No support for honest signalling of male quality in zebra finch song

Martin Bulla, Remya Sankar and Wolfgang Forstmeier

Alam et al.[1] claim to have discovered a song feature, "path length", that honestly signals male fitness and is therefore preferred by all females. However, their data and analyses provide no statistical support for this claim. (1) The key finding — that long-path songs are difficult to learn (Fig. 4c) — is a statistical artefact: regressing *y minus x* on *x* creates an illusory effect where none exists. (2) Their path-length estimates have a measurement error of 45-73%, which undermines their conclusions, including the claim that females prefer long-path songs. (3) This claim is based on playback experiments that use only three artificial stimulus pairs, which, given the measurement error, cannot reliably contrast long and short paths. In sum, there is no evidence that path length functions as an honest fitness indicator. Our re-evaluation highlights the importance of validating new methods and accounting for random noise in small datasets. Finally, we emphasise that in species where females are known to disagree on who is attractive[2-6], searching for a trait that determines male attractiveness is unwarranted.

1. Statistical artifact

Alam et al.[1] introduce a novel song feature, "path length", which represents "the spread of song in latent space, defined by the minimum path length connecting song syllables". They hypothesise that long-path-length songs are more difficult for juveniles to imitate, and claim that juveniles do indeed struggle to learn them. This claim is based on a reported significant negative correlation between the difference in path length (pupil minus tutor) and the tutor's path length (Fig. 4c). However, this conclusion is based on a statistical artefact: the inclusion of the tutor path length in both the dependent and independent variables, i.e. the regression of *y minus x* on *x*.

Notably, Alam et al.[1]'s initial test of their hypothesis revealed that "all pupils learned reasonably well… and there was no correlation between adult similarity [i.e. the acoustic similarity of pupil's song to the tutor's song] and the path length of the tutor's song (Fig. 4b)", effectively rejecting the hypothesis. They then modified their analysis, claiming that "comparing the change in path length of the pupil's song with the tutor's song… a significant negative correlation [emerges] (Fig. 4c)". As a result, they concluded that "[j]uvenile birds tutored by birds with short-path-length songs were able to match or exceed the path length of their tutor, whereas birds tutored by birds with long-path-length songs struggled to match those path lengths by adulthood". This reported correlation is spurious.

To illustrate, we have generated 1,000 random, uncorrelated values for *x* and *y* (our Fig. 1, top left). Regressing the difference (*y*-*x*) on *x* produces a strong negative correlation (r = -0.7; Fig. 1, top middle), while summing *y and x* and regressing on *x* produces a strong positive correlation *(r* = 0.8; Fig. 1, top right). These correlations arise solely because *x* is included in both axes of the analysis.

The same issue applies to the Alam et al.[1] data. The negative relationship between *pupil minus tutor path length* and *tutor path length* (their Fig. 4c) arises solely because the tutor path length is included on both axes (our Fig. 1, bottom). Critically, the absence of any relationship — linear or otherwise — between pupil and tutor path length (our Fig. 1, bottom left) suggests either that pupils do not learn path length from their tutors or that Alam et al.'s path length is not a biologically meaningful song parameter.

In sum, our findings question the validity of Alam et al.'s conclusions regarding song learning and male quality. They also highlight the importance of testing unconventional statistical procedures on randomly generated data to ensure they do not produce spurious artefacts.

2. Reliability of the path-length metric

The reliability of the novel path-length metric of Alam et al.[1] raises significant concerns. The authors attempted to demonstrate the reliability of path-length estimates across iterations of the latent space (their Fig. 2b) generated by the Uniform Manifold Approximation and Projection (UMAP). However, using the 31 normally reared (tutored) birds in their Fig. 2c (available at[7]), generating 20 UMAP iterations, and calculating the shortest path length between syllable clusters for each male song in each iteration[8], we estimated a path-length repeatability of only 55%, indicating that 45% of the variance in the measure stems from the random-number generator used to initialise the UMAP computation. Crucially, much of this 55% repeatability reflects differences between males in song syllable count (3–8 syllables per song). When adjusted for syllable count, the repeatability drops to mere 27%. This approximates the repeatability for data where all males have the same number of syllables, indicating that in such data 73% of the variance in path length reflects random noise inherent to the UMAP computation. The same issue applies to other datasets from Alam et al.[1], including the tutored birds from their Fig. 2b (Supporting Table 1 in[8]), thus undermining the reliability of the metric.

