Perch 2.0 transfers 'whale' to underwater tasks

Andrea Burns[*] Google DeepMind
Lauren Harrell[*] Google Research
Bart van Merriënboer Google DeepMind
Vincent Dumoulin Google DeepMind
Jenny Hamer Google DeepMind
Tom Denton Google DeepMind

Abstract

Perch 2.0 is a supervised bioacoustics foundation model pretrained on 14,597 species, including birds, mammals, amphibians, and insects, and has state-of-the-art performance on multiple benchmarks. Given that Perch 2.0 includes almost no marine mammal audio or classes in the training data, we evaluate Perch 2.0 performance on marine mammal and underwater audio tasks through few-shot transfer learning. We perform linear probing with the embeddings generated from this foundation model and compare performance to other pretrained bioacoustics models. In particular, we compare Perch 2.0 with previous multispecies whale, Perch 1.0, SurfPerch, AVES-bio, BirdAVES, and Birdnet V2.3 models, which have open-source tools for transfer-learning and agile modeling. We show that the embeddings from the Perch 2.0 model have consistently high performance for few-shot transfer learning, generally outperforming alternative embedding models on the majority of tasks, and thus is recommended when developing new linear classifiers for marine mammal classification with few labeled examples.

1 Background: why marine bioacoustics is challenging

Marine bioacoustics presents different challenges in data collection, verification, and audio properties versus terrestrial sound understanding. Underwater audio data collection requires both specialized sensors (ceramic pressure sensors) and specialized equipment to properly water-seal the electronics with minimal impact on acoustic quality. Deployment requires remote mooring buoys or scuba divers, which are also more expensive than most terrestrial applications. Low visibility under water makes visual confirmation of the vocalizing species impossible in many cases. Finally, the acoustic signal properties also vary depending on depth, temperature, salinity, and bathymetry (underwater topography), where, depending on these conditions and the energy of the signal, sound propagation underwater can be detected over much longer ranges (Bass and Clark, 2003; Mercado III and Frazer, 1999). In this context, few-shot learning and transfer learning are valuable tools that allow for quick iteration as new sounds are discovered while not having large amounts of human labeled data.

Despite these challenges, passive acoustic monitoring is a critical tool for marine conservation and ecology (Fleishman et al., 2023), and discoveries about underwater sounds continue to be made. E.g., many mystery sounds discovered in underwater acoustic recordings have been attributed to baleen whale species such as minkes and Bryde's whales years or decades later (Allen et al., 2024; Rankin and Barlow, 2005). Humpback songs are constantly evolving and move across populations (Allen et al., 2022; Mercado and Handel, 2012), but can also be used to distinguish between subpopulations (Garland et al., 2015). Whistle, call, and echolocations are used for determining differences in killer whale specific sub-populations called ecotypes (Foote and Nystuen, 2008; Riesch and Deecke, 2011).

[*]Corresponding authors: laurenharrell@google.com, andreaburns@google.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

Thus, new machine learning models are needed to classify and annotate large recording databases at scale for underwater sounds as new discoveries are made and songs emerge across populations.

Recent advances in terrestrial bioacoustics suggest a way forward for rapidly training new models. Previous work has shown excellent few-shot transfer learning—even to new taxonomic groups—using embeddings from bird classifiers (Ghani et al., 2023). Recently, *agile modeling* has been proposed for rapid development of custom bioacoustic classifiers, allowing practitioners to investigate novel fine-grained questions easily (Dumoulin et al., 2025). First, passive acoustic data is embedded in a vector database. Vector search quickly finds important training examples, and linear probes on pre-computed embeddings are fast to train and evaluate, allowing for an efficient active learning loop.

However, the ability of the embedding model to represent target classes is key to the success of the process. In this paper, we demonstrate that new broadly trained terrestrial bioacoustic models produce high-quality classifiers on marine mammal data. Our goal is to evaluate potential embedding models for marine mammal tasks to provide guidance for developing new classifiers.

