PHILOSOPHICAL TRANSACTIONS B

Phil. Trans. R. Soc. B.
doi:10.1098/not yet assigned

Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics

Ben Williams[abc], Bart van Merriënboer[b], Vincent Dumoulin[b], Jenny Hamer[b], Abram B. Fleishman[d], Matthew McKown[d], Jill E. Munger[de], Aaron N. Rice[fg], Ashlee Lillis[h], Clemency E. White[ij], Catherine A. D. Hobbs[ij], Tries B. Razak[k], David J. Curnick[c], Kate E. Jones[a], Tom Denton[l]

[a]University College London, UK. [b]Google DeepMind, London, UK. [c]Zoological Society of London, UK.[d]Conservation Metrics Inc., Santa Cruz, USA. [e]University of New Hampshire, USA. [f]K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University. [g]Department of Public and Ecosystem Health, Cornell University. [h]Sound Ocean Science, South Africa. [i]University of Exeter, UK. [j]University of Bristol, UK. [k]IPB University, Indonesia. [l]Google Research, San Francisco, USA.

Keywords: Bioacoustics, passive acoustic monitoring, machine-learning, coral reef, marine

Summary

Machine learning has the potential to revolutionize passive acoustic monitoring (PAM) for ecological assessments. However, high annotation and compute costs limit the field's adoption. Generalizable pretrained networks can overcome these costs, but high-quality pretraining requires vast annotated libraries, limiting their current development to data-rich bird taxa. Here, we identify the optimum pretraining strategy for data-deficient domains using tropical reefs as a representative case study. We assembled ReefSet, an annotated library of 57k reef sounds taken across 16 datasets, though still modest in scale compared to annotated bird libraries. We performed multiple pretraining experiments, finding that pretraining on a library of bird audio 50 times the size of ReefSet provides notably superior generalizability on held out reef datasets, with a mean AUC-ROC of 0.881 (±0.11) compared to pretraining on ReefSet itself or unrelated audio, with a mean AUC-ROC of 0.724 (±0.05) and 0.834 (±0.05) respectively. However, our key findings show that cross-domain mixing, where bird, reef and unrelated audio are combined during pretraining, provides a superior transfer learning performance, with an AUC-ROC of 0.933 (±0.02). SurfPerch, our optimum pretrained network, provides a strong foundation for automated analysis of tropical reef and related PAM data with minimal annotation and compute costs.

*Author for correspondence (ben.williams.20@ucl.ac.uk).
†Present address: People and Nature Lab, One Pool St, UCL East, E20 2AH, UK

Introduction

Advanced monitoring tools are key to tackling the biodiversity crisis (Pimm et al., 2015). Passive acoustic monitoring (PAM) represents a powerful medium through which to gather data for ecological assessments (Gibb et al., 2019; Ross et al., 2023). Low-cost autonomous recording units are now widely available, enabling the collection of vast quantities of PAM data with considerably lower logistical and expertise costs in the field (Hill et al., 2018; Lamont et al., 2022; Shonfield et al., 2017). However, a boom in their accessibility and application has resulted in data collection scaling beyond the analytical capacity of human annotators (Gibb et al., 2020). Effective automated analysis is therefore required to alleviate this analytical bottleneck and maximise the potential of these data. Machine learning (ML) has emerged as a powerful tool with the potential to meet this demand, with state-of-the-art approaches typically leveraging deep neural networks (Stowell, 2022). The current paradigm for ML-accelerated PAM analysis employs supervised learning techniques to train classifiers which can detect target signals. These supervised learning techniques are typically used to develop species-level detectors or identify anthropogenic activities (Gibb et al., 2020; Stowell, 2022).

A key drawback of traditional supervised approaches is their reliance on large annotated libraries of validated target sounds, typically requiring hundreds of examples per class to train an accurate classifier. These libraries are costly and time-consuming to assemble, primarily due to their reliance on human annotators (Kholghi et al., 2018). Additionally, classifiers often generalize poorly to "out of-distribution" data, where new data differs significantly from the initial training set (e.g., new field sites, different microphones). To address these issues, recent efforts have sought to develop broad multi-species classifiers that are trained on large and diverse libraries of recordings. Well-established pretrained bioacoustic networks have been primarily trained on bird taxa, where large open-source annotated libraries are available (e.g., Xeno-canto, Macaulay Library)(Ghani et al., 2023; Kahl et al., 2021). These pretrained networks can sometimes be used off the shelf to classify sounds in new recordings, but this is restricted to only the classes present in their training set. Furthermore, these pretrained classifiers still underperform on novel datasets which are out-of-distribution from their training data and on classes which are under-represented in their training data (Stowell, 2022; Pérez-Granados, 2023), limiting their broad application and utility.

When a pretrained network cannot directly classify novel signals, transfer learning represents an effective alternative. Here, samples from new target classes are passed through the pretrained network, and outputs from an intermediate layer are used to produce a feature embedding representation of the samples. These fixed embeddings can then be used to train a lightweight machine learning classifier (Ghani., et al 2023; White et al., 2023). This removes the costly training phase by leveraging knowledge from the network's embedding space learned during pretraining. Additionally, strong pretrained models facilitate few-shot transfer learning, where only a small number of training examples are needed to produce a highly accurate

classifier, significantly reducing annotation costs (Kath et al., 2024). Emerging research shows networks pretrained on unrelated terrestrial bioacoustic data transfer well to similar bioacoustic domains, enabling few-shot learning (Ghani et al., 2023). However, the ability of pretrained bioacoustic networks to transfer to highly novel domains, such as aquatic environments, is largely untested (Williams et al., 2024). Substantial domain shifts may require the development of novel pretrained networks to achieve accurate few-shot transfer learning. The optimal pretraining strategies to produce these networks remain unknown, which is further compounded by the sparsity of annotated libraries for novel domains.

Coral reef ecosystems host some of the highest documented bioacoustic diversity in the ocean (Kaplan et al., 2018; McWilliam et al., 2018), yet ML-accelerated PAM analysis is significantly underdeveloped for these habitats. Coral reefs host >25% of marine biodiversity and >375M people are directly reliant on the ecosystem services these habitats provide (Hoegh-Guldberg et al., 2019; Knowlton et al., 2010). They are also among the most threatened habitats globally, with >50% of the world's reefs lost over the last 70 years and a projected loss of 90% of those remaining by 2050 (Eddy et al., 2021; IPCC 2019). The soundscapes of these habitats have been found to contain information relevant to key ecosystem attributes such as coral cover and fish community diversity, as well as representing a key ecosystem function which drives larval recruitment (Kaplan et al., 2018; Pysanczyn et al., 2023). PAM is therefore emerging as a promising tool to monitor these threatened habitats (Mooney et al., 2020), but there is a sparsity of relevant annotated data. As is common for underwater soundscapes, the majority of biological sounds on coral reefs remain un-documented and for a significant portion of those that have been recorded, the taxonomic source of origin has not been validated (Parsons et al., 2022; Rountree et al., 2019). Automated analysis of reef PAM data is therefore highly underdeveloped, with only a very limited number of studies having used ML-accelerated analysis on PAM data from these habitats (Lin et al., 2018; Ozanich et al., 2021; Williams et al., 2022). Consequently, coral reef habitats represent an excellent candidate for assessing novel few-shot transfer learning bioacoustic frameworks, with advances in this field having the potential to help address real-world conservation challenges.