To illustrate one of the consequences, consider the key experiment in which Alam et al. used three artificial stimulus pairs, comprising "long-path" and "short path" songs, each with five syllables per song. For these pairs, path-length comparisons are highly unreliable, since ~73% of the variance in path-length estimates arises from the randomness inherent in the UMAP computation rather than biological differences. To assess the impact of this uncertainty, we used Alam et al.'s dataset of 31 tutored males (their Fig. 2c[7]) and path-length estimates from 20 latent space iterations[8]. To control for variation due to syllable count, we selected only males with four-syllable songs (n = 11, the largest sample size for any syllable count) and used these to create 55 song pairs. Indeed, we found that a song classified as "long-path" in one UMAP iteration had a 38% probability of being reclassified as "short path" in another (Fig. 2). Assuming that the true path-length difference for each song pair can be approximated by averaging over 20 UMAP iterations, the probability of misclassification for Alam et al.'s three stimulus pairs is 41% for Pair #3 (path-length difference of 6), 21% for Pair #1 (difference of 15), and 6% for Pair #2 (difference of 30; Extended Data Fig. 1). Overall, the probability that all three stimulus pairs were correctly classified is only 44% (the product of the individual correct-classification probabilities).

In sum, using path lengths derived from a single UMAP iteration does not reliably characterise male songs or artificial playback stimuli, questioning the presence of any meaningful biological signal in the metric. Furthermore, UMAP is designed for dimension reduction and preservation of topological structures, not for reliable distance comparisons between clusters. As UMAP focuses on local distances, this issue is particularly pronounced for longer distances between clusters[9]. Thus, any such distance comparison should be validated before drawing any inference from it[9]. In our email correspondence with the UMAP developers (John Healy and Leland McInnes), they recommended computing the distances between cluster centroids in the original high-dimensional space instead of the low-dimensional UMAP space.

3. Limitations in experimental design

A robust playback experiment requires a large number of independent stimuli with strong contrasts in the trait of interest[10,11] (here: long versus short path lengths). This is essential to (i) separate the effect of the trait itself from chance (e.g. a particular playback stimulus sounding attractive or aversive) and to (ii) avoid non-independence of data points caused by reusing the same odd stimulus across trials[10,11]. The effective sample size of Alam et al.'s represents only three stimulus pairs, of which only Pair #2 represents a strong contrast between short and long path (Fig. 3), which limits the ability to disentangle biologically relevant effects from random noise. Furthermore, Alam et al. tested the most informative Pair (#2) the least (n = 3 females; Fig. 3), and females showed the least preference for it (Fig. 3). It is unclear why more song pairs with highly contrasting path lengths were not created and tested with a larger number of females.

Conclusions

Alam et al.'s findings are undermined by statistical artefacts, low metric repeatability, and poor experimental design. The evidence they present does not support the hypothesis that path length in zebra finch songs signals male quality or that females prefer such songs.

Our analysis highlights the critical need for scrutiny and rigorous validation of novel metrics and unconventional statistical procedures, as well as for the design of robust experiments capable of disentangling biologically relevant effects from noise, ensuring reliable interpretation of biological data.

Rethinking honest signalling in mate choice

A long-standing question in behavioural ecology is how females benefit from choosing high-quality mates by evaluating traits that honestly signal male quality. However, large meta-analyses indicate that the quality-related information content of male signals appears to be very low[12,13].

If most signals are indeed largely uninformative, females may be better off ignoring them, rather than competing for the presumed best and most-sought-after males. In species where pair members need to cooperate, females may instead aim for behavioural compatibility, which can significantly increase their fitness[2]. This raises the question of whether female mate preferences are primarily unanimous (seeking overall quality or attractiveness) or individual-specific (seeking compatibility).

The zebra finch is probably the best-studied bird species in this respect. When examining mate choice holistically rather than focusing on single traits, females in large naturalistic (unmanipulated) populations rarely agree on which males they find attractive[2-6]. This is particularly striking when examining female preferences for extra-pair copulation partners; each female has distinct likes and dislikes, suggesting that no common trait makes some males universally more attractive than others[4,5].

Thus, attempts to identify a magic X-factor[14], a single trait that deems certain males attractive, are unlikely to succeed in species without unanimous preferences.

References

1 Alam, D., Zia, F. & Roberts, T. F. The hidden fitness of the male zebra finch courtship song. *Nature* **628**, 117-121, doi: https://doi.org/10.1038/s41586-024-07207-4 (2024).

