Received: 19 August 2024 | Accepted: 31 May 2025

DOI: 10.1111/2041-210X.70089

BRITISH ECOLOGICAL SOCIETY Methods in Ecology and Evolution

RESEARCH ARTICLE

Conservation, Ecology and Artificial Intelligence: Advances and Symbiotic Solutions

# Simulated soundscapes and transfer learning boost the performance of acoustic classifiers under data scarcity

Matthew J. Weldy[1,2,3] | Damon B. Lesmeister[2] | Tom Denton[4] | Adam Duarte[5] | Ben J. Vernasco[3,5,6] | Amandine Gasc[7] | Jennifer C. Rowe[8] | Michael J. Adams[8] | Matthew G. Betts[1]

[1]Department of Forest Ecosystems and Society, Oregon State University, Corvallis, Oregon, USA; [2]Pacific Northwest Research Station, USDA Forest Service, Corvallis, Oregon, USA; [3]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA; [4]Google DeepMind, Oakland, California, USA; [5]Pacific Northwest Research Station, USDA Forest Service, Olympia, Washington, USA; [6]Whitman College, Department of Biology, Walla Walla, Washington, USA; [7]Institut Méditerranéen de Biodiversité et d'Ecologie Marine et Continentale, Marseille, France and [8]U.S. Geological Survey, Forest and Rangeland Ecosystem Science Center, Corvallis, Oregon, USA

**Correspondence**
Matthew J. Weldy
Email: matthewjweldy@gmail.com

**Funding information**
USDA Forest Service Pacific Northwest Research Station; Oregon State University Department of Forest Ecosystems and Society; USDA Forest Service Collaborative Forest Landscape Restoration Program (CFLRP)

**Handling Editor:** Sara Beery

## Abstract

1. The biodiversity crisis necessitates spatially extensive methods to monitor multiple taxonomic groups for evidence of change in response to evolving environmental conditions. Programs that combine passive acoustic monitoring and machine learning are increasingly used to meet this need. These methods require large, annotated datasets, which are time-consuming and expensive to produce, creating potential barriers to adoption in data- and funding-poor regions. Recently released pre-trained avian acoustic classification models provide opportunities to reduce the need for manual labelling and accelerate the development of new acoustic classification algorithms through transfer learning. Transfer learning is a strategy for developing algorithms under data scarcity that uses pre-trained models from related tasks to adapt to new tasks.
2. Our primary objective was to develop a transfer learning strategy using the feature embeddings of a pre-trained avian classification model to train custom acoustic classification models in data-scarce contexts. We used three annotated avian acoustic datasets to test whether transfer learning and soundscape simulation-based data augmentation could substantially reduce the annotated training data necessary to develop performant custom acoustic classifiers. We also conducted a sensitivity analysis for hyperparameter choice and model architecture. We then assessed the generalizability of our strategy to increasingly novel non-avian classification tasks.
3. With as few as two training examples per class, our soundscape simulation data augmentation approach consistently yielded new classifiers with improved performance relative to the pre-trained classification model and transfer learning

classifiers trained with other augmentation approaches. Performance increases were evident for three avian test datasets, including single-class and multi-label contexts. We observed that the relative performance among our data augmentation approaches varied for the avian datasets and nearly converged for one dataset when we included more training examples.

4. We demonstrate an efficient approach to developing new acoustic classifiers leveraging open-source sound repositories and pre-trained networks to reduce manual labelling. With very few examples, our soundscape simulation approach to data augmentation yielded classifiers with performance equivalent to those trained with many more examples, showing it is possible to reduce manual labelling while still achieving high-performance classifiers and, in turn, expanding the potential for passive acoustic monitoring to address rising biodiversity monitoring needs.

KEYWORDS

amphibian sounds, avian sounds, biodiversity monitoring, data scarcity, few-shot learning, mammal sounds, orthopteran sounds, passive acoustic monitoring

## 1 | INTRODUCTION

Passive acoustic monitoring (PAM) is an efficient and non-invasive sensor-based sampling approach that can be used to simultaneously collect data for multiple species (Shonfield & Bayne, 2017). However, large data volumes and challenges associated with correctly identifying target sounds have hindered the widespread adoption of PAM (Gibb et al., 2019; Hartig et al., 2023). Recent advances in machine learning offer a promising path towards semi-automated detection and classification of animal vocalizations (Stowell, 2022); however, these algorithms require iterative training over large annotated datasets, which take substantial resources, effort and specialized knowledge to construct. Building training datasets is a significant challenge for new PAM programmes, especially in highly biodiverse regions where conservation needs are high, but financial resources for labelling species vocalizations are low (Cui et al., 2023).