In this study, we aim to empirically identify an effective pretraining strategy to produce an efficient network which supports accurate few-shot transfer learning for a novel bioacoustic domain. Such a network should facilitate analysis of PAM data with minimal computational and annotation costs. We selected the coral reef domain due to the threatened status and high acoustic diversity of these ecosystems, with potential to provide a strong foundation for transferring to other aquatic habitats. To achieve this, we first assembled ReefSet, the largest published dataset of annotated reef recordings to date, though only 1.99% the size of comparable bird libraries at the time of writing (Xeno-canto, 2023). We then set out to determine how well existing pretrained networks perform at few-shot transfer learning on ReefSet. Next, we tested whether performance could be enhanced by i) pretraining on this significantly smaller but in-domain dataset, and ii) through cross-domain mixing during

pretraining. Finally, to assess the generalizability of this strategy across the reef domain, we tested whether the output from cross-domain mixing optimized for the coral reef domain impacts generalizability to unrelated domains.

Methods

Data compilation

To maximize the generalizability of our approach, we compiled a diverse meta-dataset of 57,084 labelled coral reef bioacoustic recordings across 37 classes and from 16 individual datasets over 12 countries (Fig. 1; Supp. 1. Table S1), hereon referred to as “ReefSet”. Each individual dataset was originally collected and labelled for different purposes using a variety of sampling strategies and hydrophone models (Supp. 1. Table S2). During the annotation of each dataset, longer recording periods were cut into samples of shorter windows (1.88 sec) to fit two within two window lengths of the industry standard networks YAMNet and VGGish (Table 1) at the time of curation. Samples were labelled by human annotators using aural and visual inspection of each sample's spectrogram. While many classes were of a known origin, others were unknown but typically presumed to originate from fish. All labelled samples were then re-sampled to 16 kHz and written out as a separate waveform audio file.

To amalgamate class labels across datasets, each sample was first given a single primary label: biophony, anthrophony, geophony or ambient. Here, the ambient label was used for negative samples in the Florida-boats, Kenyan and Indonesian datasets where the annotation strategy used a positive label class (motorboat, fish noise and bomb fishing respectively) alongside a strongly labelled negative class. A single secondary label was then applied to all other samples using existing labels from the datasets, with merging of labels under a common name where sounds matched across multiple datasets (e.g motorboats). Within any one dataset, only samples where one sound class was present were used. In some cases, co-occurrence of sound classes from another datasets may have been present in a sample, but this was minimised due to the short sample length. Later, for evaluation, classifiers were trained on a maximum of 32 samples per class for each dataset, with a minimum of 10 samples from each class held out for testing. Therefore, classes with less than 42 samples in any given dataset were merged by only applying a primary label (biophony, geophony or anthrophony). Where the count of samples merged under the primary label class still did not total 42 or more samples in a given dataset, these samples were discarded. This yielded the final meta-dataset of 57,074 labelled samples, split across the four primary labels: biophony (79.20%), anthrophony (10.39%), geophony (0.09%) and ambient (10.32%), with 33 secondary labels (Supp. 1. Table S2).

Evaluating existing pretrained networks

We identified four pretrained networks widely adopted for use in acoustic transfer learning (Table 1). All four networks employ a convolutional neural network architecture. The first two were VGGish (Hershey et al., 2017) and YAMNet (Google Research), both trained on general-purpose audio datasets. VGGish was trained on the YouTube-70M dataset, a dataset consisting of 20B weak multi-label samples across 31K classes. YAMNet was trained on AudioSet, a large ontology of 2.1M human labelled acoustic events across 521 classes gathered from YouTube (Gemmeke et al., 2017). The second two were BirdNET (Kahl et al., 2021) and Perch (Ghani et al., 2023) which were both primarily trained on bird recordings from the Xeno-Canto repository. Perch was trained on the full corpus of Xeno-Canto (XC) bird recordings split into 2.9m samples 5 sec in length, and was configured with hierarchical taxonomic output heads for species, genus, family, and order classes. BirdNET was trained on a smaller set of bird classes than Perch overall, but included bird samples from the Cornell Lab of Ornithology's Macaulay Library (Macaulay, 2023) alongside 101 additional classes such as human speech, dogs and amphibian species.

For input to each network, audio samples were upsampled where required to match the input sample rate of the respective model (Table 1). As samples were shorter than the input window size of BirdNET and Perch, zero-padding was applied to the tail end of each sample. As samples were twice the length of the input window size of VGGish and YAMNet, samples were split into two windows, feature embeddings were calculated for each window and the mean across both taken.

To evaluate the transfer learning capabilities of the four pretrained networks, for each of the 16 datasets in ReefSet a pretrained network was configured with a final fully connected linear layer with corresponding output heads for the classes present in the respective dataset. This final layer was trained for 128 epochs using a batch size of 32, learning rate of 0.001 and categorical cross entropy loss. This process was repeated using 4, 8, 16 and 32 training samples per class, with ten repeats using a new random seed for the train-test split and initialization of each. For each seed, all remaining samples were set aside for testing, with a minimum of 10 per class. The mean area under the receiver operator curve (AUC-ROC) was calculated for the test set across the ten repeats for each of the four training sample counts (van Merriënboer, 2024).

Pretraining with in-domain data

State of the art bioacoustic few-shot learners commonly utilize convolutional neural network (CNN) architectures (Nolasco et al., 2022). We therefore adapted the pretraining protocol used for Perch, where an EfficientNet CNN classifier is trained and used to extract high quality feature embedding representations for transfer learning (Ghani et al., 2023). During training,

datasets were filtered to remove any samples with the primary label ‘ambient’, in order to eliminate samples that may contain unintentional positive matches for classes in other datasets. For input to the network, samples were upsampled to 32 kHz and log-mel PCEN spectrograms calculated from each (Supp. 1). Repeat padding was applied to samples, where the signal was repeated until they met the 5s input shape. Samples were shuffled and two augmentations were implemented throughout training: random normalization with a minimum and maximum gain of 0.15 and 0.25, and, MixUp with a mix in probability of 0.75 (Xu et al., 2018). The Perch network architecture was adapted to be configured with hierarchical output heads for each primary label (biophony, geophony, anthrophony), with the 35 secondary labels nested within this minus any exclusive to held-out data. All training runs were completed for 200K steps.

For the first stage of the experiment a hyperparameter sweep was performed where models were trained on 14 of the 16 datasets with two held-out for validation (Supp. 1). Learning rate, EfficientNet architecture and batch size were probed during the sweep. The core pretraining stage of the experiment was then performed using the optimum hyperparameters from the sweep (Supp. 1. Table S3).

To rigorously evulate the performance of few-shot transfer learning on unseen datasets we used a Leave-one-dataset-out (LODO) approach. During LODO, pretraining was first undertaken using 15 datasets from ReefSet, maintaining one held-out dataset for evaluation. This was repeated in all combinations one by one to produce 16 pretrained models. In the second stage of LODO, evaluation of each model was performed on its respective held-out dataset following the same few-shot transfer learning protocol as described for the existing pretrained networks. Given all data originated from reef habitats, evaluation data could be considered in-domain whilst being out of distribution.

Pretraining with cross-domain mixing

We first tested mixing the full XC Bird catalog of 2.9M samples used to train Perch with the more modest ReefSet, approximately 1.99% of the size in total sample count (Fig. 2). The XC Bird dataset was used without modifications to the original pretraining of the Perch model (Ghani et al., 2023). This mixing provided a total of 10,165 target classes, minus any exclusive to held-out reef data for LODO evaluation. To integrate the XC Bird dataset into the training procedure, the model was configured with the same output heads for ReefSet, alongside additional species, genus, family and order output heads for the XC Bird dataset, with a loss weighting of 0.1 for the latter three compared to standard heads.

Next, alongside ReefSet and the XC Bird dataset we mixed in Freesound Dataset 50K (Fonseca et al., 2022), a dataset based on the AudioSet ontology consisting of 108.2 hrs of annotated

sound events across 200 classes (Fig. 2). Freesound is comprised of audio from a more general selection of sound events, comparable to the domain used to train VGGish and YAMNet, but with fully open-source access to the audio whereas AudioSet must be scraped from YouTube. Our only adjustment to the Freesound dataset was to remove all samples with the label 'bird' (3.38% of the dataset) to mitigate overlap with more taxonomically detailed labels in the XC Bird dataset. The network was then configured following the XC Bird cross domain mixing strategy, alongside an additional set of output heads for Freesound for a total of 10'364 target classes.