2 Perch 2.0

Perch is a strong pretrained bioacoustics foundation model (van Merriënboer et al., 2025) that was trained on over 1.5 million labeled recordings from Xeno-Canto, iNaturalist, the Tierstimmenarchiv, and FSD50K, covering over 14,500 species (mostly birds, but also includes insects, mammals and amphibians). The model uses a convolutional architecture (EfficientNet-B3) trained on log-mel spectrograms using a classification loss. Additionally, the model used a form of self-distillation and a self-supervised loss (in the form of source recording prediction) with the goal of producing strong embeddings that are linearly separable for a wide range of bioacoustics tasks. Embeddings from the Perch model have shown successful generalization to tasks other than species classification (e.g., individual identification and call type/dialect classification) and to species that were not represented in the training data (e.g., bats).

3 Datasets

Perch 2.0 was tested on the Watkins Marine Mammal Sound Database (WMMSD) as part of the BEANS benchmark (Hagiwara et al., 2023). However, the Watkins database consists of focal recordings from a broad range of conditions (including microphone recordings from ships or land), and thus may not reflect model transfer to passive hydrophone recordings. To establish Perch 2.0's performance for underwater audio tasks more thoroughly for ecology and passive acoustic monitoring use cases, we evaluate its performance on three marine audio validation sets: NOAA PIPAN, ReefSet, and DCLDE 2026.

The DCLDE 2026 dataset (Palmer et al., 2025) consists of recordings from multiple providers and hydrophone systems for training models to distinguish between killer whale (*Orcinius orca*) subpopulations and other noises in the northeastern Pacific. The data include over 200,000 annotations averaging 0.73s in length, where 99% of the annotations are shorter than 3.0s. For evaluation in this work, the annotations are broken into three different label sets as distinct tasks:

- Species: humpback, orca, abiotic, and undetermined biological sounds. The specific ecotype annotations for orcas are combined into a single label for this task.
- Ecotype: orca-NRKW, -OKW, -SAR, -SRKW, and -TKW. Ecotypes are locally adapted populations with distinct traits, e.g., NRKW refers to northern resident killer whales (orcas) which live in the northeast part of the Pacific Ocean. This classification is only over orcas, as there are no ecotype annotations for humpbacks.
- Known species: humpback and orca. The specific ecotype annotations for orcas are combined into a single label for this task, and labels for killer whales where the annotator is uncertain are dropped.

NOAA PIPAN data is an annotated subset of the NOAA Passive Acoustic Archive (NOAA Pacific Islands Fisheries Science Center, 2021). Specifically, it contains recordings from the NOAA Pacific Islands Fisheries Science Center deployments. The labeled segments were extracted from annotations provided in Allen et al. (2021, 2024) in addition to additional annotations provided by NOAA that

includes new classes unseen by the models. The classes used in the NOAA PIPAN evaluation set include anthropomorphic noise, unknown whale species, and the following baleen whale species: common minke whale, humpback whale, sei whale, blue whale, fin whale, and Bryde's whale. All examples from NOAA PIPAN are 30s in length and are considered weakly labeled. The ReefSet data are described in detail in Williams et al. (2025), but include a mix of biological reef noises (such as croaks, crackles, growls), classes for some species/genera (e.g., damselfish, dolphins, and groupers), as well as anthropomorphic noise and wave classes. ReefSet examples are all 1.88s in length.

4 Method

Evaluation method. We use a transfer learning protocol with linear probing to compare embedding performance for each dataset task. This evaluation protocol is established for data that are potentially 'weakly' labeled such that the target class is contained in a longer file; for example, the NOAA PIPAN data contains 30-second files labeled at the file level, but we don't know the specific timestamp where the target species vocalization occurs. We apply the same protocol for each model. First, we compute an embedding for each recording in the dataset by chunking it into fixed-sized windows (either 5s or 3s, depending on the model specifications) with the hop size set to the same size as the window, embedding all resulting windows with the model.

