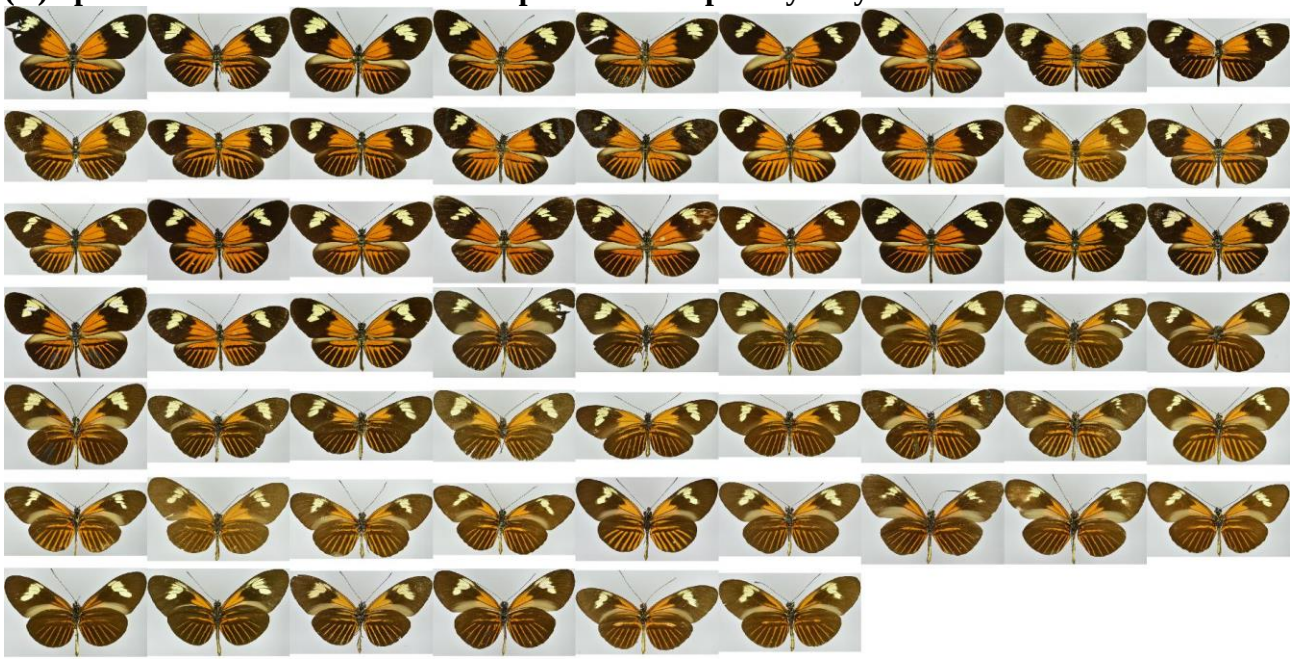


Fig. S3. Heatmap showing mean pairwise phenotypic and geographic distances between 38 subspecies of *H. erato* (black labels) and *H. melpomene* (gray labels). Upper matrix (including diagonal) shows mean pairwise distances for subspecies calculated from the 64-dimensional phenotypic embedding generated using a deep convolutional triplet network across 2468 butterfly images. Lower matrix shows mean pairwise geographic distance between butterfly specimens. Key shows correspondence between heat-map colours and distance, rescaled to vary between zero and one from original values (upper, squared Euclidean phenotypic distance and lower, Euclidean geographic distance, Tables S5-S6), from blue (most similar) to red (least similar). Black borders on squares indicate traditionally hypothesised co-mimics (*12*). Numbers adjacent to subspecies names show the number of butterfly specimens in the image database.

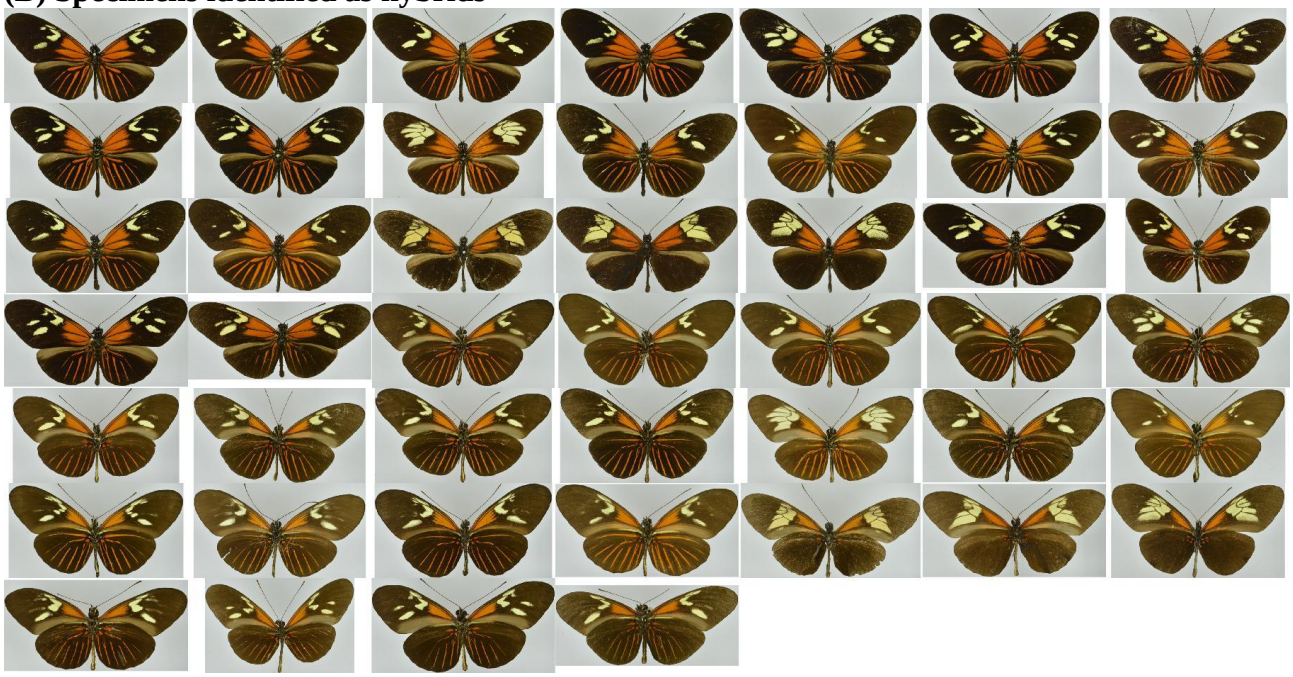
1 *Heliconius melpomene aglaope*

(A) Specimens identified as valid subspecies or accepted synonym

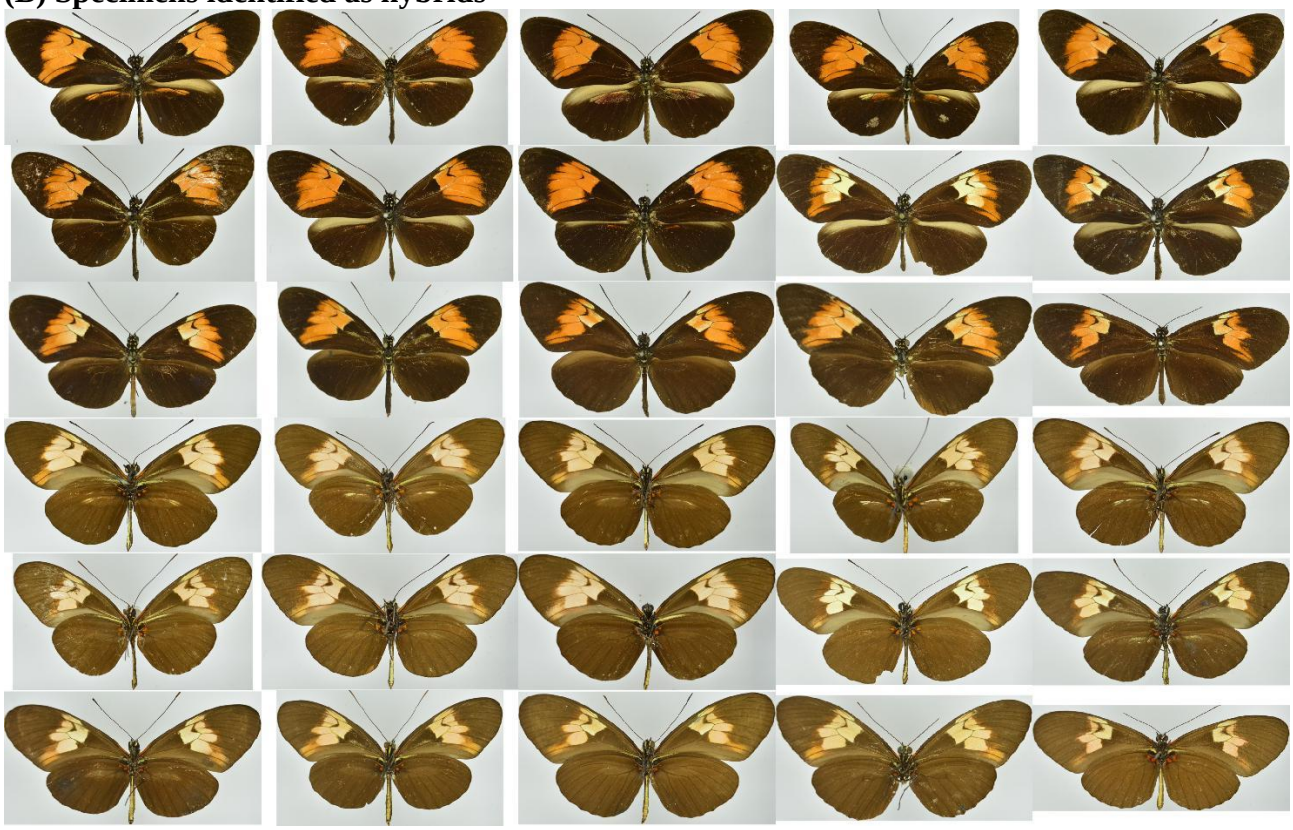


(B) Specimens identified as hybrids

2 *Heliconius erato amalfreda*
(B) Specimens identified as hybrids



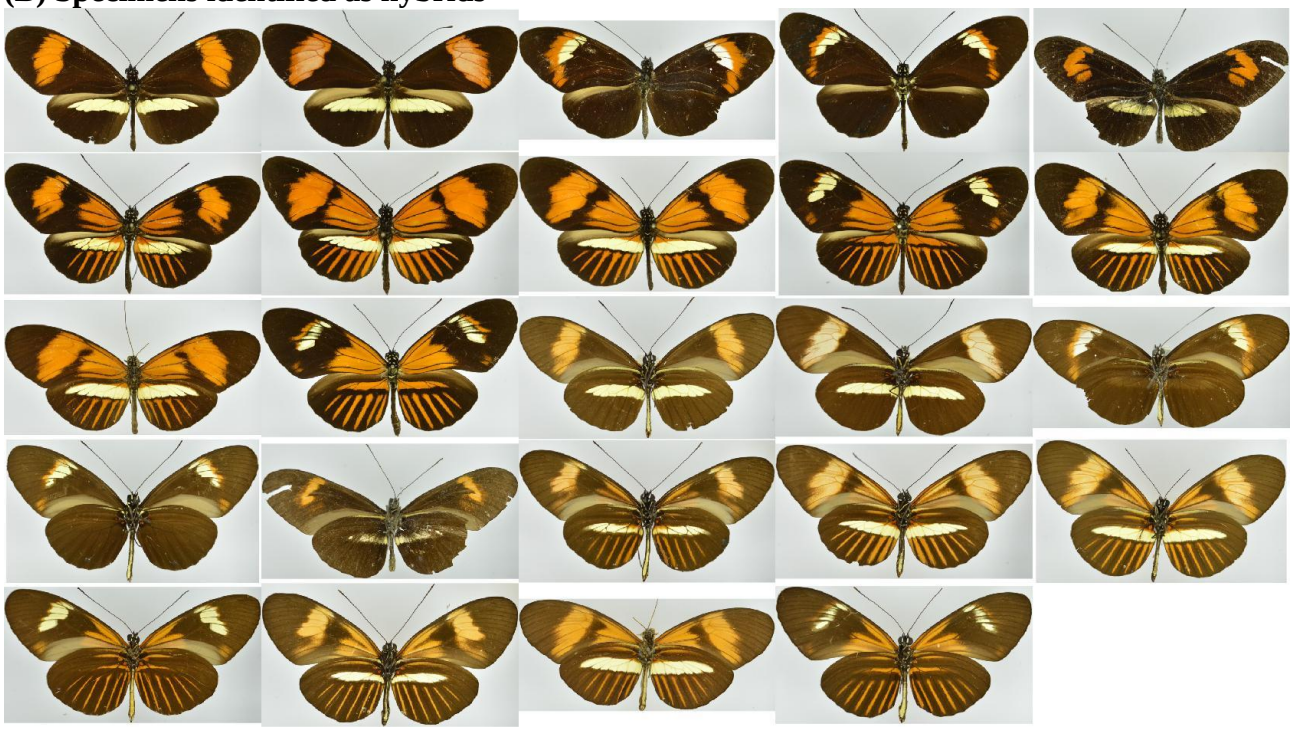
3 *Heliconius melpomene amandus*
(B) Specimens identified as hybrids