During pretraining for both the domain mixing strategy experiments, individual datasets were cycled back in once all samples from them had been used once. As with the ReefSet pretraining strategy, a hyperparameter sweep was performed, with an additional parameter for dataset weighting (Supp. 1). All other components remained unchanged, with the LODO approach used for pretraining and evaluation.

Evaluating SurfPerch on novel bioacoustic domains

Using SurfPerch, the resultant network after optimising the highest performing strategy in our pretraining experiments, we mirrored the evaluation protocol used in Ghani et al. (2023) to assess the ability of Perch to generalize to novel bioacoustic domains. Novel domains originated from bird, frog, bat and marine mammal recordings. Sample counts ranging from 4 to 256 training samples per class were used, with 10 repeats of each. SurfPerch was evaluated in an "off the shelf" manner, with no hyperparameter sweeps or pretraining used to optimize for the novel bioacoustic domains being tested. These novel datasets originated from the bird, bat, frog and marine mammal domains, see Ghani et al. (2023) for further details on the data. As with the other pretraining experiments, a new network was configured with a final classification head to match the target classes for each respective dataset, which was then fine-tuned whilst the rest of the weights were kept frozen. Fine-tuning was conducted for 128 epochs using a batch size of 32, learning rate of 0.001 and categorical cross entropy loss. The fine-tuned networks were then evaluated on held out test sets from their respective dataset. Fine-tuning was performed across multiple counts of training samples per class for each dataset, ranging from 4 to 256, with ten repeats for each count using a new random seed to select the training data.

Results

Pretrained bioacoustic networks outperform networks pretrained on general audio

Mean AUC-ROC scores revealed that the pretrained networks ranked consistently across all four training sample counts. In ascending order the mean and standard deviations of these were: BirdNET (0.908 ±0.09), Perch (0.881 ±0.11), YAMNet (0.834 ±0.05), VGGish (0.813 ±0.05) (Fig. 3). These results revealed the two networks pretrained primarily on the bird domain outperformed the two trained on the more general YouTube data. As expected, the mean AUC-ROC of all pretrained models improved and standard deviation of this declined as the number of training samples per class increased from 4 through to 32 (Fig 3; Fig. S1).

Considering the constituent datasets within ReefSet on an individual basis, these presented a range of apparent difficulty and complexity (Fig. 4; Fig. S1). The datasets from Thailand, the Philippines and Indonesia, which only required binary classification between one anthropogenic and one biophonic class, generally presented easier challenges with mean AUC-ROC scores of 0.994 (±0.007), 0.960 (±0.010) and 0.935 (±0.056) respectively across all four pretrained networks and sample counts. More challenging datasets were those which required the prediction of multiple classes where samples were labelled with secondary biophony labels alongside samples labelled only with the primary biophony label class. These more challenging tasks included the Kenya, Belize and Tanzania datasets with mean AUC-ROC scores of 0.703 (±0.043), 0.791 (±0.065) and 0.812 (±0.037) respectively across all four pretrained networks and sample counts. Using just four training examples per class with BirdNET, the overall best performing pretrained network, the lowest and highest AUC-ROC scores were reported for the Kenya and Thailand datasets, with mean AUC-ROC scores of 0.746 (±0.062) and 0.996 (±0.006) respectively across the ten random seeds used for each.

Existing pretrained networks outperform pretraining on limited in-domain data

Our second experiment revealed that few-shot transfer learning capabilities of a CNN pretrained on our highly in-domain but smaller ReefSet meta-dataset were considerably lower than that of all four existing pretrained networks. This strategy reported a mean AUC-ROC score of 0.724 (±0.05) across all four training sample counts, lower than any of the pretrained networks. The ReefSet only pretraining strategy had a 200.72% and 47.80% higher AUC-ROC error (area above the ROC) than BirdNET and VGGish, the highest and lowest performing pretrained networks respectively.

Cross-domain pretraining improves generalizability

Our third experiment revealed cross-domain mixing of the small in-domain ReefSet dataset with a large set of out-of-domain bird bioacoustic data provided considerable improvements. Using the LODO pretraining and evaluation procedure once again, we observed a mean AUC-ROC score of 0.895 (±0.03) across all four training sample counts using this cross-domain

pretraining strategy (Fig. 3). This represented a notable improvement over pretraining on the in-domain ReefSet alone, which had a 163.90% higher AUC-ROC error. Importantly, this also achieved a 12.32% improvement in AUC-ROC error upon pretraining with the XC Bird dataset alone, represented by the pretrained Perch model. The only pretrained network which still outperformed this cross-domain pretraining strategy was BirdNET, with our ReefSet and XC Bird cross-domain pretraining strategy having a 12.24% higher mean AUC-ROC error.

Our fourth experiment revealed that expanding the diversity of data used in cross-domain pretraining further enhanced few-shot transfer learning capabilities on the novel coral reef domain. Using the LODO train and evaluation protocol, this triple-domain pretraining achieved the highest mean AUC-ROC scores of any strategy, with a mean AUC-ROC of 0.928 (±0.02) across all four training sample counts (Fig. 3). This represented a 31.26% reduction in error compared to cross-domain pretraining with the ReefSet and XC Bird datasets. Importantly, this strategy also outperformed BirdNET, the previously highest scoring network, which had a 21.68% higher AUC-ROC error. Using just four training samples per class, this triple-domain pretraining strategy achieved a mean AUC-ROC of 0.900 (±0.02) across the 16 datasets.

Final trials using the triple-domain strategy revealed modifications to the bias, gain and smoothing parameters of the PCEN spectrogram (Supp. 1), alongside pretraining for 1m steps, further improved performance. Following the LODO pretraining and evaluation protocol, a mean AUC-ROC score of 0.933 (±0.02) was reported using these adjustments, representing the strongest overall performance. This improvement corresponded to 26.84% and 6.60% mean improvements upon the AUC-ROC error of BirdNET and our initial triple-domain pretraining trial respectively (Supp 1. Fig. S2). Finally, we produced SurfPerch, the open-source version of this model (Supp. 2), by pretraining with this triple-domain strategy including the full ReefSet meta-dataset, using all 16 source datasets.

Targeted cross-domain pretraining does not improve generalizability to non-target bioacoustic domains

Whilst cross-domain mixing reported notable generalizability improvements to the reef domain, we observed this strategy negatively impacts performance on alternative bioacoustic domains which were not optimized for during pretraining (Fig. 5). We observed that Perch and BirdNET outperformed SurfPerch in all six of the novel domains. The lowest performance gap between SurfPerch and Perch, the overall best performing pretrained network at the novel challenges, was observed for the Godwit Calls and Watkins Marine Mammals datasets, with mean AUC-ROC scores 0.019 lower than Perch for both datasets across all training sample counts per class. The largest performance gap between SurfPerch and Perch was observed for the Yellowhammer dialect dataset, with a mean AUC-ROC score 0.084 lower than Perch across all training sample counts per class. However, SurfPerch did outperform both YAMNet and

VGGish across all datasets. As the training samples per class counts increased, the performance gap between SurfPerch and Perch decreased for each, with a difference between mean AUC-ROC scores across all datasets of 0.067 for 4 training samples per class, and 0.012 for the maximum training sample count per class (32 or 256).