2 Ihle, M., Kempenaers, B. & Forstmeier, W. Fitness Benefits of Mate Choice for Compatibility in a Socially Monogamous Species. *PLoS Biol* **13**, e1002248, doi:https://doi.org/10.1371/journal.pbio.1002248 (2015).

3 Wang, D., Forstmeier, W. & Kempenaers, B. No mutual mate choice for quality in zebra finches: Time to question a widely-held assumption. *Evolution* **71**, 2661–2676, doi:https://doi.org/10.1111/evo.13341 (2017).

4 Wang, D., Forstmeier, W., Martin, K., Wilson, A. & Kempenaers, B. The role of genetic constraints and social environment in explaining female extra-pair mating. *Evolution* **74**, 544-558, doi:https://doi.org/10.1111/evo.13905 (2019).

5 Wang, D., Forstmeier, W., D'Amelio, P. B., Martin, K. & Kempenaers, B. Is female mate choice repeatable across males with nearly identical songs? *Anim Behav* **181**, 137-149, doi: https://doi.org/10.1016/j.anbehav.2021.09.001 (2021).

6 Forstmeier, W. & Birkhead, T. R. Repeatability of mate choice in the zebra finch: consistency within and between females. *Anim Behav* **68**, 1017-1028, doi:https://doi.org/10.1016/j.anbehav.2004.02.007 (2004).

7 Alam, D. tut.zip, v5. *Texas Data Repository*, doi: https://doi.org/10.18738/T8/WBQM4I/Q92O9A (2024).

8 Bulla, M. & Sankar, R. Supporting information for 'No support for honest signalling of male quality in zebra finch song'. *GitHub*, doi:https://github.com/MartinBulla/rebuttal_alam_2024 (2025).

9 Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nature Reviews Methods Primers* **4**, doi:https://doi.org/10.1038/s43586-024-00363-x (2024).

10 Kroodsma, D. E. Using appropriate experimental designs for intended hypotheses in 'song' playbacks, with examples for testing effects of song repertoire sizes. *Anim Behav* **40**, 1138-1150, doi:https://doi.org/10.1016/S0003-3472(05)80180-0 (1990).

11 Kroodsma, D. E., Byers, B. E., Goodale, E., Johnson, S. & Liu, W.-C. Pseudoreplication in playback experiments, revisited a decade later. *Anim Behav* **61**, 1029-1033, doi: https://doi.org/10.1006/anbe.2000.1676 (2001).

12 Dougherty, L. R., Rovenolt, F., Luyet, A., Jokela, J. & Stephenson, J. F. Ornaments indicate parasite load only if they are dynamic or parasites are contagious. *Evol Lett* **7**, 176-190, doi:10.1093/evlett/qrad017 (2023).

13 Dougherty, L. R. Meta-analysis reveals that animal sexual signalling behaviour is honest and resource based. *Nat Ecol Evol* **5**, 688-699, doi:https://doi.org/10.1038/s41559-021-01409-z (2021).

14 Howe, N. P. & Thompson, B. AI hears hidden X factor in zebra finch love songs. *Nature*, doi:https://doi.org/10.1038/d41586-024-00864-5 (2024).

15 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Usinglme4. *J Stat Softw* **67**, doi:https://doi.org/10.18637/jss.v067.i01 (2015).

Code availability

Code to generate the results is available along with the display item at https://martinbulla.github.io/rebuttal_alam_2024/[8].

Data availability

Data to generate the results are available at https://github.com/MartinBulla/rebuttal_alam_2024[8].

Acknowledgements

We are grateful to all colleagues that invested into discussing the matters with us. M.B. was supported by the Research Excellence in Environmental Sciences Project (REES 003) from the Faculty of Environmental Sciences, Czech University of Life Sciences Prague.

Authors and Affiliations

Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21 Prague 6 – Suchdol, Czech Republic
Martin Bulla (bulla.mar@gmail.com)
Max Planck Institute for Biological Intelligence, Eberhard-Gwinner-Str., 82319 Seewiesen, Germany
Wolfgang Forstmeier (wolfgang.forstmeier@bi.mpg.de)
Remya Sankar (remya.sankar@bi.mpg.de)

Competing interests

The authors declare no competing interests.

Contributions

M.B. and W.F. contributed equally. RS performed the latent space (UMAP) computations and helped in finalizing the paper.