When faced with these challenges, PAM managers typically adopt one of three approaches based on project objectives and available resources. First, project support staff can exhaustively search PAM recordings for target vocalizations. Second, projects may use existing classification and detection algorithms, such as BirdNET (Kahl et al., 2021), Perch (Ghani et al., 2023) or PNW-Cnet (Ruff et al., 2023), and dedicate time to reviewing predictions and calibrating outputs. This involves fine-tuning species-specific detection thresholds and adjusting sampling protocols to manage the risk of false-positive and false-negative predictions (Cole et al., 2022; Wood & Kahl, 2024). Third, some projects may invest considerable resources to annotate acoustic files and train custom local classification algorithms from scratch (Gaylord et al., 2023; Ruff et al., 2023).

Alternatively, PAM managers can adopt an intermediate transfer learning approach that leverages both pre-trained models and local data to iteratively develop custom acoustic classifiers tailored to specific project needs. Transfer learning is a machine learning technique that adapts the knowledge from pre-trained models to improve performance on related tasks (Pan & Yang, 2010). A common approach to transfer learning uses the penultimate layer output from a pre-trained model—an 'embedding'—as input to train a new classifier for a different task (Oquab et al., 2014). This approach has been used extensively in computer vision tasks (Kornblith et al., 2019), acoustic classification (Kong et al., 2020) and natural language processing (Houlsby et al., 2019). In ecological acoustic classification, transfer learning has recently shown promise (Ghani et al., 2023; Incze et al., 2018; Nolasco et al., 2023), and pre-trained acoustic classification models like BirdNET (Kahl et al., 2021) and Perch (Ghani et al., 2023) offer strong pre-trained model options. For example, Dufourq et al. (2022) used pre-trained image classification models to develop new acoustic classifiers for Hainan gibbon (*Nomascus hainanus*), black-and-white ruffed lemur (*Varecia variegata*), thyolo alethe (*Chamaetylas choloensis*) and pin-tailed whydah (*Vidua macroura*). Transfer learning remains underutilized in PAM due to its novelty in ecology, perceived complexity and the misperception that substantial training data are required; however, improved transfer learning strategies could potentially shorten the development cycle and lead to increased adoption of active learning frameworks (Zhao et al., 2020).

Here, we test whether simulated soundscape data augmentation can be used with transfer learning-based acoustic classifiers to substantially reduce training data requirements. We used three annotated avian acoustic datasets, ranging in complexity from 1 to 10 classes. We compared transfer learning-based classifier performance under different model architectures, explored a range of hyperparameter choices and demonstrated that the simulated soundscape approach outperforms simpler data augmentation techniques. We test the performance of our custom acoustic classifiers relative to the baseline classification performance of two pre-trained avian classification models, BirdNET version 2.4 which targets a global set of species (Kahl et al., 2021) and PNW-Cnet version 4 a local model developed to support Northwest Forest Plan PAM in the Pacific Northwest, USA (Lesmeister & Jenkins, 2022; Ruff et al., 2023). Last, we explore the generalizability of our approach with three non-avian annotated acoustic datasets, ranging in complexity from 1 to 11 classes that vary in overlap with BirdNET's training data.

## 2 | MATERIALS AND METHODS

### 2.1 | Datasets

We assembled six acoustic evaluation datasets, independent of training data for BirdNET version 2.4, ranging in complexity from a single-class classification problem to an 11-class multi-label classification problem. Three datasets, which we refer to as avian datasets, were used to develop strategies for transfer learning under data scarcity (Table 1). The other three datasets, which we refer to as non-avian datasets, were used to explore the generalizability of our strategies when applied to non-avian taxa. The evaluation datasets were all annotated at the recording level (e.g. one annotation per recording that indicates the presence of at least one vocalization from a species). All the audio examples were resampled to 48 kHz using Librosa (version 0.10.0; McFee et al., 2023) and, in some cases, reduced from stereo to mono.