4 *Heliconius melpomene amaryllis*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids



5 *Heliconius erato amphitrite*
(B) Specimens identified as hybrids

6 *Heliconius melpomene burchelli*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

7 Heliconius erato colombina

(A) Specimens identified as valid subspecies or accepted synonym

8 *Heliconius erato cyrbia*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

9 *Heliconius melpomene cythera*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

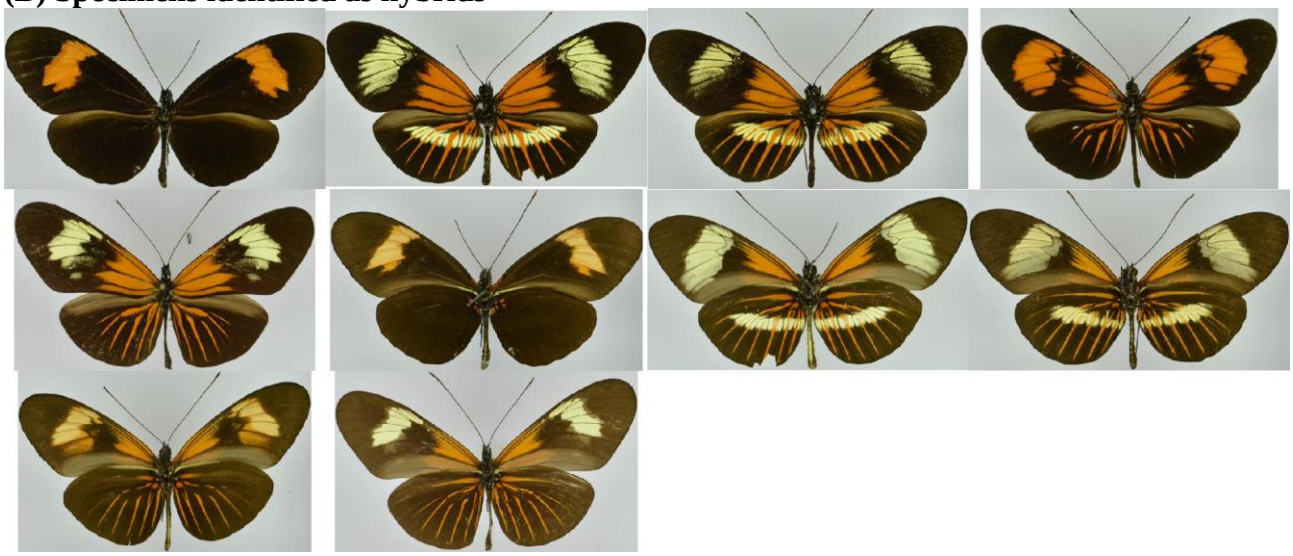
10 *Heliconius erato demophoon*

(A) Specimens identified as valid subspecies or accepted synonym

11 *Heliconius erato dignus*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids



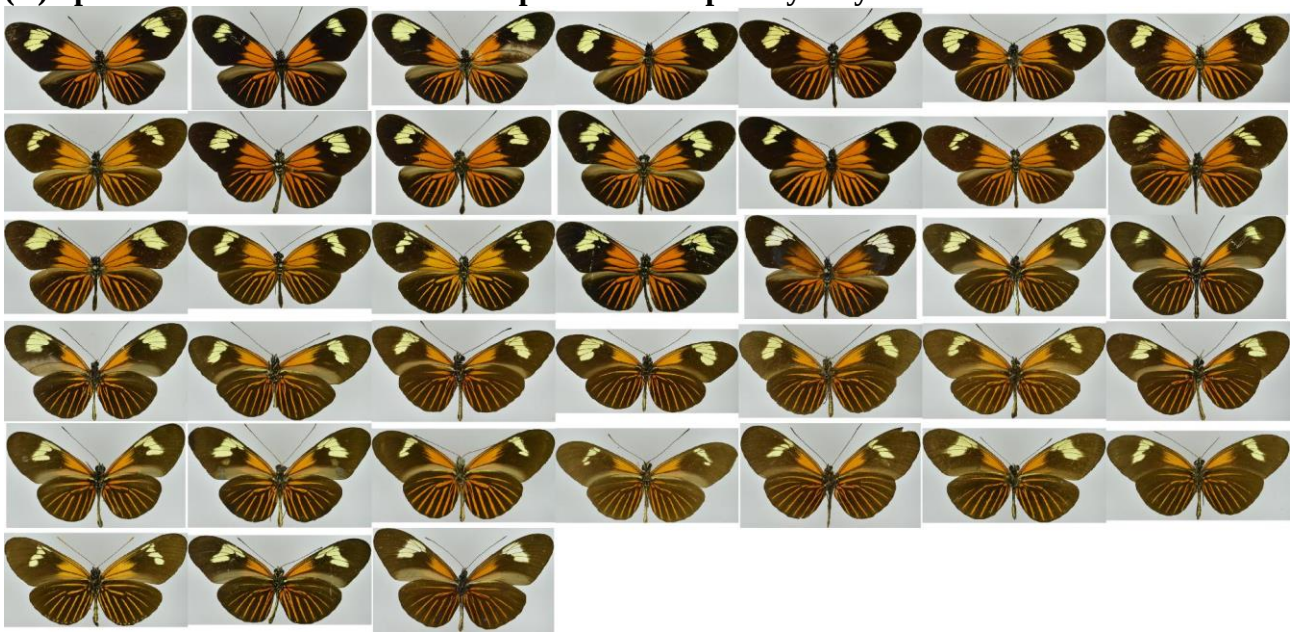
12 *Heliconius melpomene ecuadorensis*

(A) Specimens identified as valid subspecies or accepted synonym

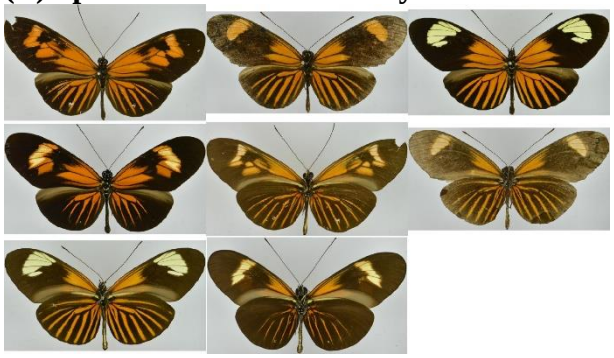
(B) Specimens identified as hybrids

13 *Heliconius erato emma*

(A) Specimens identified as valid subspecies or accepted synonym

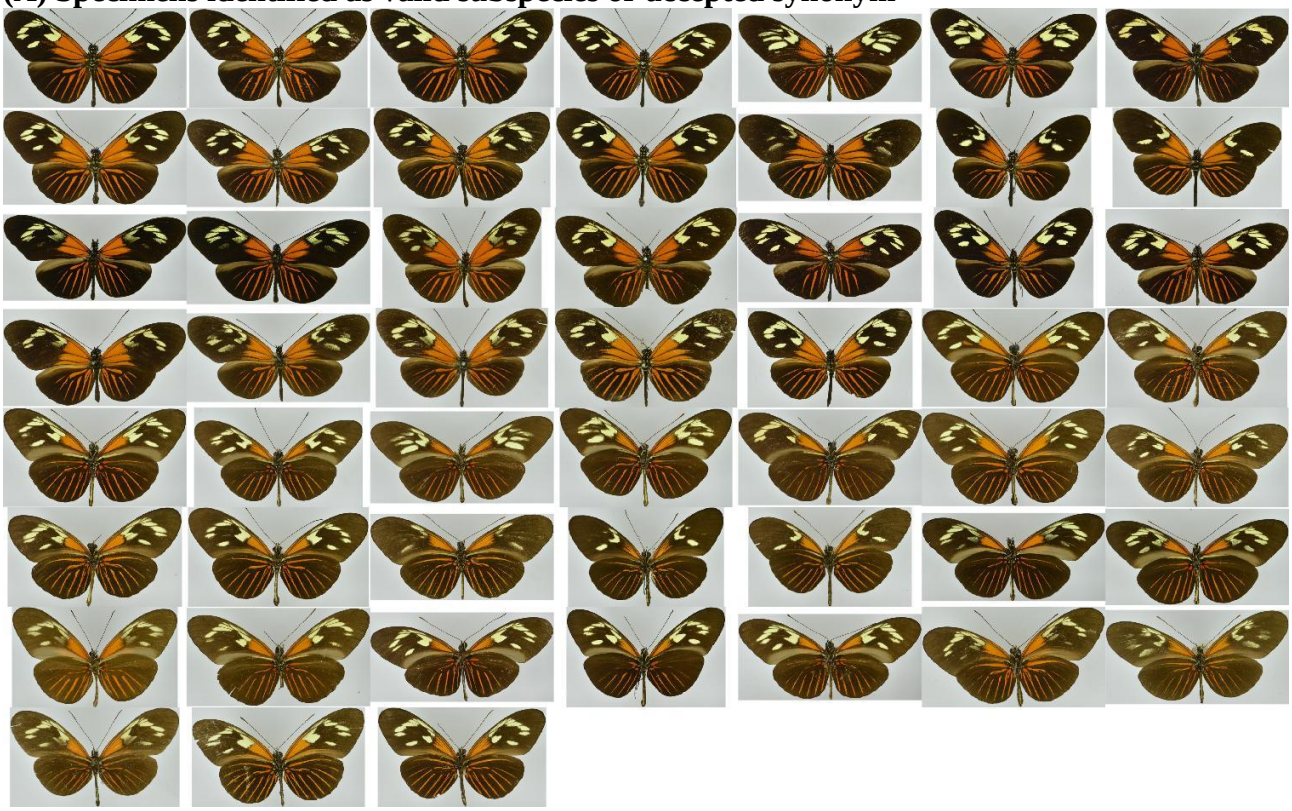


(B) Specimens identified as hybrids



14 *Heliconius erato erato*

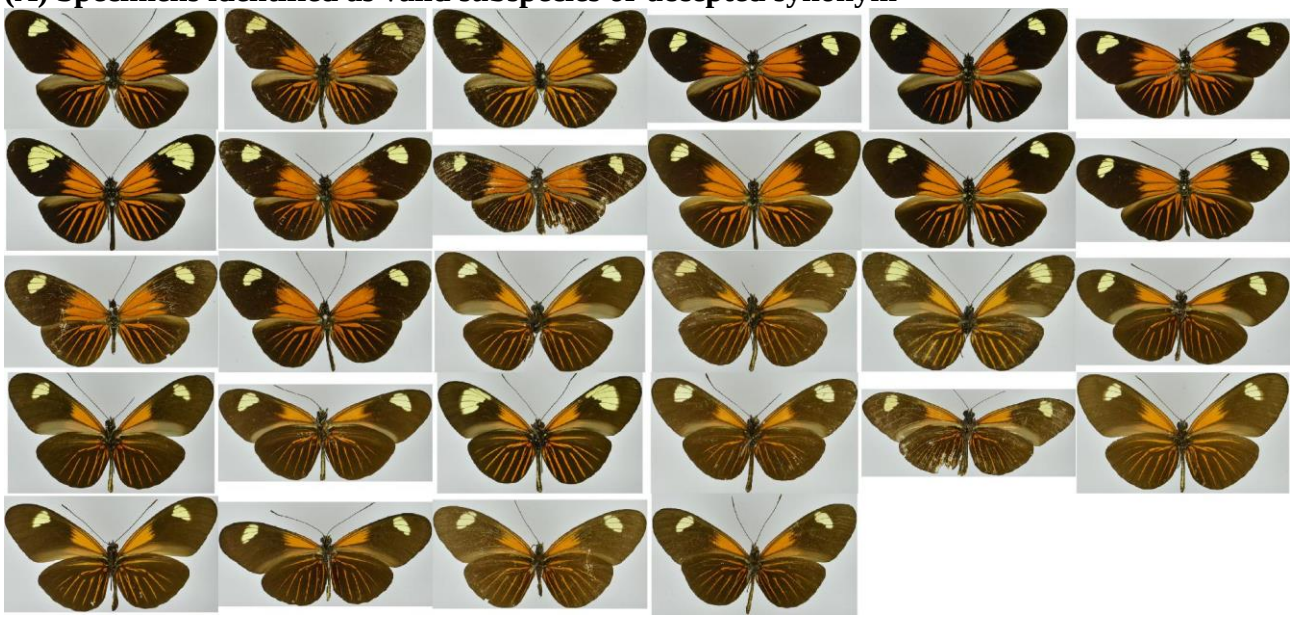
(A) Specimens identified as valid subspecies or accepted synonym



