SUPPLEMENTARY MATERIALS

Test 1 Methods – Novel image matching for known individuals

The first test conducted involved evaluating the algorithm's predictive accuracy with 100 images randomly selected of known individuals. These images were of animals that had a known ID in the historical Risso's dolphin catalog but were not the same image in the historical catalog that had been provided to the dorsal fin matching algorithm. This approach was to ensure the algorithm was tested on previously unseen images of known individuals. The purpose of this test was to evaluate how the algorithm matched novel images of known individuals, determine where the correct matches appeared in the output queue, assess the similarity scores, and identify potential missed matches. A missed match was defined as one unique ID in the historical catalog matching another unique ID in the historical catalog, where one individual was given two different IDs and treated as two different individuals.

Of the 100 random images pulled, 95 were cropped into a square shape centered around the dorsal fin and edited to enhance contrast and bring details in pigmentation and dorsal ridging. The remaining five images were left uncropped and unedited in order to evaluate whether image formatting and post-processing influenced the algorithm's performance. All images were organized into a single folder for batch upload to the matching platform.

One-way ANOVAs and linear regressions were used to examine whether image quality or distinctiveness metrics were associated with the algorithm's top similarity score for these images, and the MAP score was calculated. Residual diagnostics on ANOVA indicated that assumptions were reasonably met (Shapiro-Wilk $p = 0.7532$; Levene's test $p = 0.177$), and non-parametric alternatives produced consistent results.

Test 2a Methods – Full historical catalog

Following the results of Test 1, which revealed several missed matches, every image in the historical catalog (n = 2,421 images) was compared against itself in a many-to-many match run. The purpose of this test was to identify missed matches across the entire historical dataset, and assess the typical queue position in which the correct matches appeared, using a larger and more diverse sample set. These were all a "same image" match (i.e., the exact image that was already in the historical catalog was matched to itself, with the expectation that the exact same image should appear first in the queued match options). For manual review, only the top five candidates returned by the algorithm were evaluated as possible matches, based on the results from Test 1 showing that most matches appeared within the first two results.

This test was meant to validate the consistency of the algorithm's performance with identical images and document the typical queue position of missed matches, so no statistical comparisons were made for image quality or distinctiveness within Test 2a. These image quality and distinctiveness characteristics are outlined in Test 2b, which focuses specifically on the missed matches identified during Test 2a.

Test 2b Methods – Missed match rank analysis

While the 2,421 images taken through the algorithm revealed a number of missed matches, every instance of a missed match was not necessarily found with every photo of that individual ID in the first five options presented. Each individual ID could have more than one image in the historical catalog. There could be up to 12 images if it was a Risso's dolphin that had a variety of photos from ahead, laterally and behind, was a tagged animal, or had mostly poor quality images of the dorsal for measurement but a few good quality images showing other parts of the body for eye-based matching.

A detailed investigation was conducted into the missed matches found in Test 2a, by running each of the algorithm-found missed matches through the algorithm again, this time through the entirety of presented possible matches, to determine whether, or where, the known missed match appeared. This allowed further evaluation of how far users may need to search through the algorithm's queue to find correct missed matches.

A total of 203 images were re-tested in this step. One-way ANOVAs and linear regressions were used to examine whether image quality or distinctiveness metrics were associated with the algorithm's top similarity score for these images, and the MAP score was calculated. Residuals were approximately normal (Shapiro-Wilk $p \approx 0.015$), and non-parametric Kruskal-Wallis tests for each factor yielded consistent significance patterns.

Test 3 Methods – Updating Risso's Dolphin Catalog

During Tagless BRS and opportunistic survey efforts in 2022-2024, Cascadia collected images of Risso's dolphins from four different vessels operating in the vicinity of Catalina Island. Across 18 days of survey effort in the area (comprising of 27 boat days), 108 total sightings of Risso's dolphins were documented. The images collected from these sightings were compiled and assigned image quality and distinctiveness metrics through the Risso's database by a human cataloger. Unlike previous catalog updates, these fins were not measured using ImageJ, as this step was no longer necessary given the algorithm's prior success. These images were then run against the multi-species algorithm for matching with the goal of updating Cascadia's Risso's dolphin catalog. Each image was compared against the first 10 match options presented by the algorithm's queue. To evaluate which factors may influence algorithm performance, linear regression and ANOVA were applied to test the effects of image quality and individual distinctiveness metrics on the algorithm similarity score, and residual diagnostics indicated that

model assumptions were reasonably met, with residuals approximately normal (Shapiro-Wilk $p = 0.75$) and homogeneity of variance was satisfied (Levene's test $p = 0.18$). The MAP score was calculated to quantify overall accuracy and to account for cases where the algorithm may not identify a known individual, the MAP was recalculated by introducing a dummy "new individual" prediction at a score of 0.5. If this dummy prediction ranked first, meaning the top algorithm score was <0.5, the precision for that image was 1.0; if it ranked second, meaning the top score was >0.5, the precision was 0.5. This adjustment allowed all images, including those without confirmed matches, to contribute to the overall MAP score. Unlike earlier catalog updates, where the matching process involved two human reviewers (one conducting the initial search for matches, and a second reviewing any unfound matches to ensure the first person did not miss any), here only a single human matcher reviewed the results. Given the algorithm's demonstrated reliability in prior testing, there was enough confidence in its success to forgo the need for a second person to verify whether the algorithm may have missed any unfound matches.

Test 4 Methods – Poor Quality Images

During the initial image compilation process, photographs of individual Risso's dolphins from each sighting were imported into a database and assigned a set of image quality and distinctiveness metrics, as described previously. These metrics were used to both determine the suitability of each image for inclusion in the catalog and assist in eye-based matching by categorizing the animals. Images were often flagged as poor quality in these early stages due to factors such as suboptimal angle, visibility, exposure or sharpness, and did not move forward in the matching process from that point.

For this test, we selected 145 images that had been previously flagged as too "poor quality" for eye-based matching in the compiling stage, with an overall quality score of 2 or

higher (range 2-2.5). These images were compared to the historical catalog through the multi-species matching algorithm to evaluate whether the algorithm was able to identify matches that human matchers had previously deemed unsuitable for catalog inclusion. The MAP score was calculated for both found matches and using the dummy "new individual" prediction calculation, and linear regression and ANOVA analysis were conducted to assess whether image quality or distinctiveness metrics may have influenced the algorithm similarity score.

Match Type Assessment Methods

To evaluate comparative matching performance, all confirmed matches from the multiple datasets were grouped together into three categories: matches identified by both the algorithm and the human (Y/Y), matches identified by the algorithm but missed by the human (Y/N), and images not identified by either the algorithm or the human (N/N). These categories were used to assess whether the algorithm performed comparably to, better than, or worse than human matchers. To further explore potential performance patterns, image quality and distinctiveness metrics were proportionally compared across the three match types to identify specific attributes that may influence algorithmic versus human matching success.