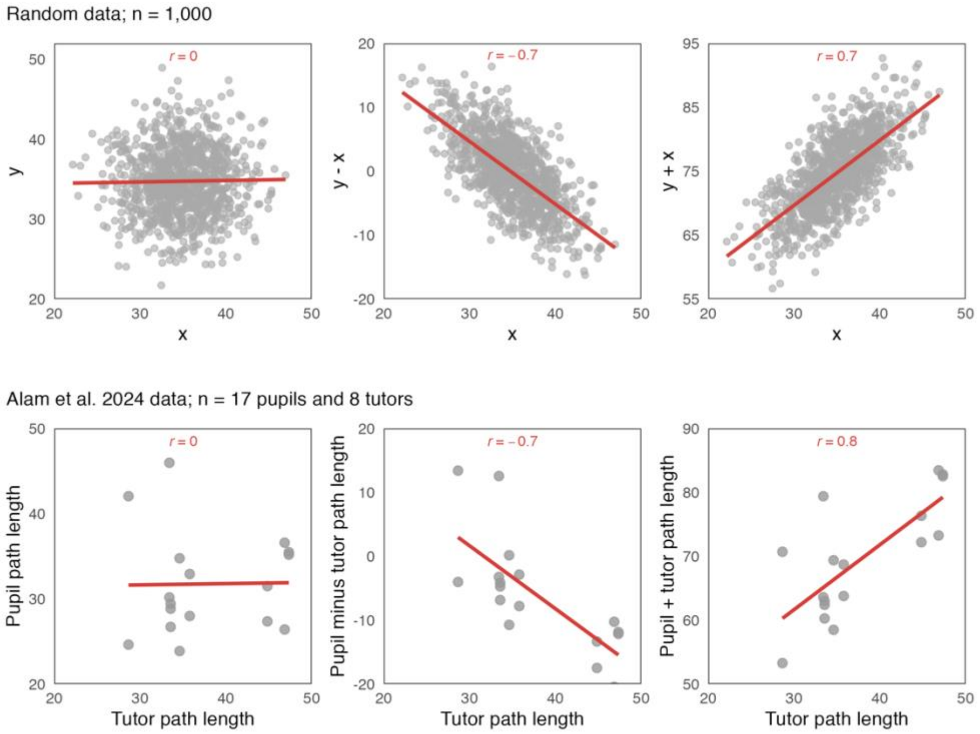


Figure 1 | Illustration of illusory relationships when an x-variable is included in both axes. The **top** panels are based on the 1,000 randomly sampled values of x and y, the **bottom** panels depict data from Fig. 4c of Alam et al. 2024[1]. Lines represent ordinary least-square regressions. *r* denotes a Pearson's correlation coefficient. The **left** panels highlight the absence of relationships in the data. The **middle** panels show negative relationships, and the **right** panels show positive relationships, both arising from including an x-variable also in the y-variable.

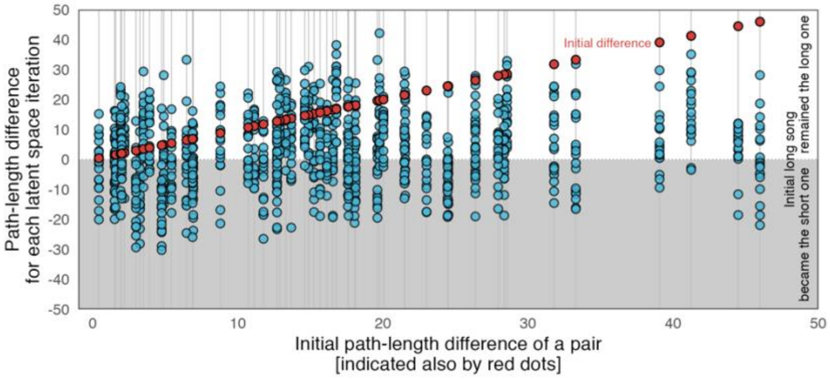


Figure 2 | Inconsistent long-path classification within song pairs across latent (UMAP) space iterations.
The x-axis represents the initial difference in path lengths between two songs of a pair (also highlighted by red points), calculated from a single latent space iteration. The y-axis shows all within-pair path-length differences across 20 latent space iterations. Grey vertical lines highlight individual pairs. We used Alam et al.'s data on 31 tutored birds from Fig. 2c[7] and generated 20 latent space iterations[8]. To control for variation due to syllable count, we selected 11 birds with four-syllable songs (the largest sample size for any syllable count), created 55 song pairs, and calculated within-pair path-length differences. A single latent space iteration served as the reference ("initial difference" in red), just like Alam et al. used only one iteration to define which song in a stimulus pair had the longer path. The blue points show the within-pair path-length differences for the remaining 19 iterations. The grey area highlights the 42% of cases where long-path song is reclassified as short-path in another iteration. Depicted are data for the reference iteration with the largest variation in path-length differences, but other iterations provide similar results (mean = 38% of reclassifications; Extended Data Fig. 2).