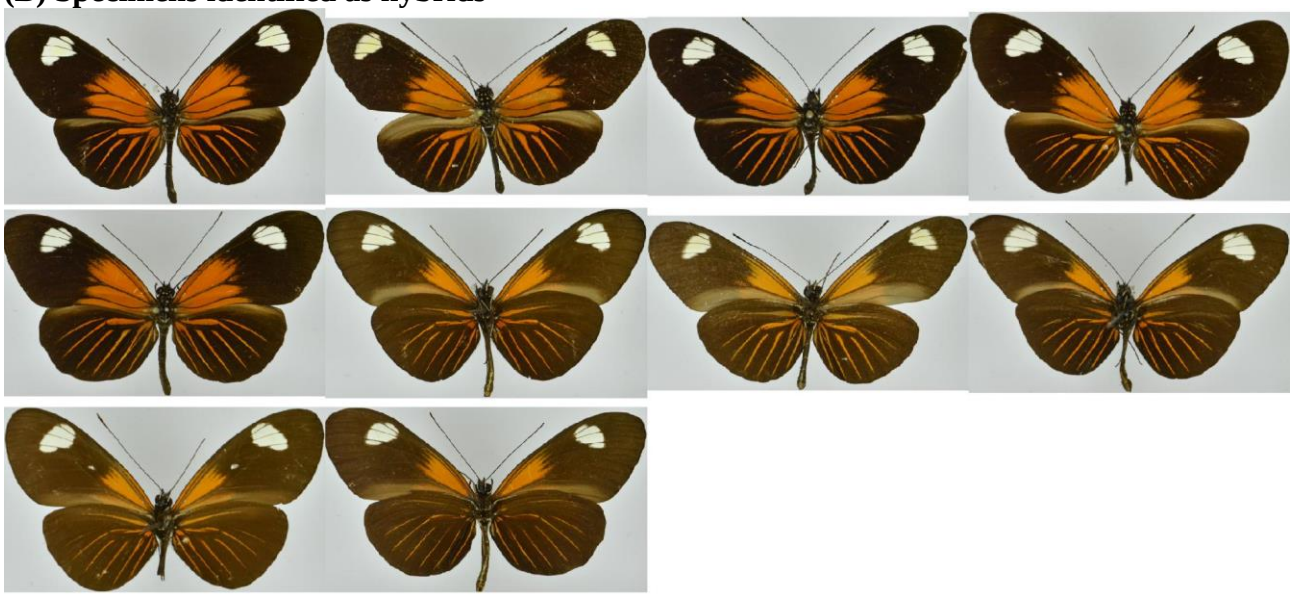
(B) Specimens identified as hybrids

15 *Heliconius erato etylus*

(A) Specimens identified as valid subspecies or accepted synonym



(B) Specimens identified as hybrids



16 *Heliconius erato favorinus*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

17 *Heliconius erato favorinus x lativitta*
(B) Specimens identified as hybrids

18 *Heliconius melpomene flagrans*

(A) Specimens identified as valid subspecies or accepted synonym

19 *Heliconius erato guarica*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

20 *Heliconius erato hydara*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

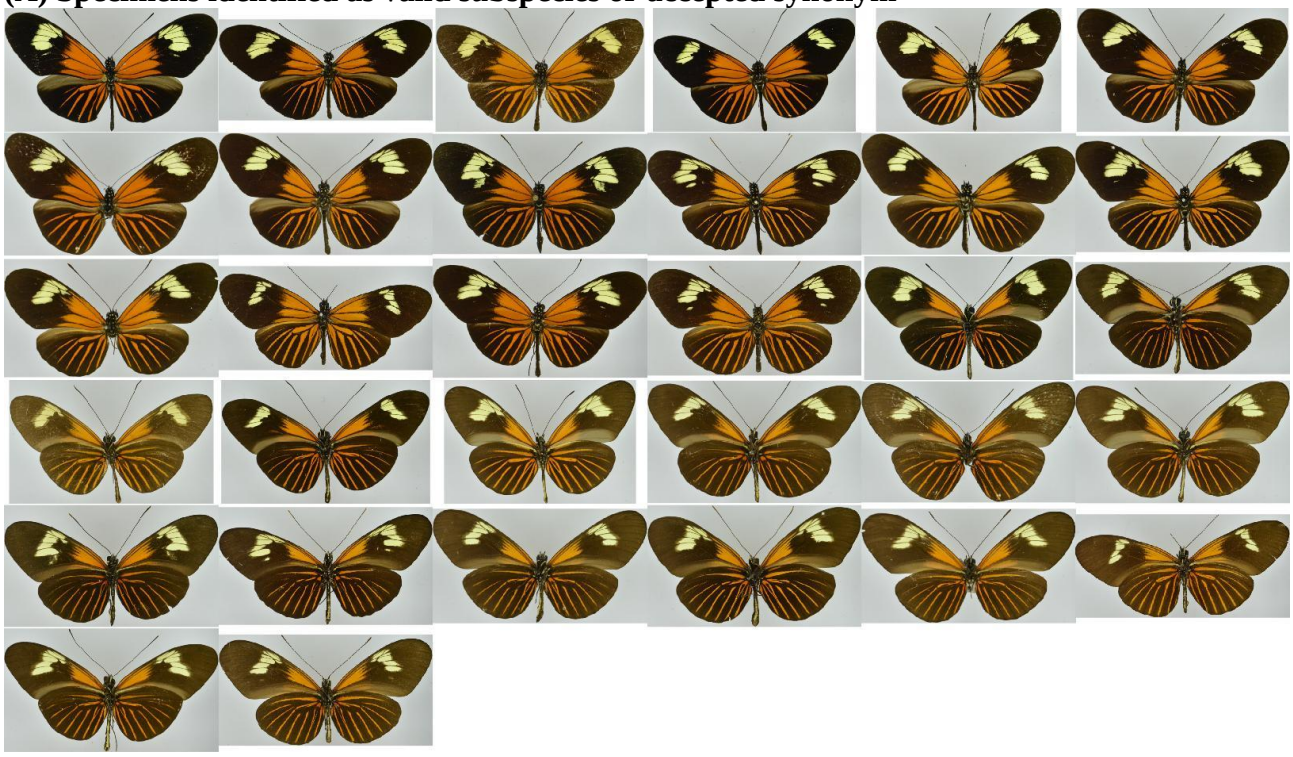
21 *Heliconius erato lativitta*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