### 2.1.1 | Avian datasets

We downloaded recordings of our avian target species from XenoCanto spatially filtered to the Pacific Northwest Region between latitudes from 37.730 to 49.030° N and longitudes ranging from 125.000 to 120.500° W, overlapping the sampling area of the PAM datasets. The XenoCanto recordings are focal recordings that feature specific target species and do not typically differentiate calls from songs or annotate background non-target species sounds (Figure 1a; van Merriënboer et al., 2024; Vellinga & Planque, 2015). In addition, the XenoCanto recordings are annotated at the recording level, which vary in length from a few seconds to multiple minutes. We annotated the XenoCanto recordings using predefined target vocalizations (Table 1; Vellinga & Planque, 2015). We then developed two independent annotated evaluation datasets for each avian dataset: one with 3-s sound

**TABLE 1** Descriptions of the avian evaluation datasets.

| Dataset | Species | Sonotype | Call description | No. of examples | No. of evaluation clips | |
|---|---|---|---|---|---|---|
| | | | | | 3 s | 12 s |
| Marbled murrelet | Marbled murrelet | marmur_call_1 | ‘keer’ | 40 | 345 | 1000 |
| | | non-target | | | 216 | 1000 |
| Blue mountains | Clark's nutcracker | clanut_call_1 | ‘kraa’ | 53 | 100 | 387 |
| | | clanut_call_2 | ‘keer’ | 39 | 100 | 62 |
| | | clanut_call_3 | ‘reek’ | 41 | 100 | 1056 |
| | American goshawk | norgos_call_1 | Screech series | 27 | 99 | 229 |
| | | norgos_call_2 | Wail | 18 | 100 | 496 |
| | White-headed woodpecker | whhwoo_call_1 | ‘pittik’ | 69 | 100 | 1034 |
| | | non-target | | | 0 | 0 |
| Passerine | Hermit thrush | herthr_song_1 | Song | 42 | 137 | 517 |
| | | herthr_call_2 | Whine | 36 | 8 | 62 |
| | Olive-sided flycatcher | olsfly_song_1 | Song | 50 | 128 | 114 |
| | | olsfly_call_1 | ‘pip’ series | 40 | 100 | 69 |
| | Spotted towhee | spotow_song_1 | Song | 278 | 22 | 125 |
| | Swainson's thrush | swathr_song_1 | Song | 42 | 134 | 331 |
| | | swathr_call_1 | ‘pwut’ | 127 | 42 | 142 |
| | | swathr_call_3 | ‘wee’ | 101 | 44 | 195 |
| | Varied thrush | varthr_song_1 | Song | 101 | 51 | 500 |
| | Wrentit | wrenti_song_1 | Song | 55 | 13 | 54 |
| | | non-target | | | 33 | 2232 |

*Note*: Each dataset is described by its name, the included sonotypes, the number of training examples obtained from XenoCanto, and the number of evaluation clips available in both 3-s and 12-s formats.