Discussion

We show that by leveraging multiple domains during pretraining, we can achieve far superior transfer learning capabilities to the data-deficient domain of marine bioacoustics. In doing so, we present a novel pipeline that could resolve the considerable bottlenecks in the analysis of tropical reef and similar acoustic domains. We began by testing the transfer learning capabilities of existing pretrained networks on marine bioacoustic data. We found networks pretrained on data from bioacoustic domains outperformed those pretrained with more general audio data from unrelated domains. Next, we show these existing pretrained networks outperformed pretraining with our highly in-domain, but smaller, ReefSet meta-dataset. We then found that cross-domain mixing using the larger out-of-domain XC Bird dataset with the smaller in-domain ReefSet improved upon transfer learning capabilities of the previous strategies. Finally, we reveal that mixing together all three domains during pretraining provides the strongest performance. Importantly, however, we observe that cross-domain mixing did not improve performance on unrelated bioacoustic domains which were not optimized for during pretraining. These findings present a powerful strategy to produce pretrained networks for bioacoustic and other acoustic domains that do not currently have adequate pretrained networks for use in transfer learning.

The product of our optimum pretraining strategy for marine bioacoustic data, SurfPerch, supports accurate few-shot learning. Furthermore, this was evaluated by only fine-tuning the final layer of the network during transfer learning. This combined few-shot transfer learning protocol therefore significantly reduces both the annotation and computational costs required to build accurate classifiers, enabling end-users to perform this on standard personal computing devices using a significantly reduced set of annotations. Example uses of tropical reef bioacoustic analyses that could be accelerated or scaled using SurfPerch include reef health assessments (Jarriel et al., 2024), tracking reef restoration success (Lamont et al., 2022), measuring marine protected area outcomes (Manna et al., 2021), or understanding fundamental processes within the soundscape such as temporal patterns and biogeographical variability (Duane et al., 2024; Raick et al., 2023). Given its potential applications, we have provided an interactive demonstration on how to implement this approach on new data entirely from a web browser (Supplementary 2), including the incorporation of an agile modeling protocol which can be used to further boost classifier performance by identifying the most relevant samples for annotation (Hamer et al., 2023). A simplified workflow for fine tuning on new data is detailed in Fig. S3.

The results we present here provide key advancements on pretraining for bioacoustics. The performance of existing pretrained networks supports similar work in Ghani et al. (2023) which reported networks pretrained on large and diverse bird bioacoustic datasets generalize better to other bioacoustic domains than those pretrained on more general audio. The reasons behind this remain an open research question, this could be due to common properties between bioacoustic domains, or, the high innate acoustic complexity and variety of bird vocalizations compared to AudioSet. The increased volume of training data also likely contributed to the improved performance of our two- and three-way cross-domain mixing strategies. However, class diversity has been empirically demonstrated as a more significant driver of generalization in multiple settings (Dhillon et al., 2020; Dumoulin et al., 2021; Luo et al., 2023). Both bioacoustic networks were pretrained on a far larger diversity of classes compared to YAMNet, whereas VGGish was trained on the largest diversity of classes but these were weakly labelled (Table 1). Whilst BirdNET was trained on a lower class diversity, its outperformance of Perch provides potential evidence that cross-domain mixing during pretraining is also a key factor in enhancing generalizability. BirdNET's (v2.3) pretraining included invertebrate, amphibian, mammal and anthropogenic sound classes. Future experiments controlling for data volume and domain diversity across datasets could help disentangle the contributions of class and domain diversity versus data quantity further. Lastly, given zero-padding was used to lengthen ReefSet samples during transfer learning to match the input length of Perch and BirdNET, the latter's shorter fixed window length may have been favourable due to reduced padding. Where padding is necessary due to cross-domain mixing, future experiments could compare models using repeat-padding versus zero-padding.

Our findings present guidance for future users aiming to leverage deep neural networks for bioacoustics. Pretraining experiments such as those presented here are inherently computationally expensive. The experiments outlined total to the training of 139 networks, on average requiring ~20 hrs on a TPUv3 pod, the equivalent of 26 '668 USD using Google Cloud spot instances at the time of writing. By comparison, fine tuning a model is arbitrary, taking <1 min on CPU with 128 training samples using a standard personal laptop. Bioacousticians with novel datasets will therefore benefit most from identifying an existing pretrained network most relevant to their domain, or testing a suite of existing networks as presented in our first experiment. Here, we show SurfPerch presents the strongest option for coral reefs and likely related aquatic domains, whereas Perch, trained exclusively on bird data, presents a better option for unrelated bioacoustic domains. Interestingly, Perch still outperformed SurfPerch on the Watkins marine mammal dataset. Many of the sounds in the Watkins dataset were in fact terrestrial marine mammal vocalizations which may partially explain this (Sayigh et al., 2016). Additionally, marine mammal calls typically consist of multiple high frequency phonemes (Erbe et al., 2017), potentially making these more comparable to the bird domain than the single phenome sounds characteristic of most fish vocalizations (e.g grunts, pops).

Where an adequate pretrained network is not available and domain specific data is sparse, we show cross-domain pretraining presents a valuable strategy to develop a suitable network for

the target domain. Future work may be able to improve upon the strategies tested here. Increasing the volume of high quality in-domain training data is typically the most valuable tool for improving model performance (Halevy et al., 2009). Relevant sound libraries to the coral reef domain with potential for growth include the FishSounds platform (Looby et al., 2023) and the proposed Global Library of Underwater Biological Sounds (Parsons et al., 2022). Furthermore, growing open-source sound event libraries are available, meaning additional bioacoustic and unrelated acoustic domains could be integrated, (Cañas et al., 2023; Humphrey et al., 2018; Mac Aodha et al., 2018). Ongoing efforts to release updated versions of existing bioacoustic networks, which incorporate an increased diversity of data from both bioacoustic and other domains into pretraining, will likely improve their generalizability to novel bioacoustic domains. Indeed, updated versions such as BirdNET v2.4, which has been trained on over twice the number of classes, could be used to validate this. More broadly, mixing data from a diverse set of domains and optimizing for a range of these during evaluation, rather than the single domain we targeted here, may present a route to developing improved foundational bioacoustic and acoustic models for wider contexts. While we restricted our pretraining in the present work to fully open-source datasets, incorporating AudioSet in its entirety, with its greater data volume and class diversity, could provide a direct route to improving generalisability during transfer learning for bioacoustic models. Additionally, including bioacoustic data in the pretraining of industry-standard models may offer reciprocal improvements in their generalizability.

Beyond enhancing the data used for training, future work could also explore methodological improvements. Marine PAM data is inherently noisy, with recordings typically comprised of multiple sources from biophony, geophony and anthropogenic components (Mooney et al., 2020). Implementing an unsupervised source separation model presents a proven tool for improving classification in noisy bioacoustic datasets, through disentangling individual signals from the broader soundscape (Denton et al., 2022; Lin & Tsao, 2019). Elsewhere, self supervised learning (SSL) presents a tool which can be used to learn informative features from unlabelled data, enabling it to exploit vast un-annotated datasets (Moummad et al., 2023). Future work could benchmark pretraining on larger in-domain datasets with SSL against cross-domain pretraining of supervised classifiers. Other changes during the transfer learning component could boost performance. Firstly, alternative lightweight classifiers (e.g two layer models, random forests) could be tested. Augmentations are another proven way to improve classification accuracy with limited training data (Stowell, 2022). Better still, agile modeling, alternatively known as active learning, can be integrated into the pipeline using a human in the loop to identify the most informative training samples (Hamer et al., 2023; Stretcu et al., 2023).

In conclusion, we leveraged cross-domain pretraining to develop a powerful tool that supports automated analysis of marine PAM recordings with low annotation and computational costs for the end-user. Our findings offer insights into replicating this for other acoustic domains where existing pretrained networks are inadequate. Combining efficient machine learning analysis such as ours with the vast scales at which PAM data can be collected has a significant potential

to boost our understanding and monitoring capacity of global biodiversity. We anticipate these technologies will facilitate the expansion of scientific frontiers towards new applications and challenges so far unrealized.