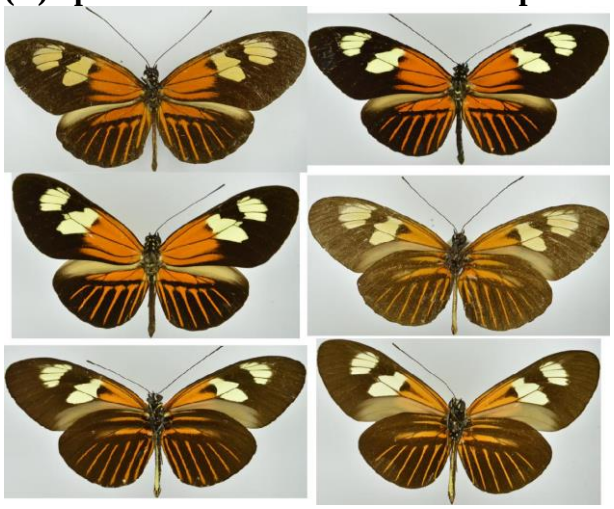
22 *Heliconius erato luscombei*

(A) Specimens identified as valid subspecies or accepted synonym

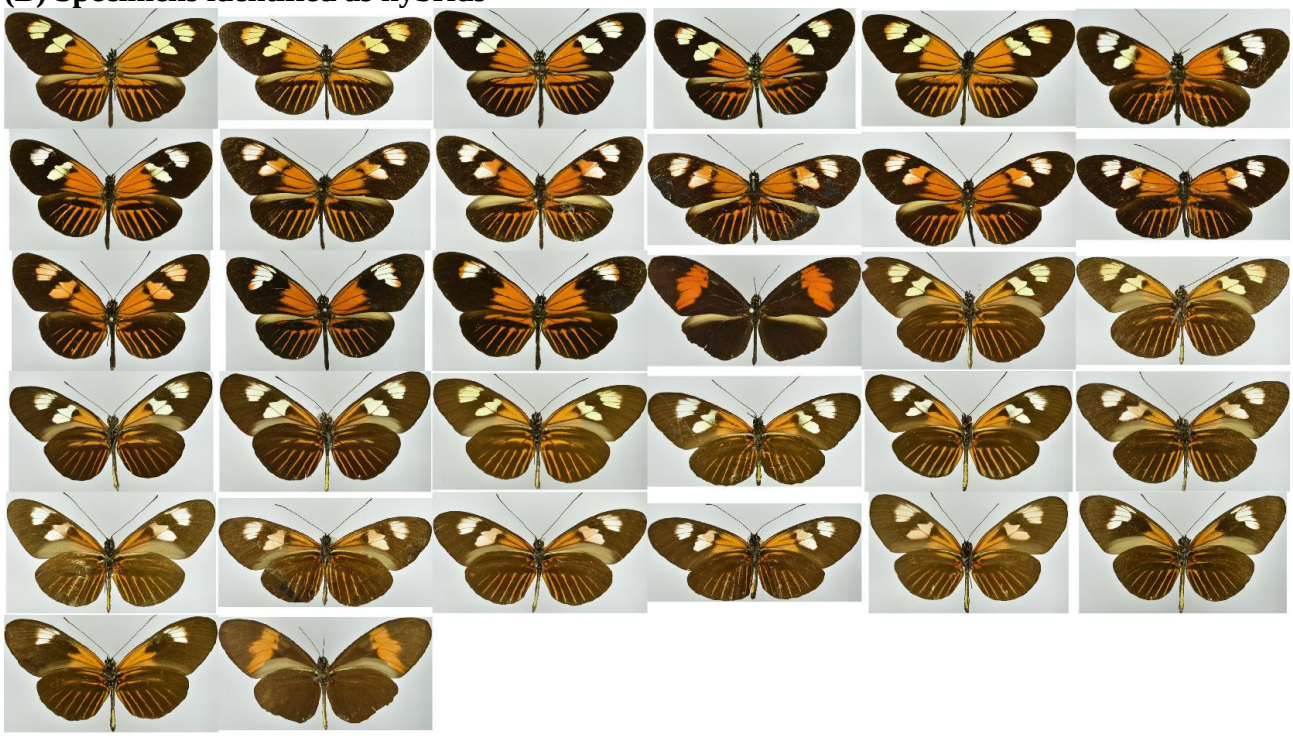


23 *Heliconius melpomene malleti*

(A) Specimens identified as valid subspecies or accepted synonym

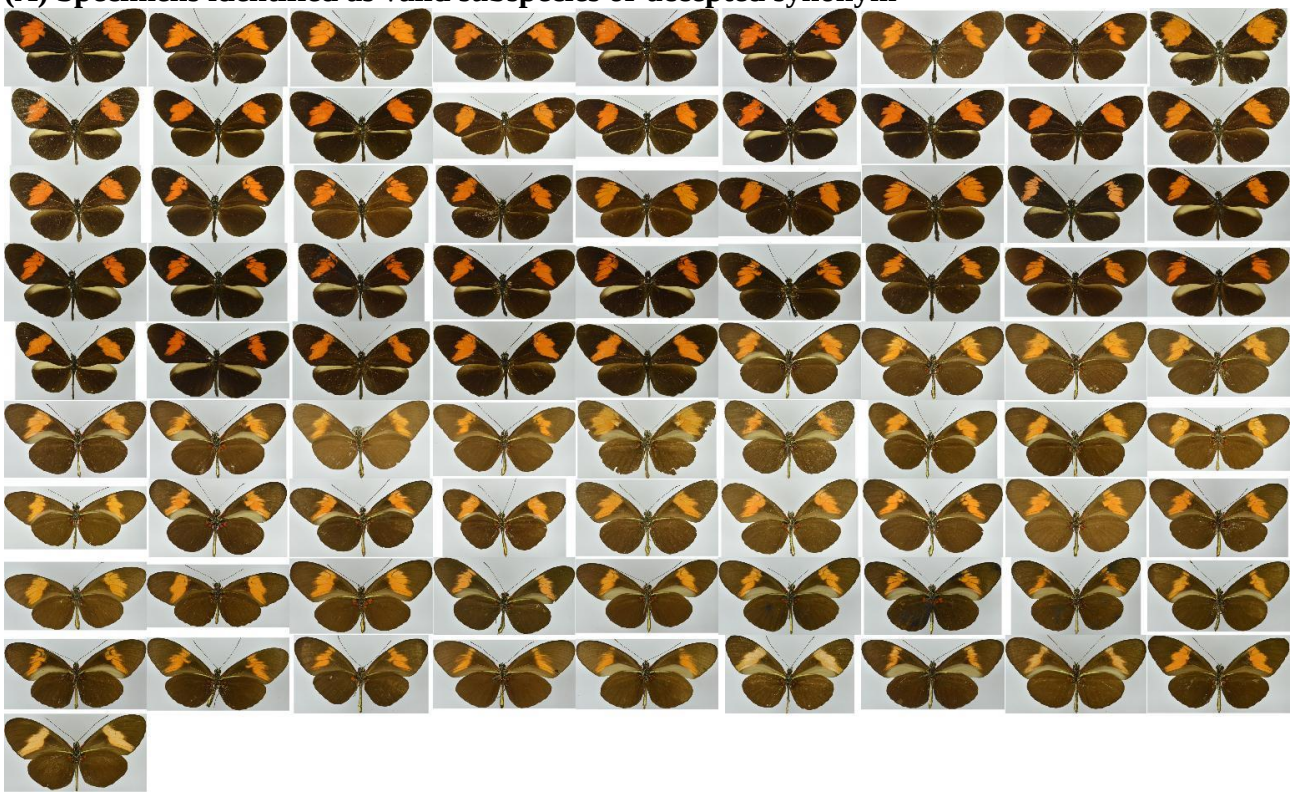


(B) Specimens identified as hybrids

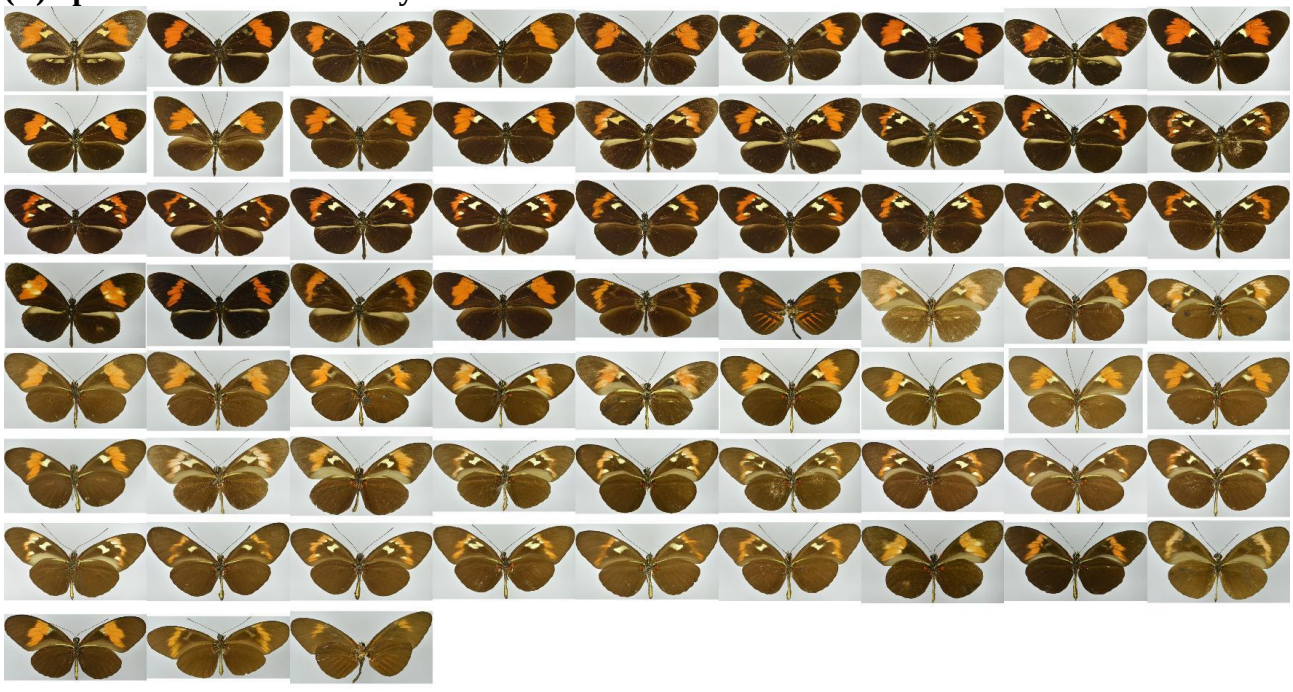


24 *Heliconius melpomene melpomene*

(A) Specimens identified as valid subspecies or accepted synonym

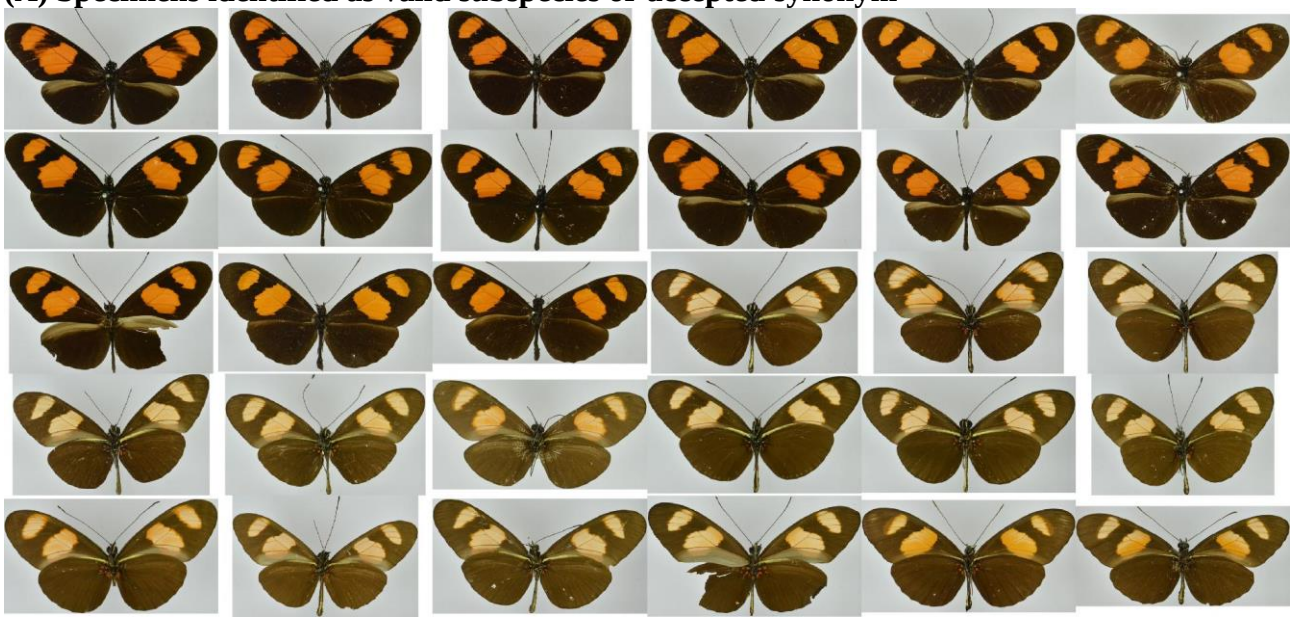


(B) Specimens identified as hybrids



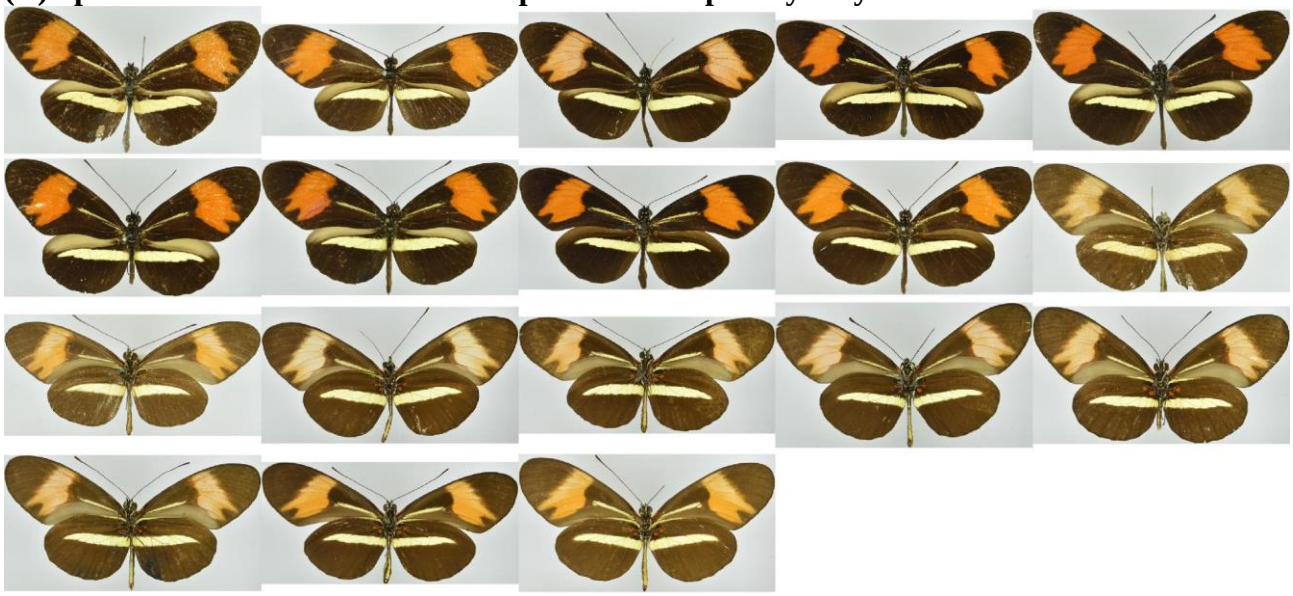
25 *Heliconius erato microclea*

(A) Specimens identified as valid subspecies or accepted synonym



26 *Heliconius melpomene nanna*

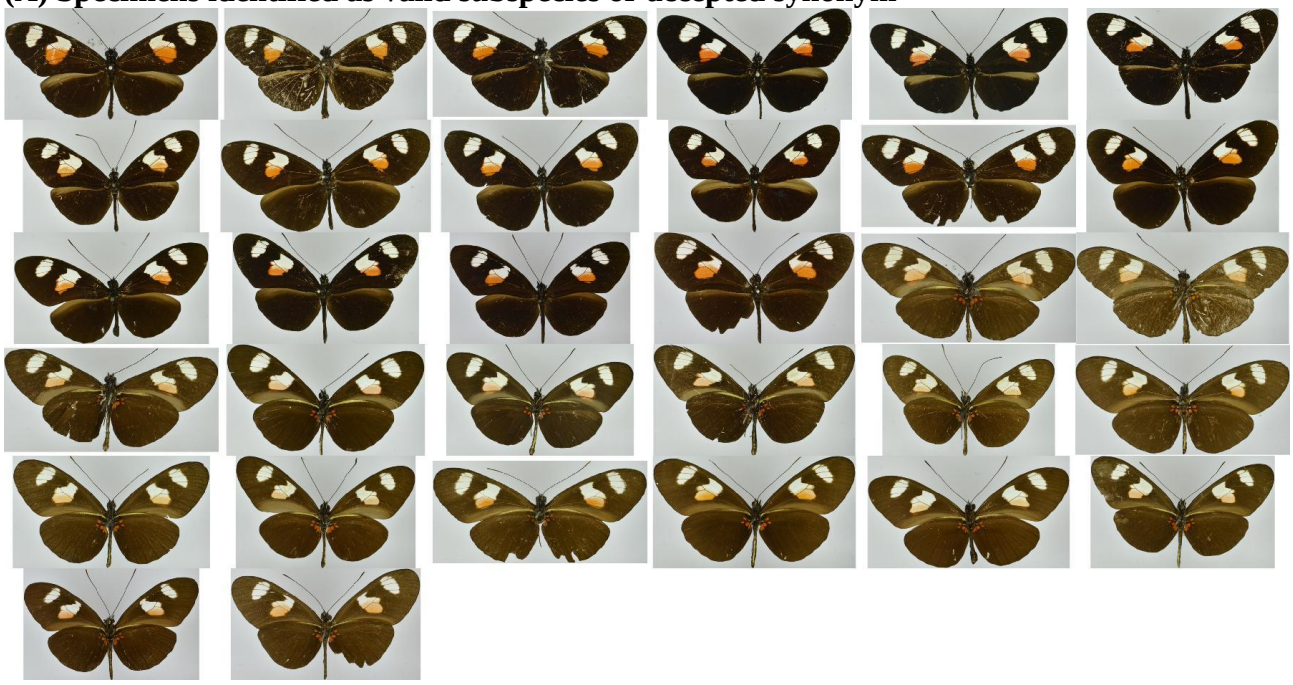
(A) Specimens identified as valid subspecies or accepted synonym