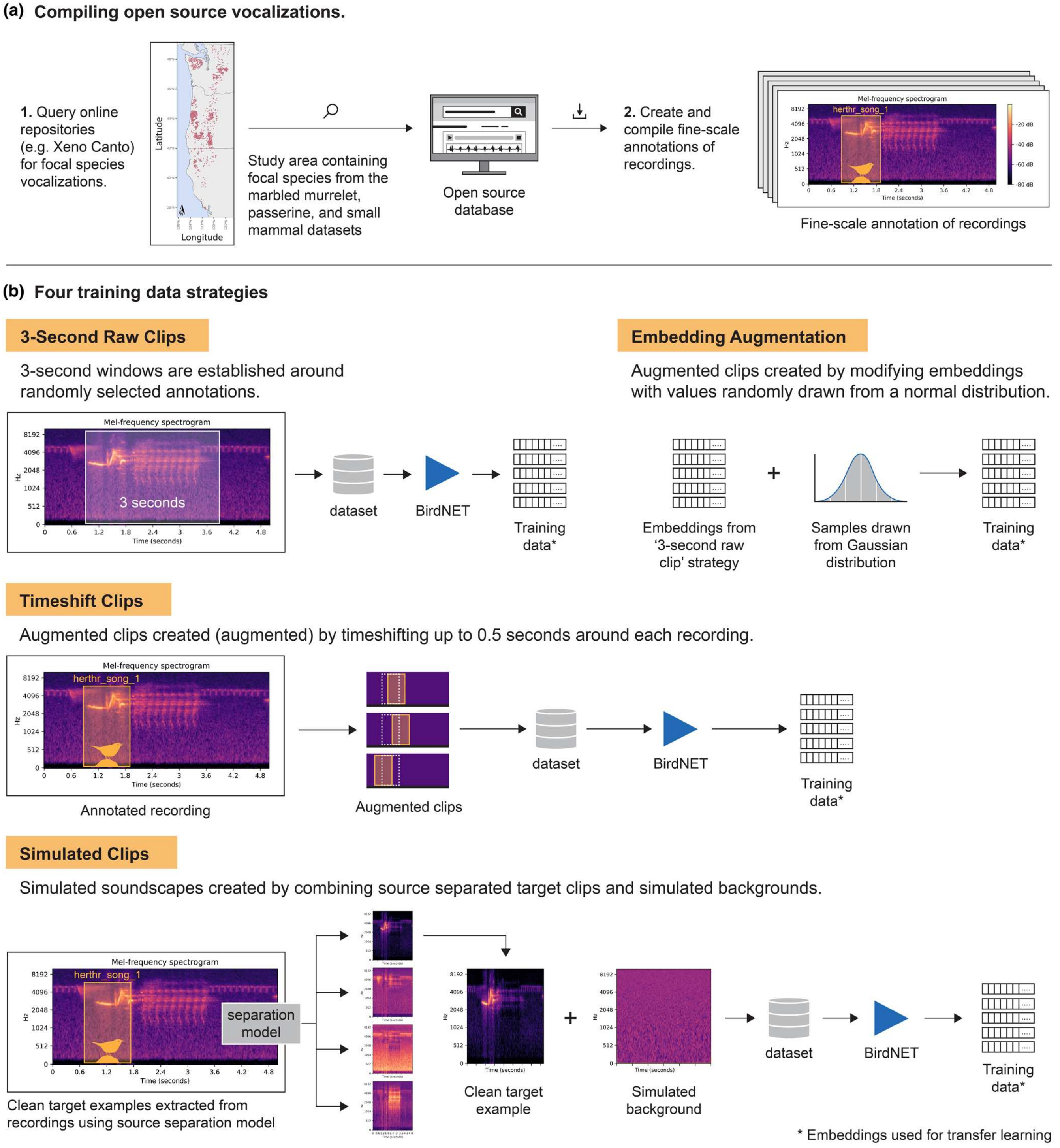
(a) Compiling open source vocalizations.
1. Query online repositories (e.g. Xeno Canto) for focal species vocalizations.
Latitude
Longitude
Study area containing focal species from the marbled murrelet, passerine, and small mammal datasets
Open source database
2. Create and compile fine-scale annotations of recordings.
Mel-frequency spectrogram
herthr_song_1
Fine-scale annotation of recordings
(b) Four training data strategies
3-Second Raw Clips
3-second windows are established around randomly selected annotations.
3 seconds
dataset
BirdNET
Training data*
Embedding Augmentation
Augmented clips created by modifying embeddings with values randomly drawn from a normal distribution.
Embeddings from '3-second raw clip' strategy
Samples drawn from Gaussian distribution
Training data*
Timeshift Clips
Augmented clips created (augmented) by timeshifting up to 0.5 seconds around each recording.
Annotated recording
Augmented clips
dataset
BirdNET
Training data*
Simulated Clips
Simulated soundscapes created by combining source separated target clips and simulated backgrounds.
separation model
Clean target examples extracted from recordings using source separation model
Clean target example
Simulated background
dataset
BirdNET
Training data*
* Embeddings used for transfer learning

FIGURE 1 Conceptual representation of (a) our training data collection and (b) our training data construction approaches including training on few raw examples ('raw'), augmenting the raw examples with standard Gaussian noise ('embedding'), time shifting ('timeshift') and soundscape simulation ('simulated clips').

clips and the other with 12-s sound clips (Table 1). The source files for both datasets were independent recordings but originated from the same PAM projects.