Acknowledgments

BW was supported by a FSBI PhD Studentship at UCL and ZSL, alongside a PhD Student Researcher placement at Google DeepMind. We wish to acknowledge multiple parties for supporting the collection of data used in this study:

- The United States Virgin Islands datasets were collected in a collaboration between Sound Ocean Science and local partner Corina Marks at Thriving Islands under a USVI Scientific Research Permit: DFW22021X.
- The Mozambique dataset was collected in a collaboration between Sound Ocean Science and local partners Dr Mario LeBrato and Karen Bowles at the Bazaruto Center for Scientific Studies under a Department of Conservation Permit: 04/GDG/ANAC/MTA/2020.
- The Tanzanian dataset was collected in a collaboration between Sound Ocean Science and local partners Dr Mario LeBrato and Karen Bowles Ulli Kloiber and Omar Nyange at Chumbe at Island Coral Park, Zanzibar, Tanzania under under CHICOP Zanzibar research permits.
- The Thailand dataset was collected under a citizen science program operated by Black Turtle Dive, Thailand.
- The Florida boats dataset was gathered under a Special Activity License (license number: SAL-21-1798-SRP) granted by the Florida Fish and Wildlife Conservation Commission. Funding was provided by a Donald R Nelson Behaviour Research Award to CW by the American Elasmobranch Society.
- The Kenyan dataset was collected in a collaboration between Mars Global and local partners Angus Roberts and Viola Roberts at the Ocean Trust, with permissions from the Lamu County Department of Fisheries.
- Soundscape data from Indonesia were collected as part of the monitoring program for the Mars Coral Reef Restoration Project, in collaboration with Universitas Hasanuddin. We thank Lily Damayanti, Pippa Mansell, David Smith and the Mars Sustainable Solutions team for support with fieldwork logistics. We also thank the Department of Marine Affairs and Fisheries of the Province of South Sulawesi, the Government Offices of the Kabupaten of Pangkep, Pulau Bontosua and Pulau Badi, and the communities of Pulau Bontosua and Pulau Badi for their support. B.W.'s fieldwork in Indonesia was conducted under an Indonesian national research permit issued by BRIN (number 109A/SIP/IV/FR/3/2023), with T.B.R. as the permit's Indonesian researcher/counterpart, and associated ethical approval given by BRIN. We thank Prof J. Jompa and Prof R.A. Rappe at Universitas Hasanuddin for logistical assistance with permit and visa applications.

- Remaining datasets were gathered in a collaboration between Conservation Metrics Inc. and the Cornell Lab of Ornithology with funding support from Oceankind and the Cornell Atkinson Center for Sustainability, Cornell Lab of Ornithology.

References

1. Pimm, S.L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R. & Loarie, S. Emerging technologies to conserve biodiversity. *Trends. Ecol. Evol.* **30**, 685-696 (2015).
2. Gibb, R., Browning, E., Glover-Kapfer, P. & Jones, K.E. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods Ecol. Evol.* **10**, 169-185 (2019).
3. Ross, S.R.J., O'Connell, D.P., Deichmann, J.L., Desjonquères, C., Gasc, A., Phillips, J.N., Sethi, S.S., Wood, C.M. & Burivalova, Z. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Func. Eco.* **37**, 959-975 (2023).
4. Hill, A.P., Prince, P., Piña Covarrubias, E., Doncaster, C.P., Snaddon, J.L. & Rogers, A. AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods Ecol. Evol.* **9**, 1199-1211 (2018).
5. Lamont, T.A., Chapuis, L., Williams, B., Dines, S., Gridley, T., Frainer, G., Fearey, J., Maulana, P.B., Prasetya, M.E., Jompa, J. & Smith, D.J. HydroMoth: Testing a prototype low-cost acoustic recorder for aquatic environments. *Rem. Sens. Ecol. Cons.* **8**, 362-378 (2022).
6. Shonfield, J. & Bayne, E.M. Autonomous recording units in avian ecological research: current use and future applications. *Avian. Cons. Ecol.* **12** (2017).
7. Stowell, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* **10**, 13152 (2022).
8. Kholghi, M., Phillips, Y., Towsey, M., Sitbon, L. & Roe, P. Active learning for classifying long-duration audio recordings of the environment. *Methods Ecol. Evol.* 9, 1948-1958 (2018).
9. Ghani, B., Denton, T., Kahl, S. & Klinck, H. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Sci. Rep.* **13**, 22876 (2023).
10. Kahl, S., Wood, C.M., Eibl, M. & Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecol. Inform*. **61**, 101236 (2021).
11. Pérez-Granados, C. BirdNET: applications, performance, pitfalls and future opportunities. Ibis **165**, 1068-1075 (2023).
12. White, E.L., Klinck, H., Bull, J.M., White, P.R. & Risch, D. One size fits all? Adaptation of trained CNNs to new marine acoustic environments. *Ecol. Inform.* **78**, 102363 (2023).
13. Kath, H., Serafini, P.P., Campos, I.B., Gouvêa, T.S. & Sonntag, D. Leveraging transfer learning and active learning for data annotation in passive acoustic monitoring of wildlife. *Ecol. Inform*. ***82***, 102710. (2024)

14. Nolasco, I., Singh, S., Vidana-Villa, E., Grout, E., Morford, J., Emmerson, M., Jensens, F., Whitehead, H., Kiskin, I., Strandburg-Peshkin, A. & Gill, L. Few-shot bioacoustic event detection at the dcase 2022 challenge. https://doi.org/10.48550/arXiv.2207.07911 (2022).
15. Williams, B., Belvanera, S.M., Sethi, S.S., Lamont, T.A., Jompa, J., Prasetya, M., Richardson, L., Weschke, E., Hoey, A., Beldade, R. & Mills, S.C. Unlocking the soundscape of coral reefs with artificial intelligence. https://doi.org/10.1101/2024.02.02.578582 (2024).
16. Kaplan, M.B., Lammers, M.O., Zang, E. & Aran Mooney, T. Acoustic and biological trends on coral reefs off Maui, Hawaii. *Coral Reefs* **37**, 121-133 (2018).
17. McWilliam, J.N., McCauley, R.D., Erbe, C. & Parsons, M.J. Soundscape diversity in the Great Barrier Reef: Lizard Island, a case study. *Bioacoustics **27***, 295-311 (2018).
18. Knowlton, N., Brainard, R.E., Fisher, R., Moews, M., Plaisance, L. & Caley, M.J. Coral reef biodiversity. Life in the world's oceans: diversity distribution and abundance Ch. 4 (Wiley-Blackwell, 2010).
19. Hoegh-Guldberg, O., Pendleton, L. & Kaup, A. People and the changing nature of coral reefs. *Reg Stud. Mar. Sci.* **30**, 100699 (2019).
20. Eddy, T.D., Lam, V.W., Reygondeau, G., Cisneros-Montemayor, A.M., Greer, K., Palomares, M.L.D., Bruno, J.F., Ota, Y. & Cheung, W.W. Global decline in capacity of coral reefs to provide ecosystem services. *One Earth* **4**, 1278-1285 (2021).
21. Intergovernmental Panel on Climate Change (IPCC) 'Summary for Policymakers', in Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty. 1–24 (Camb. Univ. Press, 2022).
22. Pysanczyn, J.W., Williams, E.A., Brodrick, E., Robert, D., Craggs, J., Marhaver, K.L. & Simpson, S.D. The role of acoustics within the sensory landscape of coral larval settlement. *Front. Mar Sci.* **10**, 1111599 (2023).
23. Mooney, T.A., Di Iorio, L., Lammers, M., Lin, T.H., Nedelec, S.L., Parsons, M., Radford, C., Urban, E. & Stanley, J. Listening forward: approaching marine biodiversity assessments using acoustic methods. *Roy. Soc. Open Sci.* **7**, 201287 (2020).
24. Parsons, M.J., Lin, T.H., Mooney, T.A., Erbe, C., Juanes, F., Lammers, M., Li, S., Linke, S., Looby, A., Nedelec, S.L. & Van Opzeeland, I. Sounding the call for a global library of underwater biological sounds. *Front. Eco. Evol*. **10**, 39 (2022).
25. Rountree, R. A., Bolgan, M. & Juanes, F. How can we understand freshwater soundscapes without fish sound descriptions? *Fisheries* **44**, 137-143 (2019).