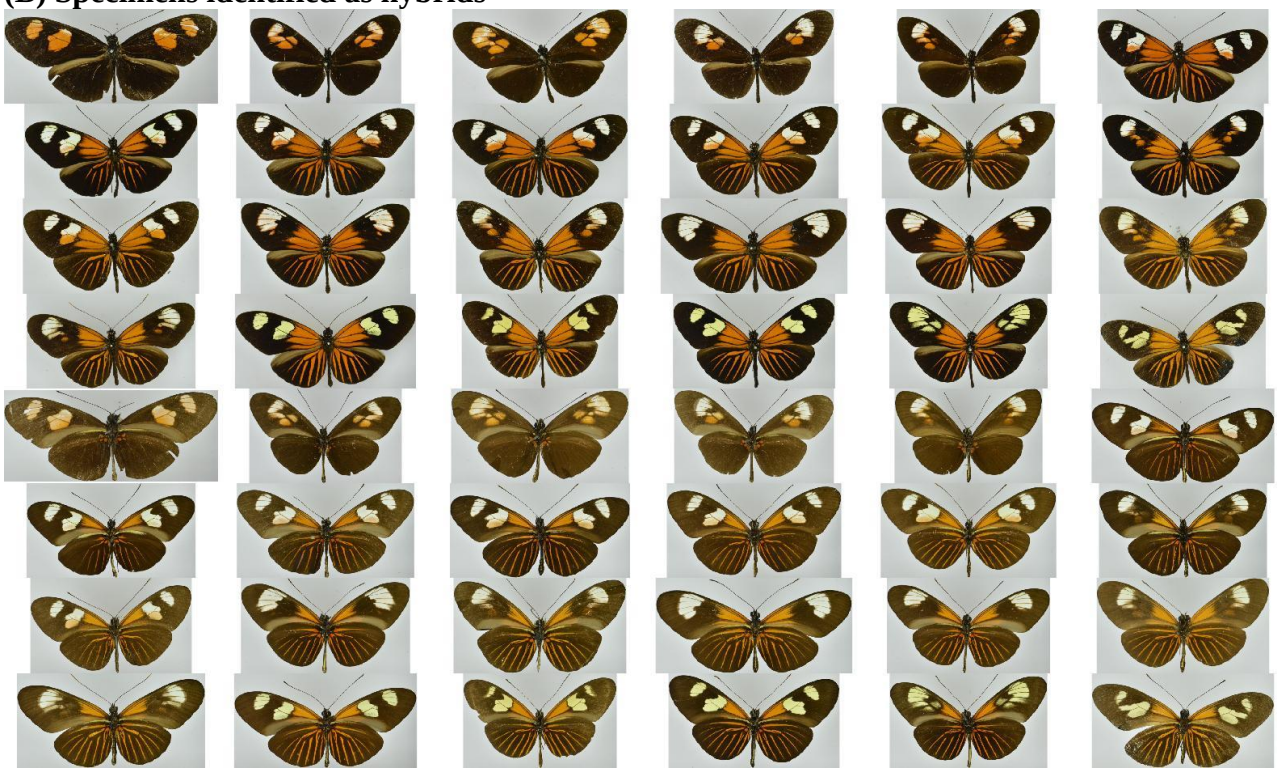
(B) Specimens identified as hybrids

27 *Heliconius erato notabilis*

(A) Specimens identified as valid subspecies or accepted synonym



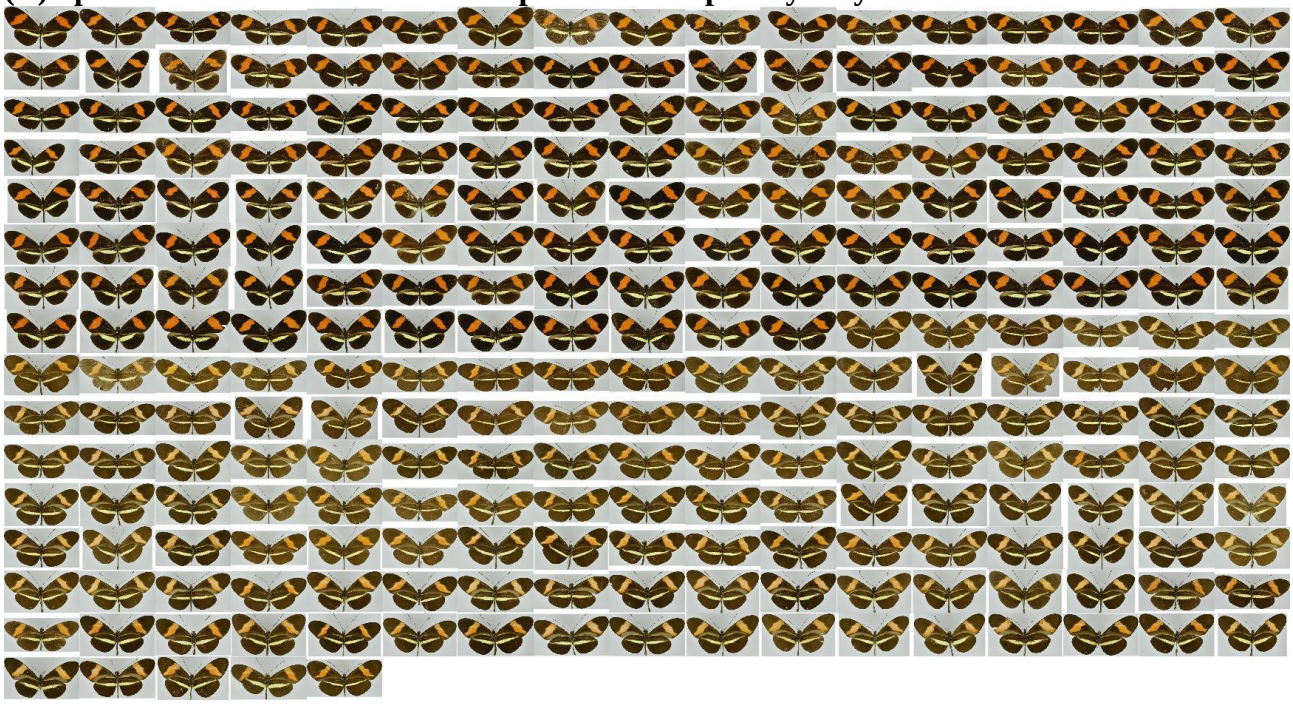
(B) Specimens identified as hybrids



28 *Heliconius melpomene penelope*
(B) Specimens identified as hybrids

29 *Heliconius erato petiverana*

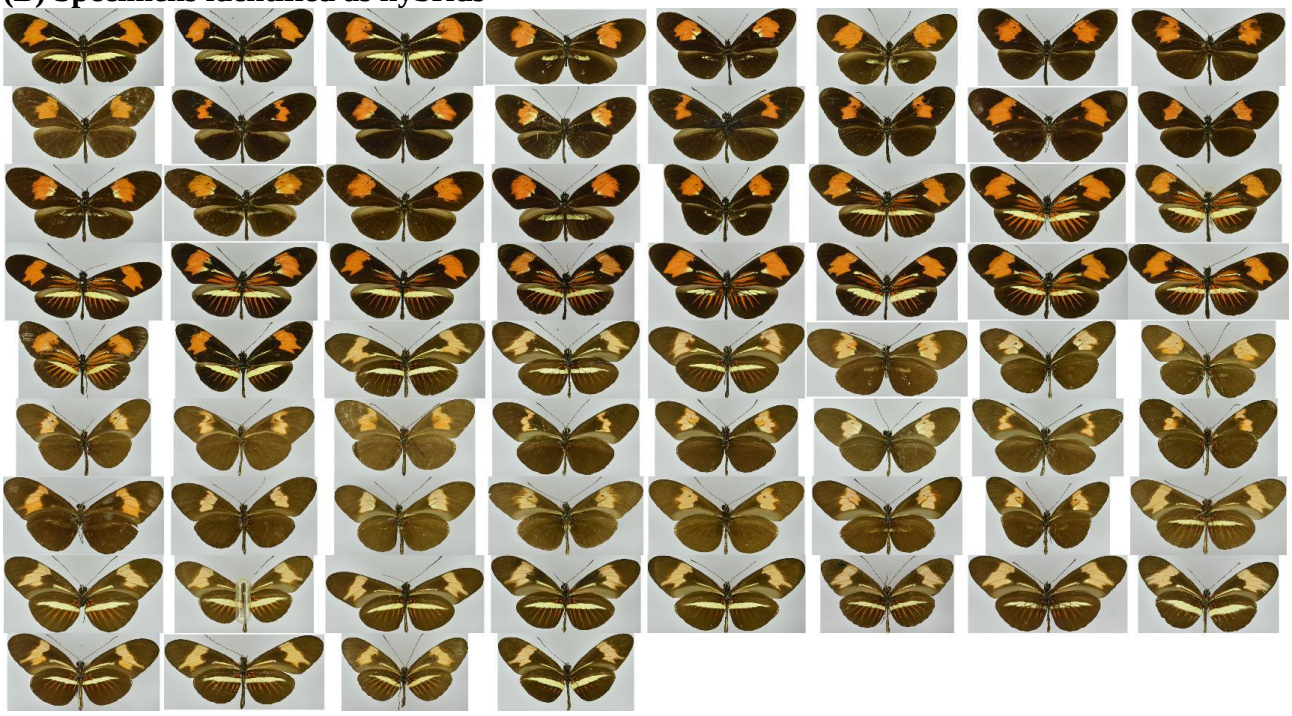
(A) Specimens identified as valid subspecies or accepted synonym



30 *Heliconius erato phyllis*

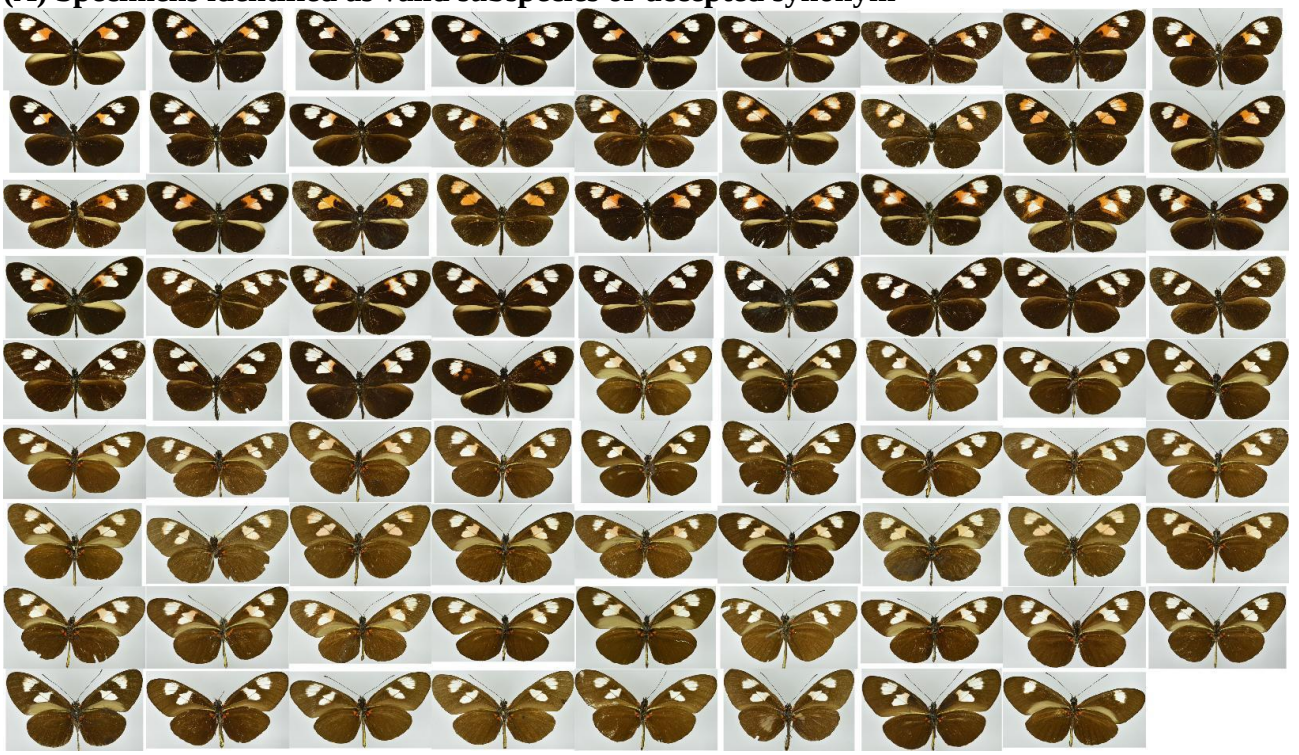
(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