The marbled murrelet (*Brachyramphus marmoratus*; eBird code: marmur) dataset is a single-class classification problem that includes 0.46 h of annotated 3-s clips and 6.67 h of annotated 12-s clips. The recordings were collected and annotated by the United States Department of Agriculture (USDA) Forest Service in association with the annual Northwest Forest Plan (NWFP) PAM of northern spotted owl (*Strix occidentalis caurina*; eBird code: stroca) populations and other old-forest-associated species (Duarte, Weldy, et al., 2024; Lesmeister & Jenkins, 2022). The recordings are selected clips from a larger 644,111-h acoustic dataset collected from March to September 2020. During 2020, PAM occurred on 1494 US Federally managed forest sites west of the Cascade Mountain Range in Oregon and Washington, USA. The acoustic data were collected using Song Meter SM4 autonomous recording units (hereafter ARU) at a sampling rate of 32 kHz and a 16-bit resolution.

The Blue Mountains dataset is a six-class multi-label classification problem that includes annotated vocalizations for three species that are either an indicator of forest management activities or are of conservation concern (Altman & Bresson, 2017): Clark's nutcracker (*Nucifraga columbiana*; eBird code: clanut), American goshawk (*Accipiter atricapillus*; eBird code: norgos) and white-headed woodpecker (*Dryobates albolarvatus*; eBird code: whhwoo). The six-class labels identify three Clark's nutcracker call types, two American goshawk call types and one white-headed woodpecker call type. The recordings were collected by the USDA Forest Service in Oregon, USA, in the northern Blue Mountains on the Wallowa-Whitman and Umatilla National Forests as part of ongoing PAM for the Northern Blues Collaborative Forest Landscape Restoration Program (Duarte, Vernasco, et al., 2024; https://research.fs.usda.gov/pnw/understory/northern-blue-mountains-wildlife-monitoring-2022-2023). The recordings are selected clips from a larger 122,052-h dataset collected on 420 US Federally managed forest sites. The acoustic data were collected using Song Meter SM4 ARUs at a sampling rate of 32 kHz and a 16-bit resolution (Wildlife Acoustics, Concord, NY, U. S. A.).

The passerine dataset is a 10-class multi-label classification problem that includes annotated vocalizations for six species: hermit thrush (*Catharus guttatus*; eBird code: herthr), olive-sided flycatcher (*Contopus cooperi*; eBird code: olsfly), spotted towhee (*Pipilo maculatus*; eBird code: spotow), Swainson's thrush (*Catarus ustulatus*; eBird code: swathr), varied thrush (*Ixoreus naevius*; eBird code: varthr) and wrentit (*Chamaea fasciata*; eBird code: wrenti). The six-class labels identify three Swainson's thrush call types, two hermit thrush call types, two olive-sided flycatcher call types, one spotted towhee call type, one varied thrush call type and one wrentit call type. These recordings were obtained from Weldy et al. (2024) and included a subset of dawn chorus recordings collected during the 2022 NWFP PAM, described above. In 2022, PAM expanded to include 2572 US Federally managed forest sites west of the Cascade and Sierra Mountain Ranges in California, Oregon and Washington, recording 1,477,751.64 h of sound. The sampling protocol and acoustic characteristics were consistent with those described in the marbled murrelet dataset.

### 2.1.2 | Non-avian datasets

The non-avian datasets are multi-label classification problems that represent a range of classification complexities, varying in number of target classes and the degree to which the evaluation data differs from BirdNET's base training data.