26. Lin, T.H. Tsao, Y. & Akamatsu, T. Comparison of passive acoustic soniferous fish monitoring with supervised and unsupervised approaches. *J. Acoust. Soc. Am* **143,** EL278-EL284 (2018).
27. Ozanich, E., Thode, A., Gerstoft, P., Freeman, L.A. & Freeman, S. Deep embedded clustering of coral reef bioacoustics. *J. Acoust. Soc. Am.* **149**, 2587-2601 (2021).
28. Williams, B., Lamont, T.A., Chapuis, L., Harding, H.R., May, E.B., Prasetya, M.E., Seraphim, M.J., Jompa, J., Smith, D.J., Janetski, N. & Radford, A.N. Enhancing automated analysis of marine soundscapes using ecoacoustic indices and machine learning. *Eco. Ind.* **140**, 108986 (2022).
29. Xeno-canto Foundation and Naturalis Biodiversity Center. xeno-canto. https://xeno-canto.org. (2023).
30. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B. & Slaney, M. CNN architectures for large-scale audio classification. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) 131-135 (2017).
31. YAMNet. Google Research. https://www.kaggle.com/models/google/yamnet/.
32. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M. & Ritter, M. Audio set: An ontology and human-labelled dataset for audio events. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 776-780). IEEE.
33. The Cornell Lab of Ornithology. Macaulay library. https://www.macaulaylibrary.org. (2023).
34. van Merriënboer, B., Hamer, J., Dumoulin, V., Triantafillou, E. and Denton, T. Birds, bats and beyond: Evaluating generalization in bioacoustics models. *Front. Bird Sc.* ***3***, 1369756 (2024)
35. Jarriel, S.D., Formel, N., Ferguson, S,R., Jensen F, H., Apprill, A. & Mooney, T,A., Unidentified fish sounds as indicators of coral reef health and comparison to other acoustic methods. Front. Remote Sens. **5**, 1338586 (2024).
36. La Manna, G., Picciulin, M., Crobu, A., Perretti, F., Ronchetti, F., Manghi, M., Ruiu, A. & Ceccherelli, G., 2021. Marine soundscape and fish biophony of a Mediterranean marine protected area. *PeerJ* ***9***, 12551 (2021).
37. Raick, X., Di Iorio, L., Lecchini, D., Gervaise, C., Hédouin, L., Under The Pole Consortium Bardout G. Fauchet J. Ferucci A. Gazzola F. Lagarrigue G. Leblond J. Marivint E. Mittau A. Mollon N. Paulme N. Périé-Bardout E. Pete R. Pujolle S. Siu G., Perez-Rosales, G., Rouze, H., Bertucci, F. and Parmentier, E. Fish sounds of photic and mesophotic coral reefs: variation with depth and type of island. *Coral Reefs* ***42***, 285-297 (2023).
38. Duane, D., Freeman, S. and Freeman, L. Moonlight-driven biological choruses in Hawaiian coral reefs. *Plos One* ***19***, 0299916 (2024).

39. Hamer, J., Laber, R. & Denton, T. Agile Modeling for Bioacoustic Monitoring. In: NeurIPS 2023 Workshop: Tackling Climate Change with Machine Learning. https://colab.research.google.com/drive/1gPBu2fyw6aoT-zxXFk15I2GObfMRNHUq?usp=sharing (2023).
40. Dhillon, G.S., Chaudhari, P., Ravichandran, A. & Soatto, S. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729* (2019).
41. Dumoulin, V., Houlsby, N., Evci, U., Zhai, X., Goroshin, R., Gelly, S. & Larochelle, H. "A unified few-shot classification benchmark to compare transfer and meta learning approaches." In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021.
42. Luo, X., Wu, H., Zhang, J., Gao, L., Xu, J. & Song, J. A closer look at few-shot classification again. International Conference on Machine Learning 23103-23123. PMLR. (2023).
43. Fonseca, E., Favory, X., Pons, J., Font, F. & Serra, X. Fsd50k: an open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 829-852 (2021).
44. Sayigh, L., Daher, M.A., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C. & Tyack, P. The Watkins marine mammal sound database: an online, freely accessible resource. *Proc. Meet. Acoust.* **27** (2016).
45. Erbe, C., Dunlop, R., Jenner, K.C.S., Jenner, M.N.M., McCauley, R.D., Parnum, I., Parsons, M., Rogers, T. & Salgado-Kent, C. Review of underwater and in-air sounds emitted by Australian and Antarctic marine mammals. *Acoust. Aust.* **45**, 179-241 (2017).
46. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. IEEE intelligent systems **24**, 8-12 (2009).
47. Looby, A., Vela, S., Cox, K., Riera, A., Bravo, S., Davies, H.L., Rountree, R., Reynolds, L.K., Martin, C.W., Matwin, S. & Juanes, F. FishSounds Version 1.0: A website for the compilation of fish sound production information and recordings. *Ecol. Infor.* **74**, 101953 (2023).
48. Cañas, J.S., Toro-Gómez, M.P., Sugai, L.S.M., Restrepo, H.D.B., Rudas, J., Bautista, B.P., Toledo, L.F., Dena, S., Domingos, A.H.R., de Souza, F.L. & Neckel-Oliveira, S. AnuraSet: A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring. *Sci. Data.* **10**, 771 (2023).
49. Humphrey, E., Durand, S. & McFee, B. OpenMIC-2018: An Open Data-set for Multiple Instrument Recognition. *ISMIR* 438-444 (2018).
50. Mac Aodha, O., Gibb, R., Barlow, K.E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G.R., Newson, S.E. & Pandourski, I. Bat detective—Deep learning tools for bat acoustic signal detection. *PLoS Comput. Biol.* **143**, 1005995 (2018).

51. Denton, T., Wisdom, S. & Hershey, J.R. Improving bird classification with unsupervised sound separation. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 636-640 (IEEE, 2022).
52. Lin, T.H. & Tsao, Y. Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval. *Rem. Sens. Ecol. Cons*, **6**, 236-247 (2020).
53. Moummad, I., Serizel, R. and Farrugia, N. Self-Supervised Learning for Few-Shot Bird Sound Classification. https://arxiv.org/html/2312.15824v3 (2023).
54. Stretcu, O., Vendrow, E., Hata, K., Viswanathan, K., Ferrari, V., Tavakkol, S., Zhou, W., Avinash, A., Luo, E., Alldrin, N.G. & Bateni, M. Agile modeling: From concept to classifier in minutes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 22323-22334 (2023).
55. Xu, K., Feng, D., Mi, H., Zhu, B., Wang, D., Zhang, L., Cai, H. & Liu, S. Mixup-based acoustic scene classification using multi-channel convolutional neural network. In Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III 19, 14-23. Springer International Publishing. (2018).

Tables

Table 1. Details of the four pretrained networks used to evaluate transfer learning performance on ReefSet. Real-time-factor inference speed reflects how many times faster each network is at processing the audio's real time duration on a CPU, further details are in Supp. 1.