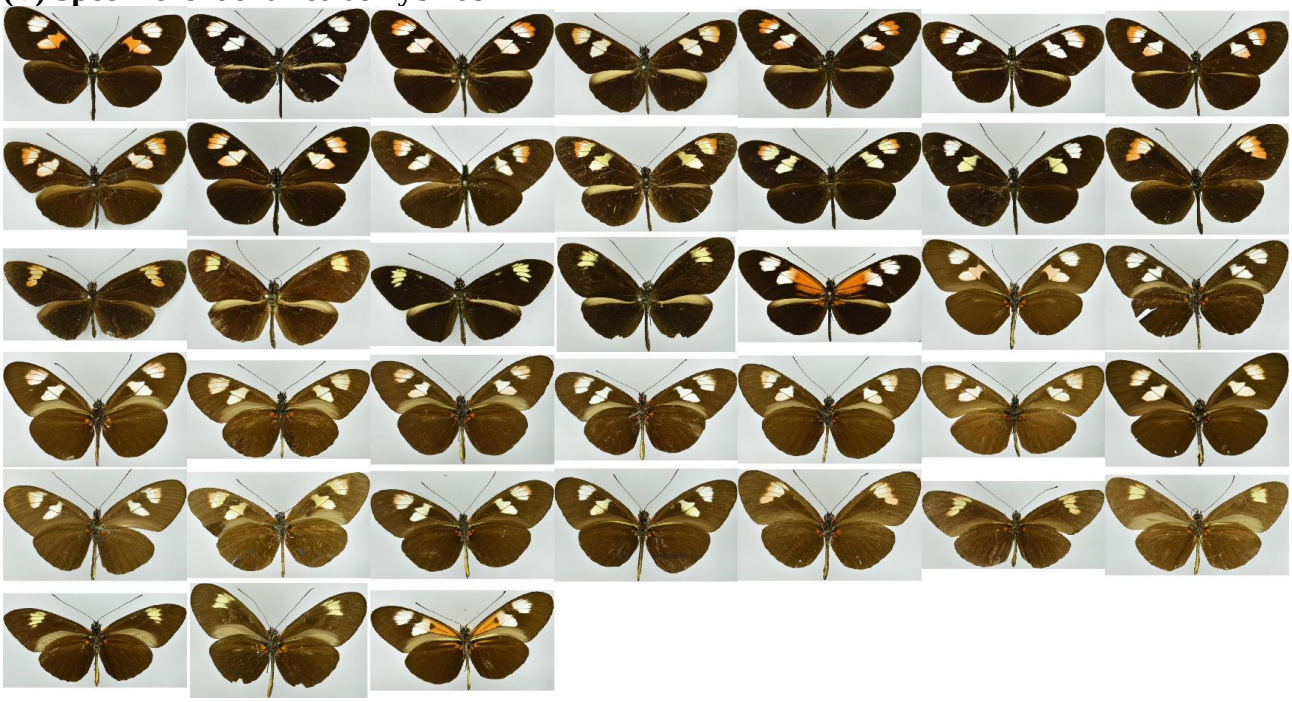


31 *Heliconius melpomene plesseni*

(A) Specimens identified as valid subspecies or accepted synonym

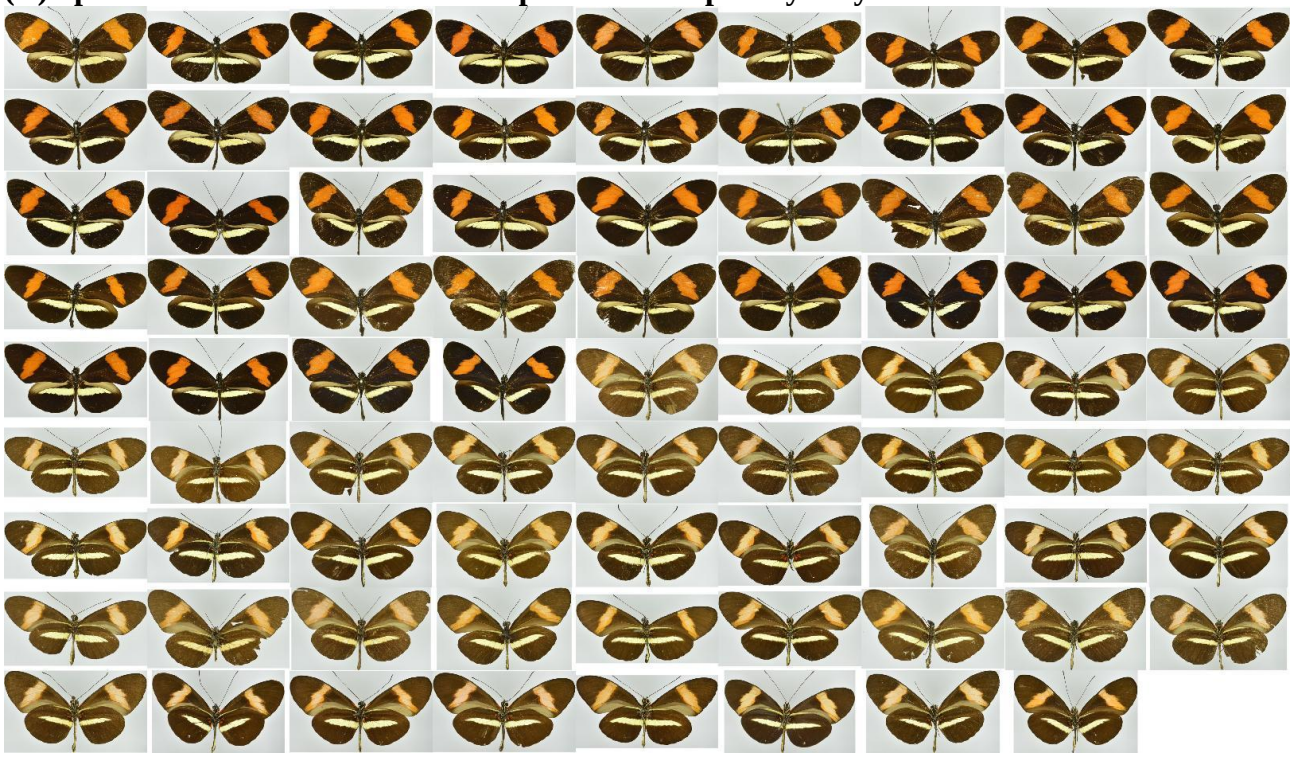


(B) Specimens identified as hybrids

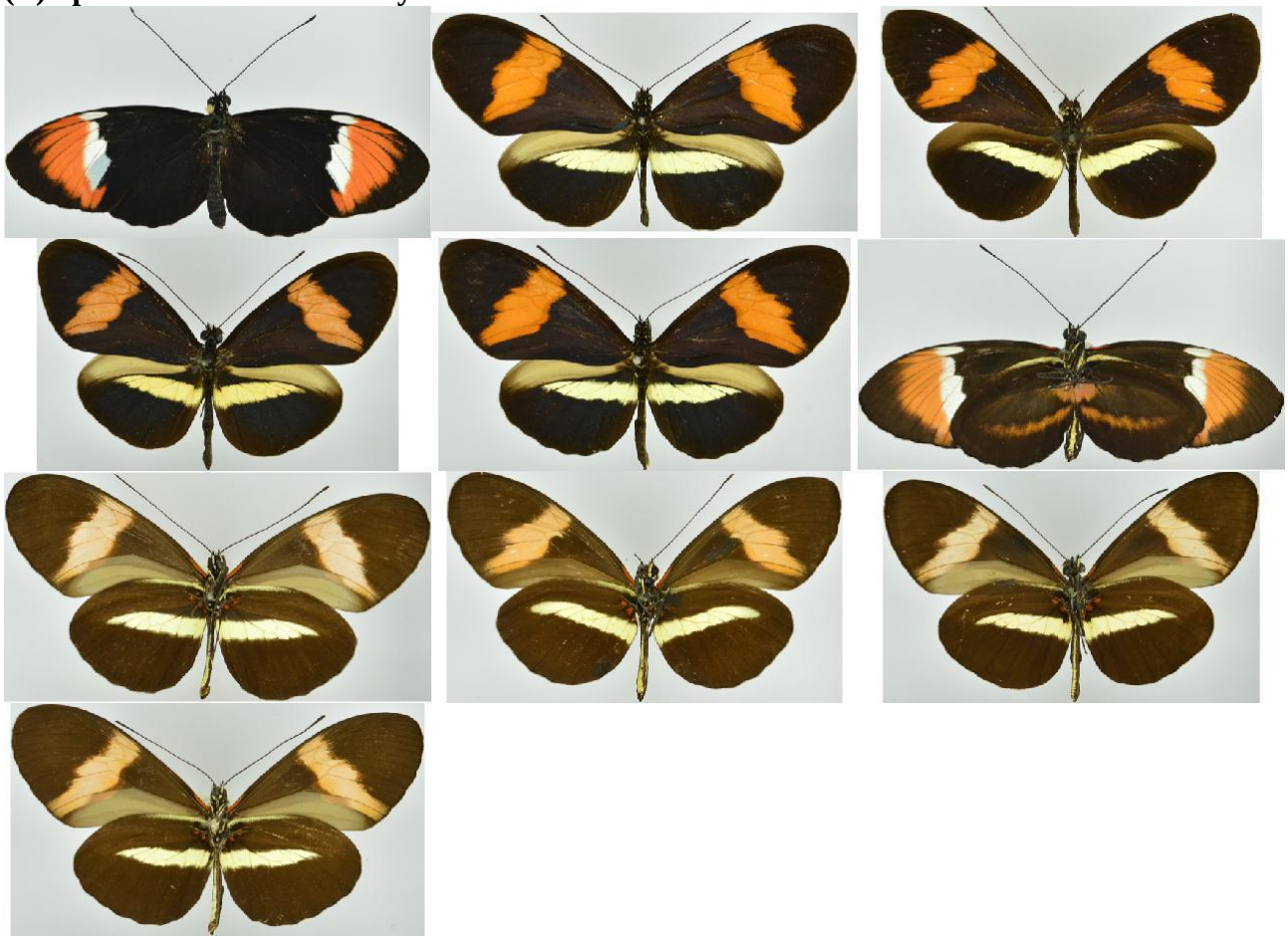


32 *Heliconius melpomene rosina*

(A) Specimens identified as valid subspecies or accepted synonym



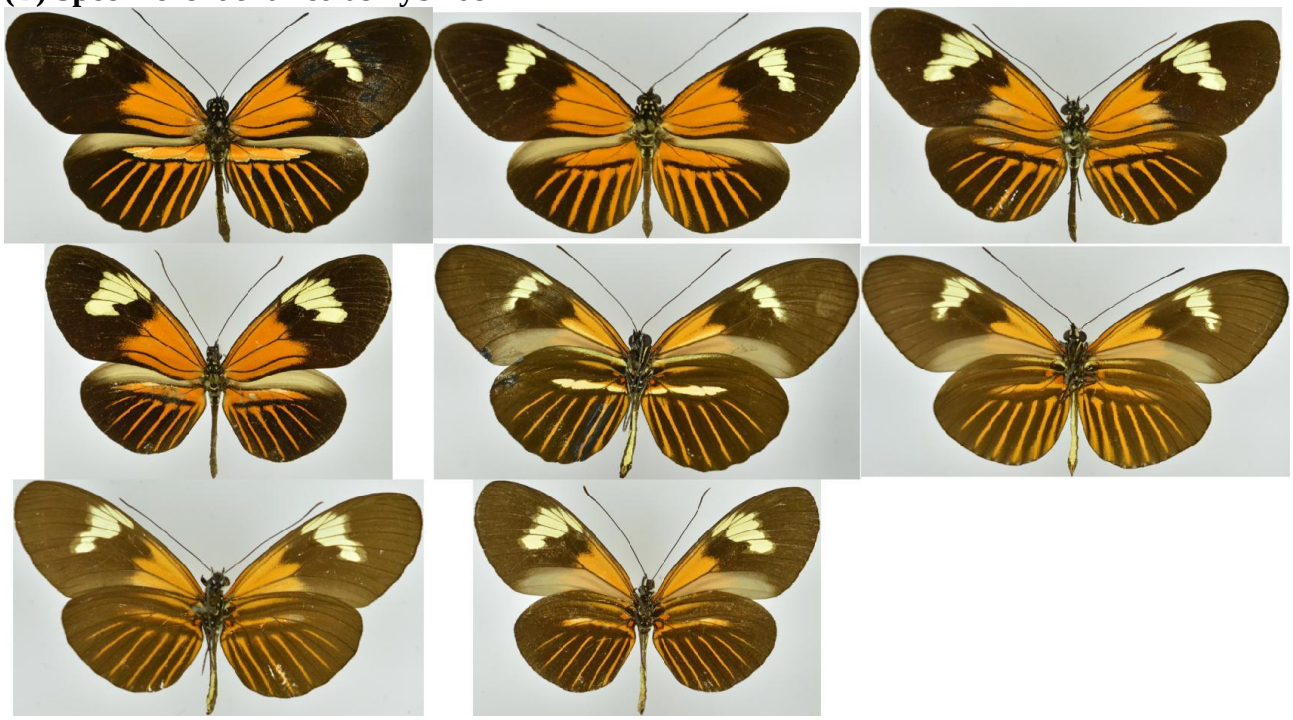
(B) Specimens identified as hybrids



33 *Heliconius melpomene schunkei*

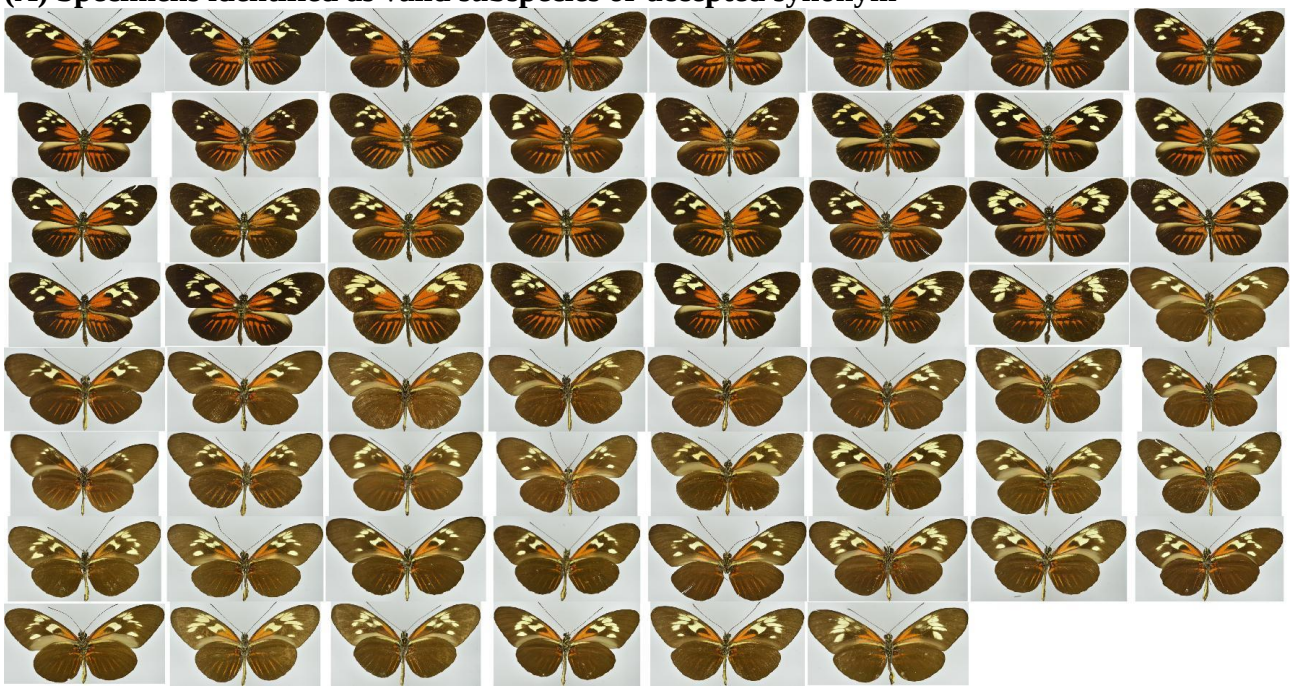
(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids