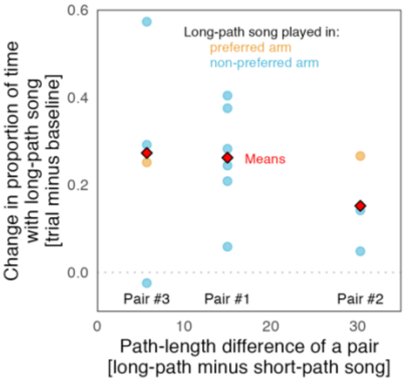
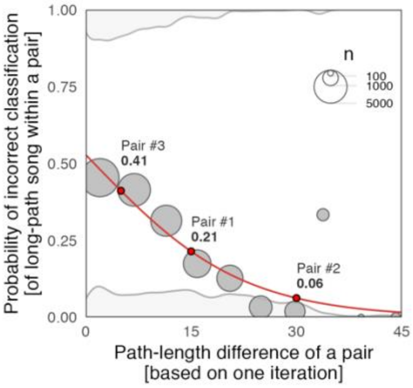
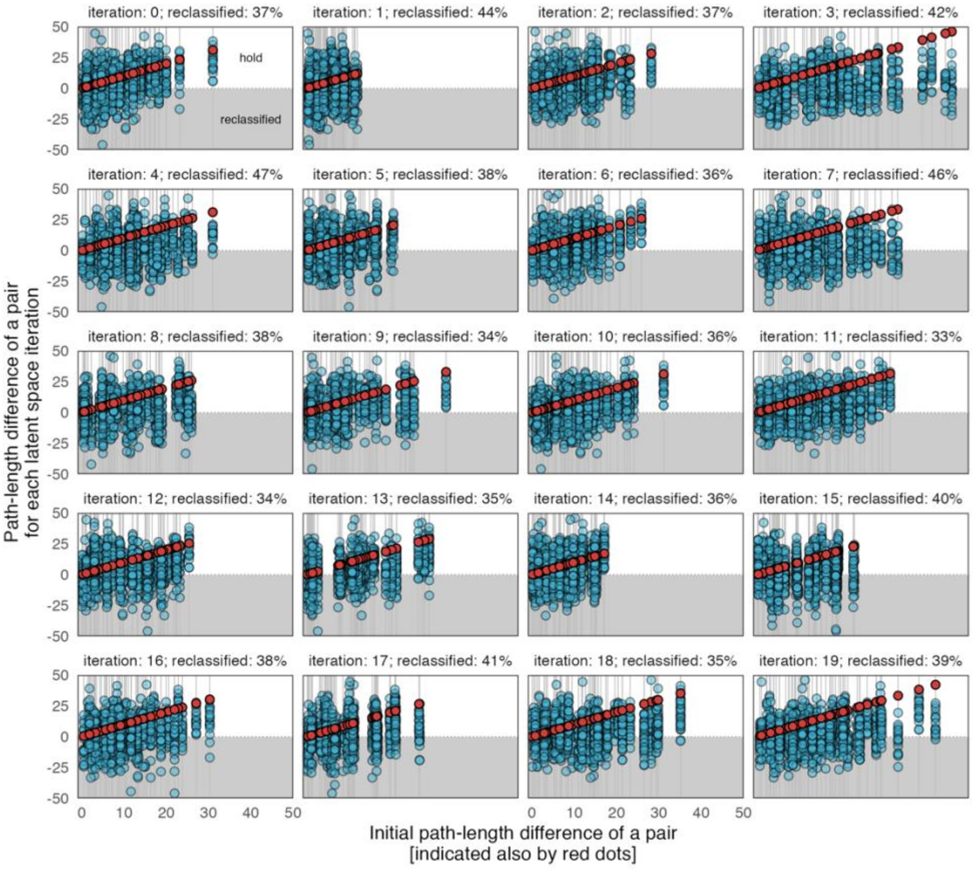