Network	Training domain	Number of training classes	Input sample rate (kHz)	Input length (sec)	Embedding dimension	Parameter count	Real-time-factor inference speed (CPU)
VGGish	AudioSet (YouTube)	31000	16	0.96	128	72.1M	41.1
YAMNet	AudioSet (YouTube)	521	16	0.96	1024	4.7M	86.04
BirdNET v2.3	Bioacoustic (primarily birds)	3337	48	3	1024	10.4M	260.68
Perch v1.4	Bioacoustic (birds)	10932	32	5	1280	80.1M	39.41

Figures

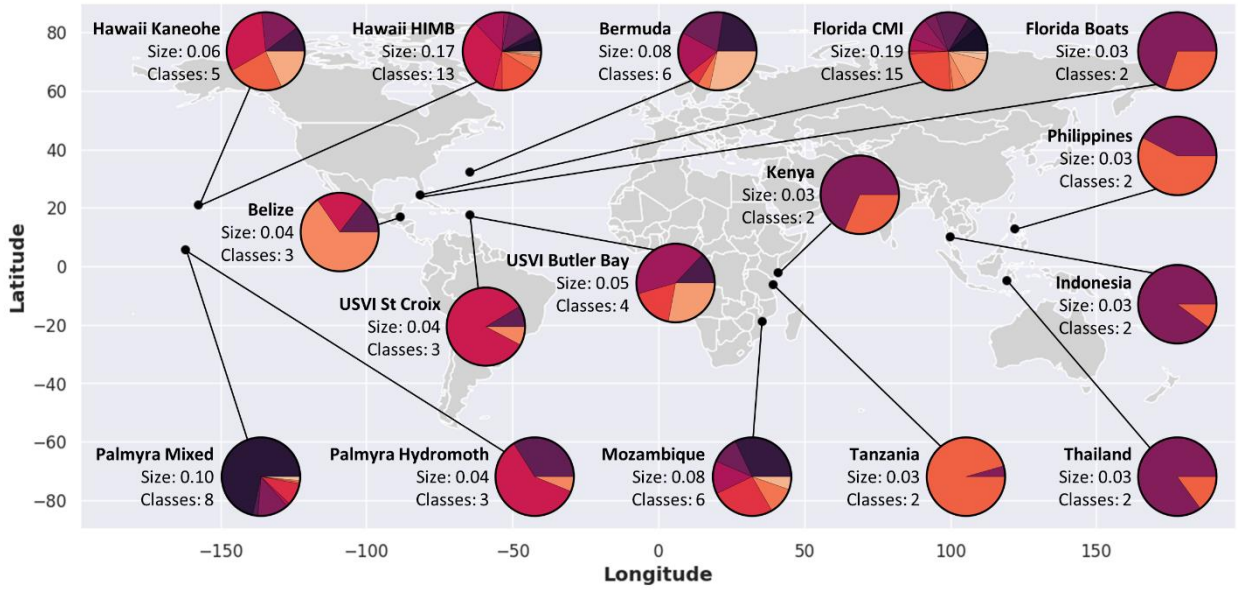

Fig. 1. Datasets used to assemble ReefSet. Size indicates the relative size of each dataset to ReefSet, summing to one. Classes indicates the number of unique labels within each dataset. Pie charts indicate the distribution of labels within the dataset, with colours set independently for each dataset.

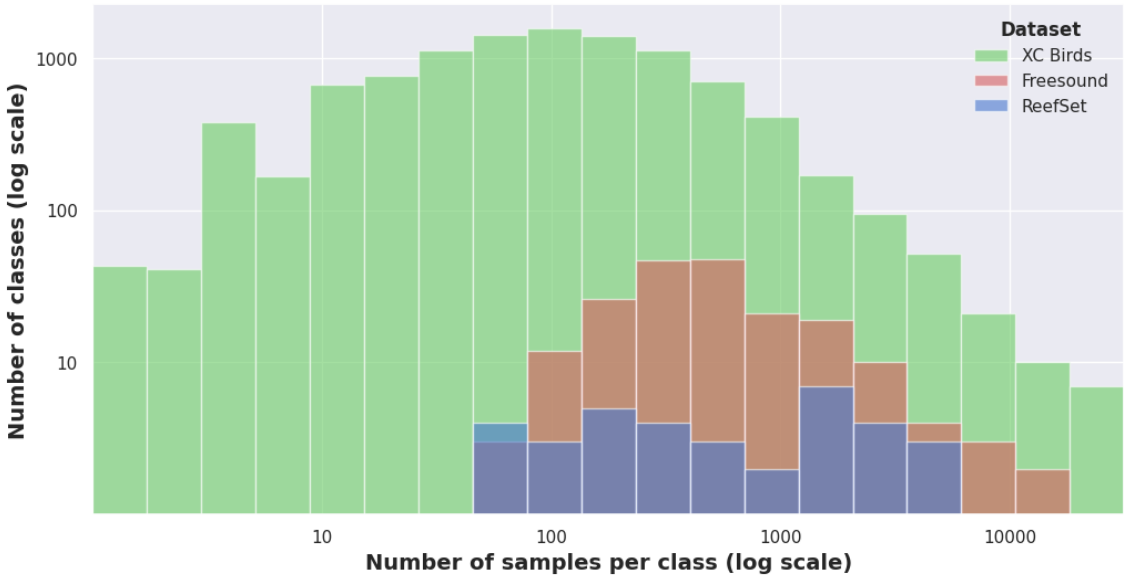

Fig. 2. Histogram of counts by class for the three datasets used for pretraining: XC Birds, Freesound and ReefSet. Bins are logarithmically spaced based on the range of counts in the XC Bird data, with bin edges determined by creating 20 equal intervals on a logarithmic scale between the minimum and maximum counts observed in the XC Bird dataset.

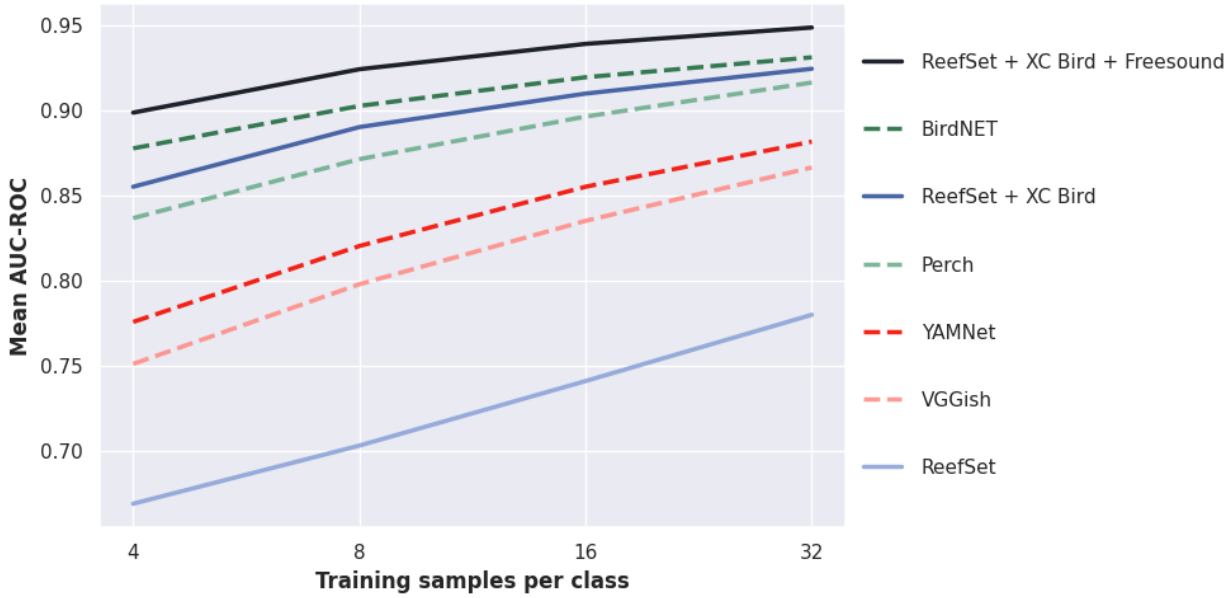

Fig. 3. Mean AUC-ROC scores reported by transfer learning evaluation for each model across all 16 datasets within ReefSet. Dashed lines represent existing pretrained networks, with names indicated in the legend. Solid lines represent the three alternative pretraining strategies for our model, with the data used for pretraining indicated in the legend.