Where the labeled recording duration exceeds the window size, we average the embeddings across the windows into a recording-level embedding that is then normalized. Mean pooling the embeddings only applies to NOAA PIPAN for all models, and DCLDE examples longer than 3.0 or 5.0 seconds. We randomly select k recording-level embeddings per class, where $k \in \{4, 8, 16, 32\}$, to train a logistic regression classifier and hold out the remaining embeddings to compute a one-vs-all ROC-AUC metric. We repeat the process 5 times with independently-sampled training sets and report the average ROC-AUC. Thus, we evaluate the embeddings of these models in a few-shot classification setup and do not use the predictions of the models directly. In the development of Perch 2.0, these tasks were used as one of the validation sources for model selection (van Merriënboer et al., 2025).

Table 1: Embedding model properties

Model name	Sample Rate	Window Size (s)	Embedding Dimension	# Model Parameters	Training Taxa
Perch 2.0	32 kHz	5.0	1536	101.8M	Broad Terrestrial
Perch 1.0	32 kHz	5.0	1280	23.9M	Birds only
SurfPerch	32 kHz	5.0	1280	24.2M	Birds + Reefs
GMWM	24 kHz	3.0	1280	4.1M	Whales
Birdnet V2.3	48 kHz	3.0	1024	20.0M	Birds, Frogs
AVES-bio	16 kHz	Variable	768[2]	94.4M	General Audio
BirdAVES (large)	16 kHz	Variable	1024[3]	315.4M	General Audio + Birds[4]

Comparison models. As our goal is to provide guidance on which pretrained embedding models should be used for agile modeling and transfer learning (with existing tools), we limit our comparisons to models supported in the Perch Hoplite Github repository[5]. We compare the performance of the mean embeddings generated by Perch 2.0, Perch 1.0, SurfPerch (Williams et al., 2025), the Google multispecies whale model (GMWM) (Allen et al., 2024; Harvey et al., 2024), BirdNet 2.3 (Kahl et al., 2021), and the AVES-bio and BirdAVES models (Hagiwara, 2023). A summary of comparison model properties can be found in Table 1.

Also note that the published SurfPerch model was trained on the ReefSet data, and similarly, a large portion of the NOAA PIPAN labeled audio data was used to train the GMWM (but the multispecies whale model does not include classes or label sets for sei whales, anthropomorphic noise, or unknown whale). In the NOAA dataset, two classes have fewer examples than can be supported with higher

[2]The AVES-bio output embeddings are computed over a sliding window, providing a 2D array of size ((window size (s) * 49) - 1, 768). We applied mean pooling across the outputs to obtain a single 768 dimensional embedding to be used in linear probing, consistent with the recommendations in the paper.

[3]Similarly, mean pooling was applied to the BirdAves data to obtain a single 1024 dimension representation.

[4](Xeno-Canto)