34 *Heliconius melpomene thelxiopeia*

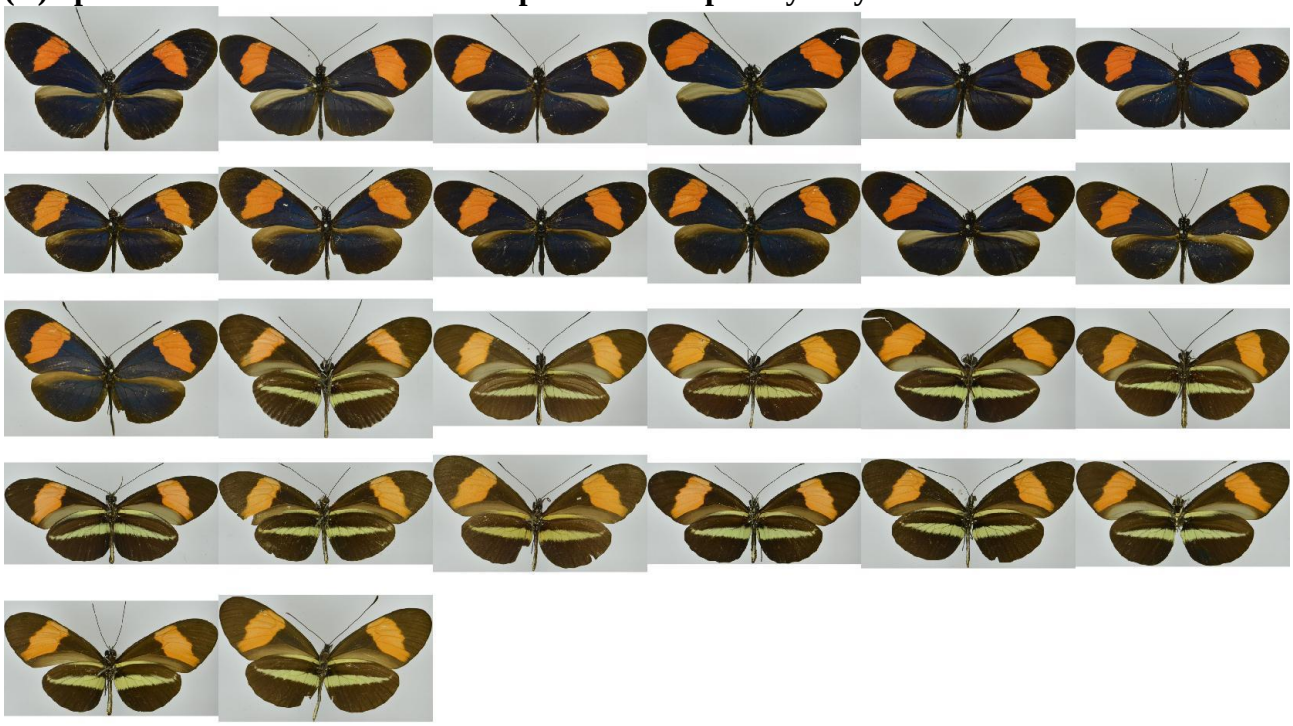
(A) Specimens identified as valid subspecies or accepted synonym



(B) Specimens identified as hybrids

35 *Heliconius erato venus*

(A) Specimens identified as valid subspecies or accepted synonym



36 *Heliconius erato venustus*

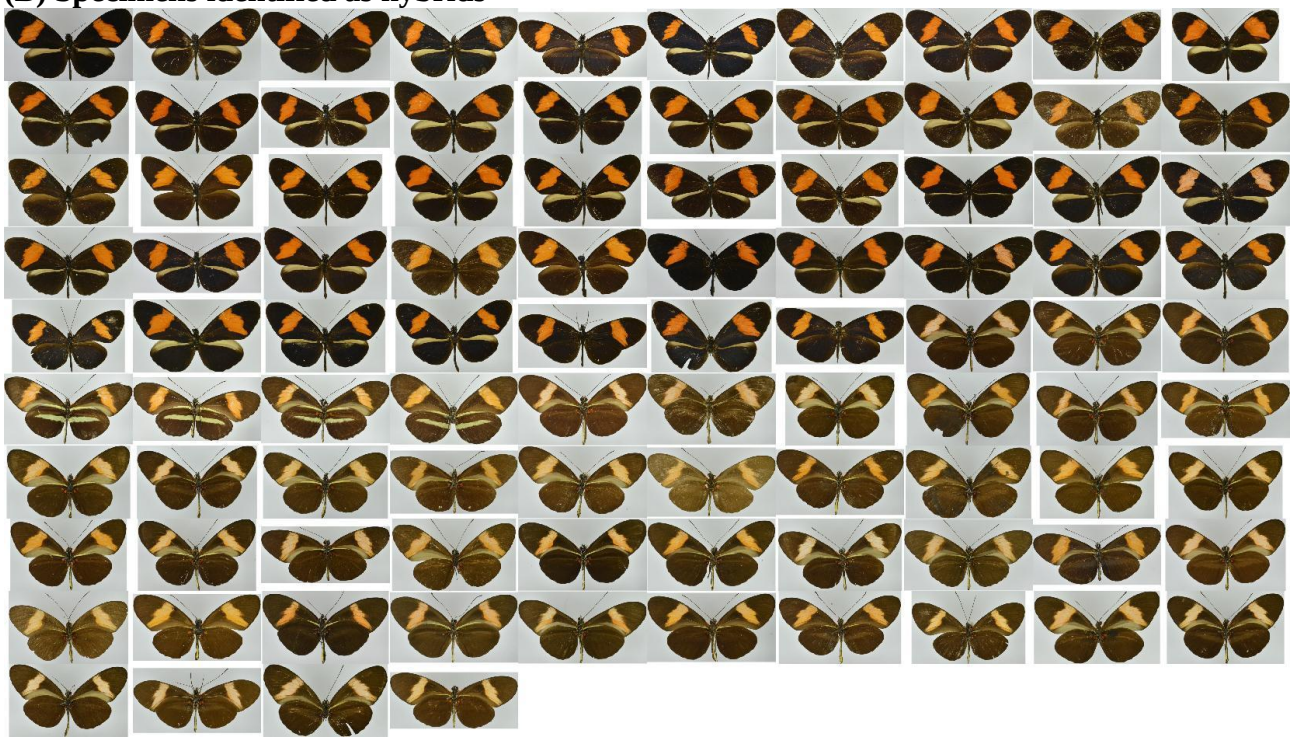
(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

37 *Heliconius melpomene vulcanus*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids



38 *Heliconius melpomene xenoclea*

(A) Specimens identified as valid subspecies or accepted synonym

(B) Specimens identified as hybrids

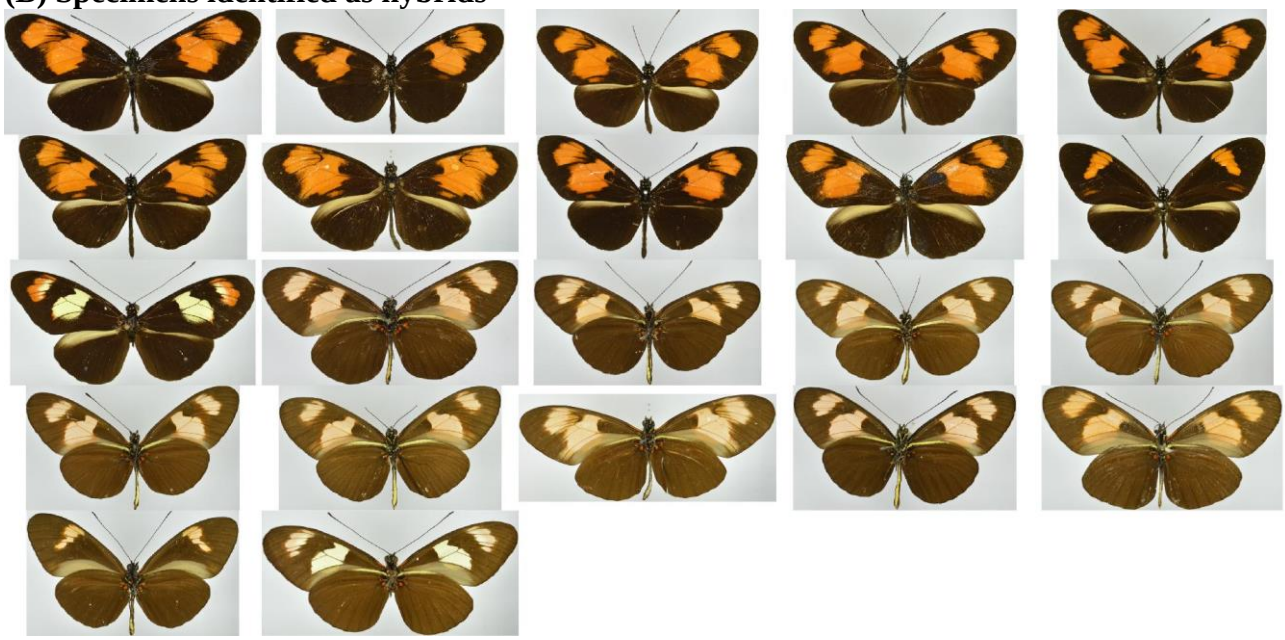


Fig. S4. Collections of specimen photographs used in this study, grouped by subspecies. Photographs show all dorsal views, followed by all ventral views (with the same specimen order, left to right, top to bottom). Subspecies identification followed NHM specimen labels and the taxonomy of Lamas (2014) (*30*). Individual image files used in deep learning, with specimen numbers corresponding Table S2, can be downloaded from the Dryad Data Repository: doi:10.5061/dryad.2hp1978. **(A)** Specimens included in the reduced dataset comprising valid subspecies and synonyms only. **(B)** Specimens excluded from the reduced dataset based on taxonomic labelling as hypothesised hybrids (e.g. based on their phenotype and/or capture locality) and a visual screen. Hybrid status was recorded (Table S2) based on additional taxonomic information from specimen labels, NHM records, the taxonomic checklist of Lamas, 2004 (*30*) and www.butterfliesofamerica.com. Photo Credits: Robyn Crowther and Sophie Ledger, Natural History Museum, London.

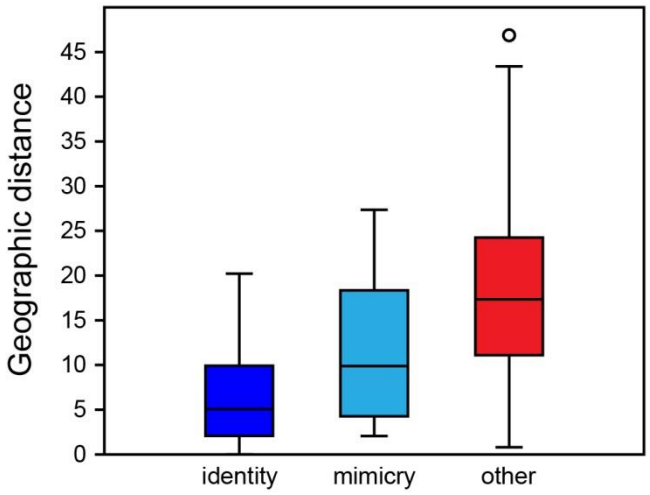


Fig. S5. Average pairwise Euclidean geographic distances between subspecies of *H. erato* and *H. melpomene*. Box plot of mean pairwise geographic distances (Table S6): within subspecies (identity), between co-mimic subspecies (mimicry) and between all other subspecies (other). Sample sizes: 38, 19 and 684 subspecies pairs, respectively. Boxes show 25-75% quartiles; horizontal lines, medians; whiskers, inner fence within 1.5 × box height; circles outliers within 3 × box height.