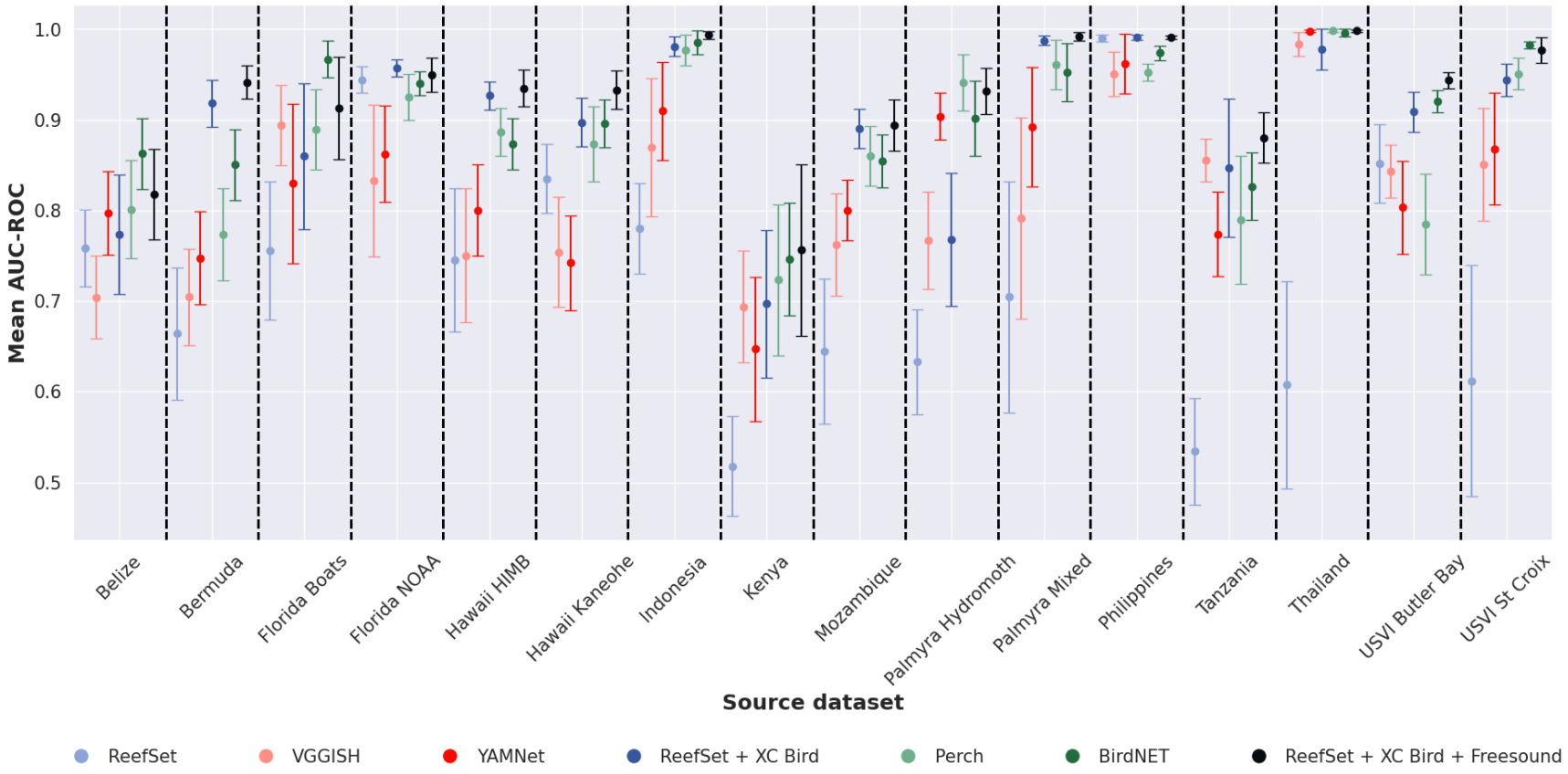

Fig. 4. Mean AUC-ROC scores reported by transfer learning evaluation on each dataset within ReefSet across all training samples per class counts used (4, 8, 16 and 32). Points represent the mean, lines represent standard deviation. Within each dataset bin, models are ordered by overall mean performance across all dataset and training sample counts, going from weakest (left) to strongest (right). In the legend, existing pretrained networks are indicated by name, whereas our alternative pretraining strategies are indicated by the datasets used during pretraining.

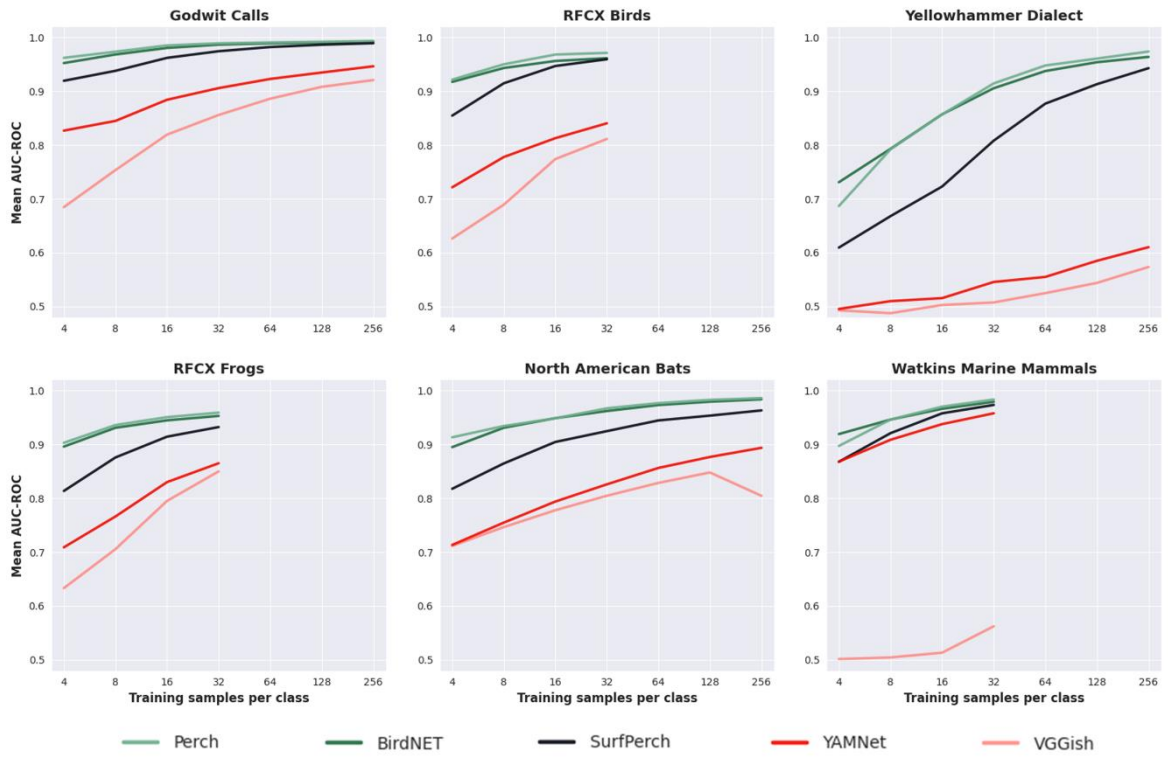

Fig. 5. Mean AUC-ROC scores reported from transfer learning evaluation of each model on six novel bioacoustic datasets. Each model and training sample per class count combination were repeated across ten random seeds.