[5]`https://github.com/google-research/hoplite`

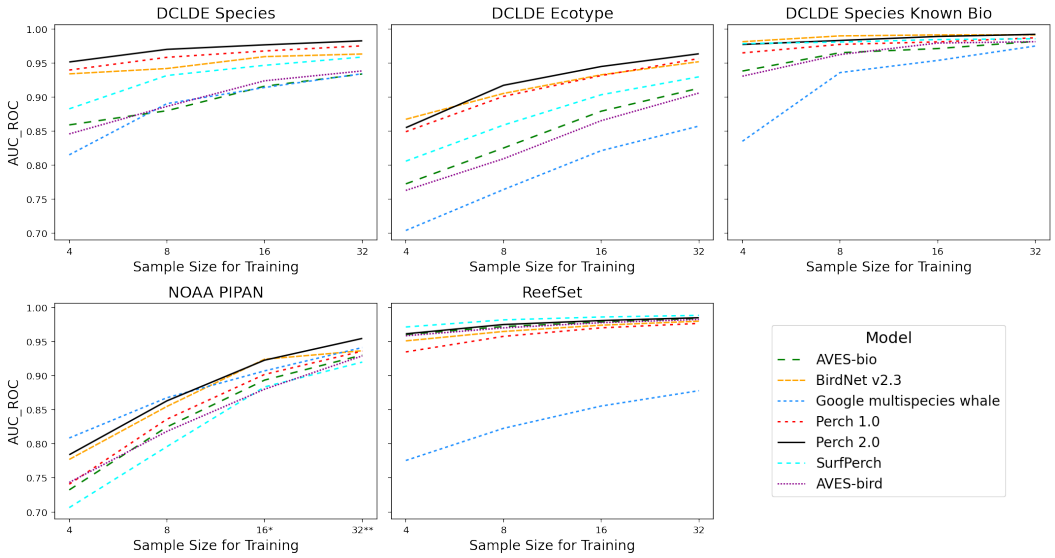

Figure 1: Performance of trained models on marine datasets, varying the number of training examples per-class. *Class "Bm" dropped for $k = 16$; **Classes 'Bm' and 'Be' dropped for $k = 32$ in NOAA PIPAN data.

k values, and thus we drop classes for which we have less than $k + 1$ recording-level embeddings. Class 'Bm' for $k = 16$ and classes 'Bm' and 'Be' for $k = 32$ are dropped for the NOAA dataset.

The data for DCLDE 2026 is unseen for all models, although GMWM was trained on killer whales. The scores in Table 2 may slightly differ from previously reported results in the literature due to differences in the linear probe estimation implementation[6]. Perch 2.0's training data includes around a dozen cetacean recordings from iNaturalist, but these were mostly phone recordings made above water and are not reflective of underwater hydrophone recordings. We bold the highest performance of the non-contaminated models (i.e., those which have no overlapping train/test data).

Table 2: Marine learning transfer tasks

Model	DCLDE 2026						NOAA PIPAN		Reefset	
	Species		Ecotype		Known Bio Species					
	$k = 8$	$k = 16$	$k = 8$	$k = 16$	$k = 8$	$k = 16$	$k = 8$	$k = 16$	$k = 8$	$k = 16$
GMWM	0.890	0.914	0.764	0.821	0.936	0.954	0.868*	0.917*	0.823	0.855
SurfPerch	0.932	0.947	0.859	0.903	0.981	0.984	0.796	0.899	0.982*	0.986*
Perch 1.0	0.958	0.968	0.901	0.931	0.977	0.981	0.836	0.905	0.958	0.970
Perch 2.0	**0.970**	**0.977**	**0.917**	**0.945**	0.983	0.989	**0.863**	**0.924**	**0.975**	**0.981**
Birdnet V2.3	0.942	0.959	0.905	0.933	**0.990**	**0.991**	0.855	**0.924**	0.965	0.974
AVES	0.880	0.916	0.825	0.879	0.965	0.971	0.825	0.893	0.972	0.979
AVES-Bird	0.886	0.924	0.809	0.865	0.962	0.979	0.818	0.880	0.970	0.978

AUC-ROC of few shot transfer learning comparing Perch 2.0 to other pretrained models for $k \in \{8, 16\}$ samples per class during training. (*) indicates the dataset was in the training data for the embedding model. The Google Multispecies Whale Model (GMWM) was also evaluated in the off-the-shelf classification setting and only reached 0.612 performance.

5 Results

Table 2 contains summaries of the results for $k = 8$ and $k = 16$ and Figure 1 shows results for $k \in \{4, 8, 16, 32\}$. Two of the selected evaluation datasets were in the training data for two of the models: GMWM was trained on NOAA PIPAN and SurfPerch was trained with ReefSet. We observe

[6]In particular, Williams et al. (2025) evaluated the SurfPerch model on ReefSet by training a linear layer using mini-batches and the Adam optimizer. In our experiments we use scikit-learn's `LogisticRegression` class, which uses the L-BFGS optimizer and weight decay by default.

that while those models did well on their training datasets, they had worse transfer to new tasks. Aside from the cases of data contamination, Perch 2.0 outperforms the published comparison models on ReefSet and all cetacean species tasks (DCLDE 2026 species, ecotypes, and NOAA PIPAN whales), except DCLDE 2026 Known Bio Species, where it comes in second to BirdNet V2.3.