The amphibian dataset includes 2494 5-s recordings annotated for two amphibian species: American bullfrog (*Lithobates catesbeianus*; code: amebul) and Pacific chorus frog (*Pseudacris regilla*; code: pacfro). We obtained training examples from California Herps, a web resource documenting the life history of California's reptiles and amphibians (Nafis, 2021; acoustic data obtained with permission). The annotated evaluation data were collected by the United States Geological Survey (USGS) and its partners at 86 wetland sites across Oregon and Washington, USA (Hill et al., 2019). The objective was to monitor bullfrog vocalization activity in the range of federally threatened Oregon spotted frogs (*Rana pretiosa*); non-target Pacific chorus frogs also occurred in the study areas. The recordings were collected using AudioMoth ARUs at a sampling rate of 48 kHz in 2020 and 16 kHz in 2021. The annotated clips were identified using the Kaleidoscope software cluster analysis feature targeting the bullfrog call's frequency range (187.5–5250 Hz; Bielinski et al., 2020) and manually reviewed. Audio collection by USGS was covered under annual USFWS Special Use Permits 20-04 and 21-01.

The cricket dataset includes 1000 2-min recordings annotated for 10 species of cricket and one cricket subfamily: *Archenopterus bouensis*, *Bullita fusca*, *Bullita mouirangensis*, *Bullita obscura*, *Calscirtus magnus*, *Koghiella flammea*, *Koghiella nigris*, *Notosciobia affnis paranola*, *Notoscioba minoris*, *Pseudotrigonidium caledonica* and *Trigonidiinae* spp. We obtained training examples from the Muséum National d'Histoire Naturelle of Paris (https://sonotheque.mnhn.fr; Sound Catalog accessed 8/01/2024). The annotated PAM data used for evaluation were recorded in New Caledonia as a component of long-term research on the effects of the invasive little fire ant *Wasmannia auropunctata* on biodiversity (Jourdan et al., 2001). Gasc et al. (2018) collected PAM recordings on 24 forest, pre-forest and shrubland sites during the dry season of 2013 and used these recordings to assess acoustic-based detection of *Wasmannia auropunctata* through changes in the acoustic calling behaviour of crickets. The recordings were collected using SongMeter SM2 and SM2+ ARUs at a sampling rate of 48 kHz and a 16-bit resolution (Wildlife Acoustics, Concord, NY, USA). We used two versions of this dataset: the first includes 11 classes, treating species-level classes separately. The second includes seven classes pooling species-level annotations at the taxonomic resolution of genus. We created the second version of this dataset to create a potentially easier classification task because we suspected that the first task would be a difficult out of domain task for BirdNET.

The small mammal dataset includes 1737 12-s recordings annotated for three sounds from two species: American pika (*Ochotona pinceps*; code: amepik) and Douglas squirrel (*Tamiasciurus douglasii*; code: dousqu). The three class labels identify two Douglas squirrel vocalizations and one American pika vocalization. The recordings were collected and annotated by the USDA Forest Service during NWFP PAM, described above. The sampling protocol and acoustic characteristics were consistent with those described in the marbled murrelet dataset.

### 2.1.3 | Annotated clip selection and annotation

The recordings for the marbled murrelet, Blue Mountains, small mammal and amphibian datasets were annotated opportunistically during the manual review of project focal species predictions. The passerine dataset recordings were selected for annotation in a stratified random sample that included three randomly selected recordings from each site from the dawn chorus period during the first hour following sunrise from May to August. For the cricket dataset, recordings were randomly selected from each site, with selections constrained to exclude recordings affected by wind or rain noise. Taxonomic experts conducted exhaustive annotations of all selected recordings, identifying the presence or absence of target sounds within designated sample windows that varied by dataset (5 s for amphibian, 3 s for Blue Mountains, 2 min for cricket and 12 s for marbled murrelet and small mammal).

## 2.2 | Experimental methodology

Our analysis consisted of two main parts. For both parts, we acquired training data from publicly available sound repositories and evaluated performance using annotated data collected during PAM. First, we developed a transfer learning strategy using BirdNET's feature embeddings in four experimental steps and assessed the performance of custom neural network linear acoustic classifiers relative to BirdNET and PNW-Cnet for shared classes. Second, we leveraged our transfer learning strategy to build acoustic classifiers for non-avian species.

We used a transfer learning approach leveraging embeddings, which are numeric representations produced by the penultimate layer of a pre-trained model. These embeddings represent characteristic features of the input data—in this case, variation in magnitude of acoustic signals across frequencies or time—that are useful for training new classifiers (Ghani et al., 2023). We used BirdNET as an embedding model throughout our experiments (Kahl et al., 2021), which maps every 3 s of audio, sampled at 48 kHz, to a 1024-dimensional numeric embedding. We trained new classifiers over embedded training data using the Adam optimizer, binary cross-entropy loss and a fixed number of gradient descent steps (Kingma & Ba, 2015).