Figure 3 | Female preference for the presumed long-path song in relation to path-length differences within stimulus pairs. The x-axis represents the difference in path lengths between the two songs of each stimulus pair. Notably, only stimulus Pair #2 (at the right end of the plot) shows a strong contrast between a very short and a very long path, yet it elicited the weakest female preference for the long-path song. The y-axis quantifies the strength of female preference for the long-path song, calculated as the change in the proportion of time spent in the choice-chamber arm with the long-path playback during the trial compared to baseline (the mean of pre- and post-trial periods). Dots represent responses of the 13 females to the stimulus pairs (each female tested only once), with dot colour indicating whether the long-path song was played in the arm preferred by the female during the pre-trial period (orange) or in the other arm (blue), illustrating the unbalanced arm-assignment. Red diamonds represent the mean female response to each stimulus pair, indicating the effective sample size of three. The dotted line indicates no preference and thus highlights the negative value for one female; such lack of preference is present in the Alam et al.'s Extended Data Fig. 6, but not in their source data for Fig. 3c where this trial contains identical (duplicated) values from a different trial. Note, since no method ensured that the artificial song stimuli retained natural syntax, some of the six stimuli may have sounded especially interesting or aversive to zebra finches, regardless of their path length.

Extended Data



Extended Data Figure 1 | Probability of incorrect long-path classification within stimulus pairs as a function of their path-length difference in a single UMAP iteration. Density plots at the top and bottom indicate the distribution of the data. Grey points represent means for ten equally spaced x-axis intervals. The red line represents the predicted probability based on the joint posterior distribution of 5,000 simulated values from a logistic regression using the `sim` function from the `arm` R-package. The three red dots, along with their values, indicate the probability of incorrect assignment for the three stimulus pairs used by Alam et al. The data for this figure were obtained as follows. We used Alam et al.'s spectrograms on 31 tutored birds from Fig. 2c[7] and generated 20 latent space iterations[8]. Within each iteration, for each male we calculated the shortest path length between syllable clusters. To control for variation due to syllable count, we selected 11 birds with four-syllable songs (the largest sample size for any syllable count). We fitted an intercept-only mixed-effects model with bird identity as a random intercept and using the `coef()` function in R we extracted an unbiased estimate of the "true inherent path length" for each male[15], approximating an average over an infinite number of iterations. To simulate Alam et al's playback scenario of three stimulus pairs contrasting presumably "long-path" against "short-path" songs, for each of the 20 iterations, we created all unique subsets of six males (n = 462), each with a slightly different set of six males. Within each subset (following Alam et al's Extended Data Fig. 5d), songs were sorted by their path length, and three song pairs created: Pair #1 contrasts the 2nd- and 5th-ranked song, Pair #2 the 1st- and 6th-ranked song, and Pair #3 the 3rd- and 4th-ranked song. This process resulted in 27,720 comparisons (three pairs per subset across 462 subsets and 20 iterations). For each pair, we noted whether the song appearing as "long-path" in a single iteration was actually "short-path" based on its true inherent value. A logistic regression was fitted to estimate how the probability of incorrect classification changes with the path-length difference in a given iteration.



Extended Data Figure 2 | Inconsistent long-path classification within song pairs across latent (UMAP) space iterations. The x-axis represents the initial difference in path lengths between two songs of a pair (also highlighted by red points), calculated from a single latent space iteration. The y-axis shows all within-pair path-length differences across 20 latent space iterations. Grey vertical lines highlight individual pairs. Each panel represents a different latent space iteration serving as the reference. We used Alam et al.'s data on 31 tutored birds from Fig. 2c[7] and generated 20 latent space iterations[8]. To control for variation due to syllable count, we selected 11 birds with four-syllable songs (the largest sample size for any syllable count), created 55 song pairs, and calculated within-pair path-length differences. A single latent space iteration served as the reference ("initial difference" in red), just like Alam et al. used only one iteration to define which song in a stimulus pair had the longer path. The blue points show the within-pair path-length differences for the remaining 19 iterations. The grey area illustrates how often the long-path song is reclassified as short-path in another iteration. The panel titles highlight the percentage of reclassifications, which was overall 38%. Data for five-syllable songs yield similar results (Supporting Fig. 1[8]).