For the DCLDE data, using GMWM's pretrained classification scores for the supported classes yields an AUC-ROC of 0.612. Using the embeddings from this model for few-shot learning the performance jumps to 0.954. The poor performance of the off-the-shelf classification head may be due to overfitting to a particular microphone or other characteristics of the training data.

6 Discussion - why do models trained on birds work so whale?

The performance of Perch 2.0 and BirdNet on marine mammals is perhaps surprising given the lack of underwater audio included in the training set, but we may hazard an explanation. First, neural scaling laws (Kaplan et al., 2020) tell us that larger models with more training data will most likely perform better. Similarly to how ImageNet-pretrained models dominated computer vision for a long time, it is reasonable to assume that bioacoustics models pretrained on large datasets with thousands of classes can be high performing, even on out-of-domain downstream tasks. Next, the 'bittern lesson' learned when training Perch 2.0 was that bird species classification in particular is a challenging supervision task (van Merriënboer et al., 2025); the small inter-class variances and large number of classes are likely to force the network to learn detailed acoustic features which are useful for a wide variety of downstream bioacoustics tasks (Hong et al., 2024). It is also notable that a wide variety of species—including birds and marine mammals—have evolved similar means of sound production, namely the myoelastic-aerodynamic mechanism (Elemans et al., 2015). This could partially explain why methods transfer remarkably well across these taxa.

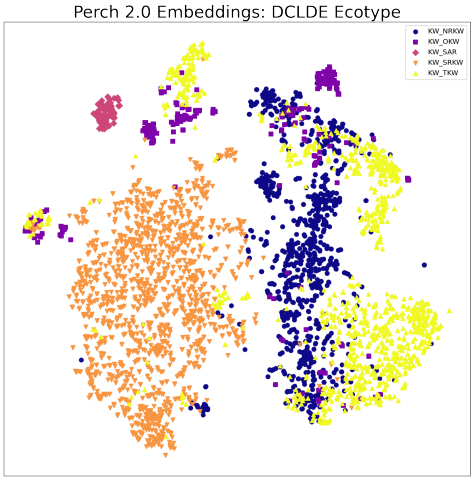

Figure 2: tSNE plot of Perch 2.0 Embeddings for DCLDE 2026 Ecotype data

In Figure 2, we illustrate the quality of Perch 2.0's embeddings with a tSNE visualization which plots samples from the DCLDE 2026 ecotype subset. The pooled embeddings are first compressed to dimension 32 with PCA prior to running tSNE. We see that there is strong separability between different ecotypes of the same species (orca), demonstrating that the representations are indeed discriminative. Additional tSNE plots for the other reported models can be found in the Appendix.

While the focus of our efforts was to understand Perch 2.0's performance on underwater transfer tasks, our results show several evaluated embedding models have high performance for transfer-learning, particularly BirdNet. While AVES performance for tasks with fewer examples was often lower, it's possible the method for pooling the embeddings could be modified for improved performance.

Overall, we observe strong few-shot transfer learning performance for general terrestrial bioacoustic models to marine tasks. We expect that using these pretrained models in an agile modeling context will greatly improve the efficiency of a wide range of marine bioacoustic tasks.

Acknowledgments

Special thanks to Ann Allen (NOAA Pacific Islands Fisheries Center) for providing additional annotations used in the NOAA PIPAN dataset, Isabelle Simpson (Google DeepMind), and Dan Morris (Google Research).

References

A. N. Allen, M. Harvey, L. Harrell, A. Jansen, K. P. Merkens, C. C. Wall, J. Cattiau, and E. M. Oleson. A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Front. Mar. Sci.*, 8:607321, Mar. 2021.

A. N. Allen, M. Harvey, L. Harrell, M. Wood, A. R. Szesciorka, J. L. McCullough, and E. M. Oleson. Bryde's whales produce biotwang calls, which occur seasonally in long-term acoustic recordings from the central and western north pacific. *Frontiers in Marine Science*, 11:1394695, 2024.