We assessed model performance relative to the annotations in the evaluation datasets using two threshold-independent metrics: area under the receiver operating characteristic curve (hereafter AUC) and average precision (AP). AUC is the probability that a randomly selected positive example scores higher than a randomly selected negative example (Fawcett, 2006; van Merriënboer et al., 2024). AP measures how well the model correctly predicts positive examples across many thresholds. For single-class binary classifiers, we report AUC and AP directly; for multi-class, multi-label classifiers, we report macro averaged AUC ($\text{AUC}_{\text{macro}}$) and AP (mAP). We repeated each experimental step 10 times and reported an average of all metrics to reduce the stochastic sensitivity of our performance estimates. In addition, because we are developing acoustic classifiers in a transfer learning context, we ensured our evaluation datasets were independent of the training data for BirdNET.

For evaluation datasets where the annotated clip lengths (e.g. 12 s) exceeded the receptive field of the classifier (3 s for BirdNET and our transfer learning classifiers), we divided each clip into non-overlapping 3-s subsets, applied the classifier to each subset and aggregated the predicted scores. Specifically, we selected the maximum score for each class among all subsets of a given clip. To evaluate BirdNET's performance on vocalization-specific annotations below the species level (e.g. call types), we treated each vocalization type within a species as a distinct class and repeated BirdNET's species-level predictions accordingly. For example, the passerine dataset includes two vocalization-specific classes for hermit thrush: the hermit thrush song (herthr_song_1) and the hermit thrush call (herthr_call_1; Table 1). To evaluate BirdNET's performance for these two classes, we repeated BirdNET's species-level hermit thrush prediction for both classes.

### 2.2.1 | Part 1: Transfer learning strategy

In Experiment 1, we estimated linear classifier performance for all combinations of three hyperparameter value sets (i.e. an ablation), including batch size (16, 32, 64, 128), learning rate (0.1, 0.01, 0.001) and the number of gradient descent steps (100, 500, 1000, 2000). For all hyperparameter combinations, we compiled training datasets by selecting up to 100 vocalizations for each class from the annotated XenoCanto recordings, without restricting selection to one vocalization per original recording, and paired them with an equal number of simulated background clips. We then embedded the datasets and fitted linear classifiers using each hyperparameter value combination. We evaluated the relative performance of the trained linear classifiers using the 3-s evaluation datasets. We assessed the overall performance of each hyperparameter value by averaging performance metrics across the hyperparameter combination replications and ranking the average performance by the number of times each hyperparameter value was included in the top 10 combinations. Our hyperparameter search was not exhaustive; however, we sought a reliable combination of hyperparameters that yielded consistent performance without overfitting to the simulated training data. We adopted the optimal hyperparameter values in subsequent experimental steps.

In Experiment 2, we evaluated four approaches to constructing training datasets under five levels of imposed data scarcity (2, 4, 8, 16, 32 training examples per sound type; Figure 1b). For each level of data scarcity, we first randomly selected annotated XenoCanto examples of each class, without restricting selections to one vocalization per original recording, and used these examples to construct four training datasets. We then embedded the datasets and fit linear classifiers. The first data construction approach ('raw') uses 3-s sound windows extracted from around the selected XenoCanto annotations in the original recordings. The second approach ('embedding') augments the raw examples to 100 examples per class by adding randomly generated standard normal Gaussian noise ($\mu=0$, $\sigma=1$) to the embeddings of the 'raw' training data. The third approach ('timeshift') augments the 'raw' training dataset by shifting the acoustic window by up to 0.5 s around the selected annotations in the original XenoCanto recordings. The fourth approach ('simulated clip') implements stochastic soundscape simulation using the python (version 3.10.9) scaper package (Salamon et al., 2017). We evaluated the relative performance of the linear classifiers using the 12-s evaluation datasets.