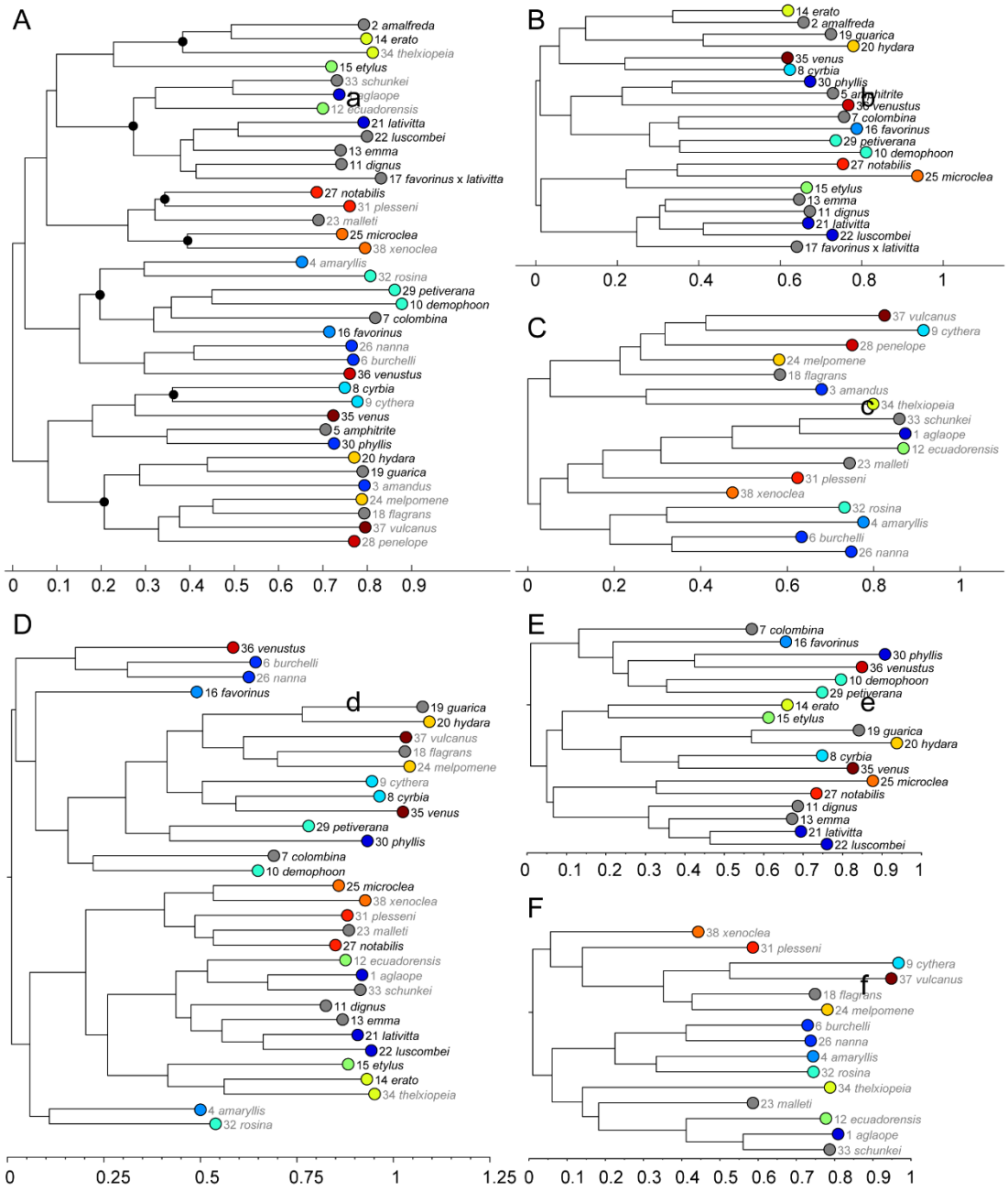


Fig. S6. Neighbor-joining trees of phenotypic distance between subspecies of *H. erato* and *H. melpomene*. (A, D) All subspecies. **(B, E)** Only subspecies of *H. erato*. **(C, F)** Only subspecies of *H. melpomene*. Subspecies label colour indicates species (black *H. erato*, grey *H. melpomene*). Leaf node colours correspond to mimicry groups of Fig. 1. Subspecies numbers correspond to Table S1. Black internal nodes (**A**) show independent clades containing interspecies co-mimics. **(D-F)** Phylogenies reconstructed from average subspecies distances calculated after exclusion of hybrid specimens (Table S2).

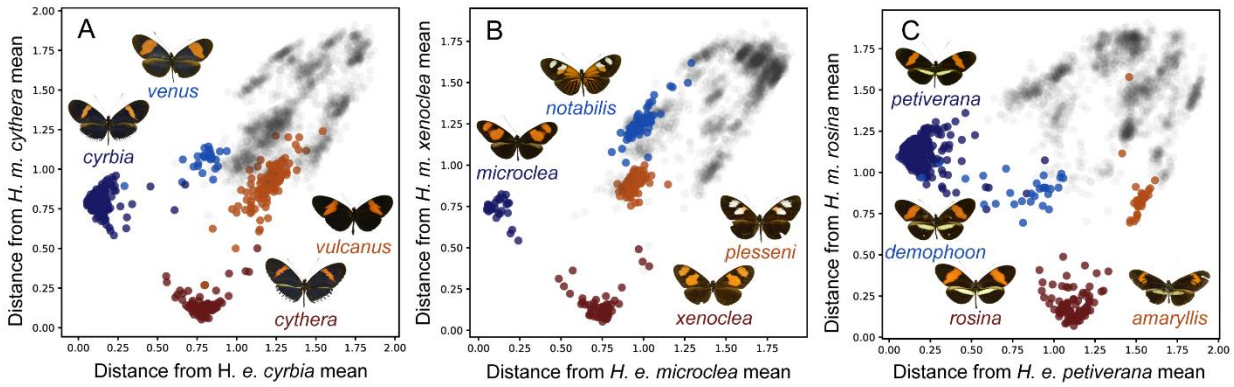


Fig. S7. Comparative analyses of the extent of phenotypic convergence in mimicry. The locations of six focal subspecies **(A-C)** dark blue, *H. erato*: *cyrbia*, *microclea*, *petiverana*; dark red, *H. melpomene*: *cythera*, *xenoclea*, *rosina*) in phenotypic space are compared alongside their six nearest conspecific subspecies **(A-C)** *H. erato*, light blue; *H. melpomene*, light red). Subspecies are illustrated by dorsal photographs of the butterfly closest to the mean location for the subspecies. Grey points indicate images of the other subspecies in the dataset. Axes show the squared distance from the mean location of the focal co-mimic, summed across all 64 spatial embedding axes. As polarised towards the focal taxa, these comparative analyses indicate cases of mutual convergence **(A-B)** as well as implied divergence by *H. erato* where the two species ranges cease to overlap **(c)** in Central-North American *H. erato petiverana* (*23*). In **(A)** the generally less abundant species *H. melpomene* has converged further towards *H. erato*, in line with the frequency dependent fitness benefits predicted in Müllerian mimicry. In another case, **(B)** *H. erato microclea* shows greater convergence on its co-mimic *H. melpomene xenoclea* than *vice versa*. Subspecies mean convergent distance from the focal co-mimic of the other species (distance = conspecific – focal conspecific) and corresponding Mann Whitney *p* values for *H. erato* and *H. melpomene* respectively: **(A)** distance = 0.26, *p* = 1.0195E-15, distance = 0.41, *p* = 5.1718E-31; **(B)** distance = 0.52, *p* = 8.2445E-16, distance = 0.20, *p* = 2.1749E-22; **(C)** distance = -0.22, *p* = 3.2368E-16, distance = 0.41, *p* = 1.1133E-19. Values with reversed evolutionary polarities (focal conspecifics, a-c: *venus*, *vulcanus*; *notabilis*, *plesseni*; *demophoon*, *amaryllis*): **(A)** distance = 0.23, *p* = 1.3176E-14, distance = 0.08, *p* = 3.8215E-12; **(B)** distance = 0.19, *p* = 4.4761E-11, distance = 0.51, *p* = 5.003E-27; **(C)** distance = 0.23, *p* = 6.1463E-17, distance = -0.42, *p* = 1.1133E-19. Statistical values with hybrids excluded (and standard evolutionary polarities) for *H. erato* and *H. melpomene* respectively: **(A)** distance = 0.29, *p* = 2.37E-15, distance = 0.40, *p* = 1.35E-10; **(B)** distance = 0.51 , *p* = 1.44E-11, distance = 0.22, *p* = 4.70E-16; **c,** distance = -0.21, *p* = 6.20E-16, distance = 0.41, *p* = 3.24E-10.

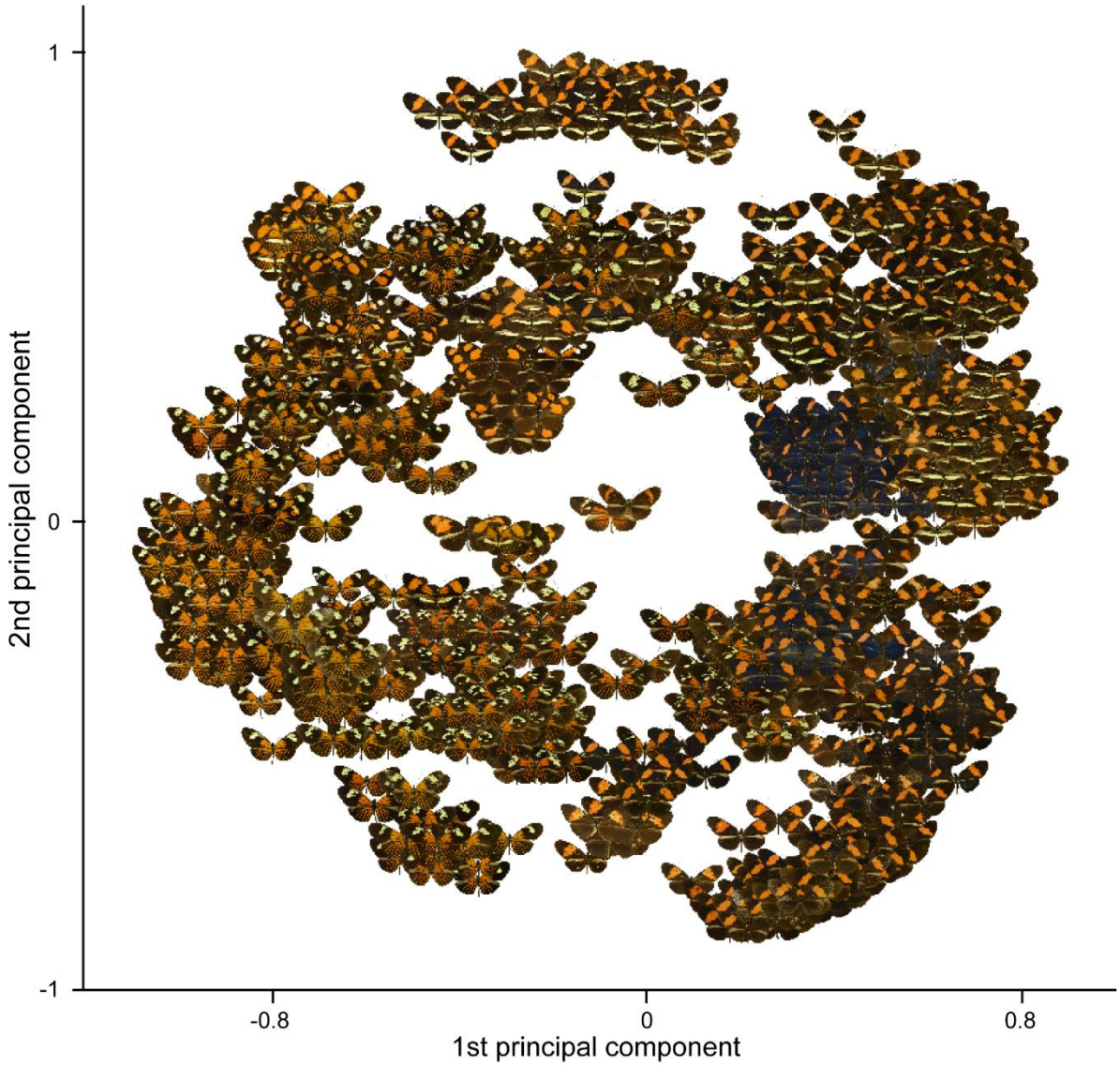


Fig. S8. Principal component visualization of *Heliconius* butterflies. Dorsal photographs of 1234 *Heliconius* butterflies visualised in the space of PCA scores calculated from the deep learning spatial embedding coordinates as described for Fig. 2.