J. A. Allen, E. C. Garland, C. Garrigue, R. A. Dunlop, and M. J. Noad. Song complexity is maintained during inter-population cultural transmission of humpback whale songs. *Scientific reports*, 12(1): 8999, 2022.

A. H. Bass and C. W. Clark. The physical acoustics of underwater sound communication. In *Acoustic communication*, pages 15–64. Springer, 2003.

V. Dumoulin, O. Stretcu, J. Hamer, L. Harrell, R. Laber, H. Larochelle, B. van Merriënboer, A. Navine, P. Hart, B. Williams, T. A. C. Lamont, T. B. Razak, Mars Coral Restoration Team, S. Brodie, B. Doohan, P. Eichinski, P. Roe, L. Schwarzkopf, and T. Denton. The search for squawk: Agile modeling in bioacoustics. *arXiv [eess.AS]*, May 2025.

C. P. H. Elemans, J. H. Rasmussen, C. T. Herbst, D. N. Düring, S. A. Zollinger, H. Brumm, K. Srivastava, N. Svane, M. Ding, O. N. Larsen, S. J. Sober, and J. G. Švec. Universal mechanisms of sound production and control in birds and mammals. *Nat. Commun.*, 6(1):8978, Nov. 2015.

E. Fleishman, D. Cholewiak, D. Gillespie, T. Helble, H. Klinck, E.-M. Nosal, and M. A. Roch. Ecological inferences about marine mammals from passive acoustic data. *Biological Reviews*, 98 (5):1633–1647, 2023.

A. D. Foote and J. A. Nystuen. Variation in call pitch among killer whale ecotypes. *The Journal of the Acoustical Society of America*, 123(3):1747–1752, 2008.

E. C. Garland, A. W. Goldizen, M. S. Lilley, M. L. Rekdahl, C. Garrigue, R. Constantine, N. D. Hauser, M. M. Poole, J. Robbins, and M. J. Noad. Population structure of humpback whales in the western and central south pacific ocean as determined by vocal exchange among populations. *Conservation Biology*, 29(4):1198–1207, 2015.

B. Ghani, T. Denton, S. Kahl, and H. Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, Dec. 2023.

M. Hagiwara. AVES: Animal vocalization encoder based on self-supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, June 2023.

M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian. BEANS: The benchmark of animal sounds. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, June 2023.

M. Harvey, A. Allen, and L. Harrell. Google multispecies whale model, 2024.

G. Z. Hong, Y. Cui, A. Fuxman, S. H. Chan, and E. Luo. Why fine-grained labels in pretraining benefit generalization? *Transactions on Machine Learning Research*, 2024.

S. Kahl, C. M. Wood, M. Eibl, and H. Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecol. Inform.*, 61(101236):101236, Mar. 2021.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv [cs.LG]*, Jan. 2020.

E. Mercado and S. Handel. Understanding the structure of humpback whale songs (l). *The Journal of the Acoustical Society of America*, 132(5):2947–2950, 2012.

E. Mercado III and L. N. Frazer. Environmental constraints on sound transmission by humpback whales. *The Journal of the Acoustical Society of America*, 106(5):3004–3016, 1999.

NOAA Pacific Islands Fisheries Science Center. Pacific islands passive acoustic network (PIPAN) 10kHz data, 2021.

K. J. Palmer, E. Cummings, M. G. Dowd, K. Frasier, F. Frazao, A. Harris, A. Houweling, J. Kanes, O. S. Kirsebom, H. Klinck, H. LeBlond, L. Laturnus, C. Matkin, O. Murphy, H. Myers, D. Olsen, C. O'Neill, B. Padovese, J. Pilkington, L. Quayle, A. R. Vuibert, K. Trounce, S. Vagle, S. Veirs, V. Veirs, J. Wladichuk, J. Wood, T. Yack, H. Yurk, and R. Joy. A public dataset of annotated orcinus orca acoustic signals for detection and ecotype classification. *Sci. Data*, 12(1):1137, July 2025.

S. Rankin and J. Barlow. Source of the north pacific "boing" sound attributed to minke whales. *The Journal of the Acoustical Society of America*, 118(5):3346–3351, 2005.

R. Riesch and V. B. Deecke. Whistle communication in mammal-eating killer whales (orcinus orca): further evidence for acoustic divergence between ecotypes. *Behavioral Ecology and Sociobiology*, 65(7):1377–1387, 2011.

B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton. Perch 2.0: The bittern lesson for bioacoustics, 2025. URL `https://arxiv.org/abs/2508.04665`.

B. Williams, B. van Merriënboer, V. Dumoulin, J. Hamer, A. B. Fleishman, M. McKown, J. Munger, A. N. Rice, A. Lillis, C. White, C. Hobbs, T. Razak, D. Curnick, K. E. Jones, and T. Denton. Using tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 380(1928):20240280, June 2025.

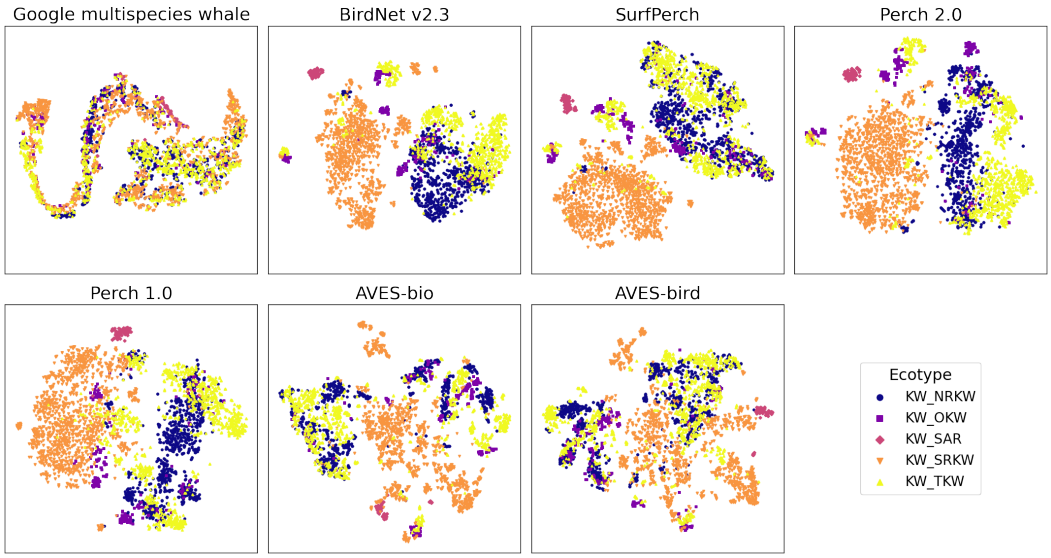

Figure 3: tSNE plots on the DCLDE 2026 Ecotype dataset which contains five ecotype variants of the killer whale (orca) species. Plots were generated with Sklearn PCA and tSNE libraries, with embeddings first projected to 32 dimension vectors prior to tSNE being applied.

7 Appendix

7.1 Additional tSNE Visualizations

In Figure 3 we include tSNE plots on the DCLDE 2026 Ecotype dataset for all evaluated models (GMWM, SurfPerch, Perch 1.0, Perch 2.0, Birdnet V2.3, AVES-bio, and AVES-Bird). Somewhat unsurprisingly, given the quantitative results of Table 2, the Google multispecies whale model embeddings do not show strong linear separability.

All other models show reasonable clustering of orca ecotype, but AVES-bio and BirdAves models show a greater amount of class entanglement, especially with the southern resident killer whale (SRKW) ecotype, which other models have more distinctly separated. Perch 1.0 also does not separate SRKW as well as Perch 2.0, SurfPerch, or BirdNet.

Finally, we see the AVES models do not well separate the transient (TKW) and northern resident killer whales (NRKW). On the other hand, Perch 2.0 appears to have the best boundary between the NRKW and TKW ecotypes.