The scaper python package provides tools to programmatically generate novel audio soundscapes through additive layering of sounds, where the parameters describing the placement and relative loudness of sounds are randomly sampled from user defined probabilistic distributions. Each generated soundscape consists of randomly sampled foreground examples (e.g. in this case isolated avian vocalizations) layered over a randomly selected background sound. We compiled a collection of foreground vocalizations from publicly available sound repositories for each sound of interest. The collected vocalizations were then preprocessed with a source separation model (Denton et al., 2022), which splits multi-source audio recordings into four separate channels (Figure 1b). From the source-separated multi-channel output, we manually selected the isolated target sound. We simulated 400 background sound examples of four types of noise (100 each) and included one clip of silence. The noise types included Gaussian noise ($\mu=0$, $\sigma=1$), mixtures of Gaussian noise with Butterworth low-pass filtered noise, and impulse augmented examples of both. The Butterworth low-pass clips included order one and two filters with cut-off frequencies ranging from 500 Hz to 5 kHz. The impulse augmented examples included one to five short, high-intensity spikes added to the audio (Figure 1b).

In Experiment 3, we performed an ablation over eight classification model architectures. These included a single-layer linear classifier (a neural network with no hidden layers) and 3 two-layer multilayer perceptrons (MLP), which are neural networks consisting of an input layer, one fully connected hidden layer and an output layer. The hidden layers in these MLPs consisted of 512, 1024, 2048 units with rectified linear unit activations, which introduce potential non-linearities to the models. Additionally, we tested four modified versions of these architectures that included a dropout layer—a regularization technique which can reduce overfitting—with a dropout rate of 0.3, as the penultimate layer. We simulated 100,000 3-s audio clips for each avian dataset using the simulated clip approach described in Experiment 2 and embedded the clips. We randomly selected 1000 of the embeddings and fit the eight model architectures to the selection. We evaluated the relative performance of the classifiers using the 12-s evaluation datasets.

In Experiment 4, we evaluated the effect of increasing the number of simulated examples (128, 256, 512, 1024, 2048, 4096, 8192, 16,384, 32,768) and the effects of two additional acoustic augmentations. The acoustic augmentations included pitch shifting up and down by a random amount sampled from a uniform distribution ranging from −2 to 2 semitones and time stretching by a random factor sampled from a uniform distribution ranging from 0.8 to 1.2 times the original clip length. In addition to 100,000 3-s embeddings generated in Experiment 3 (simulated with no acoustic augmentations), we simulated three additional sets of 100,000 3-s embeddings for each avian dataset by simulating acoustic clips while applying pitch shifting, time stretching and their combination to the preprocessed examples. We randomly selected a fixed number of embeddings for each embedding set and fitted single-layer linear classifiers to each selection. We evaluated the relative performance of the linear classifiers using the 12-s evaluation datasets.

We then developed new classifiers for each avian dataset using our transfer learning strategy developed in Experiments 1–4. For shared classes, we compared the performance of our custom classifiers to the off-the-shelf performance of BirdNET and PNW-Cnet for vocalization-specific classification. The custom classifiers were linear classifiers with one layer of dropout trained with 8192 simulated clips generated using four known examples and no pitch shifting or time stretching.

### 2.2.2 | Part 2: Generalization to non-avian sounds

We applied our simulation-based transfer learning strategy to develop new acoustic classifiers for three non-avian datasets representing a range of potential complexities. The amphibian dataset, the least complex, includes two species that are part of BirdNET's training dataset. The small mammal dataset is slightly more complex than the amphibian dataset because American pika and Douglas squirrel vocalizations are not included in the BirdNET training dataset. However, BirdNET does include at least two other squirrel species with similar vocalizations. The cricket dataset is more complex than the other non-avian datasets; it includes 10 species from at least seven genera and one subfamily that are not part of the BirdNET training dataset.

## 2.3 | Visualization of BirdNET embeddings

We visualized BirdNET's 1024-dimensional feature embeddings using t-distributed stochastic neighbour embeddings (t-SNE; Hinton & Roweis, 2002). t-SNE is a dimensionality reduction technique that attempts to preserve local distances between data points while mapping high-dimensional data into lower dimensions (van der Maaten & Hinton, 2008). We first filtered the 3-s evaluation datasets to background clips with no annotations and clips with one annotation and embedded these clips with BirdNET. We then mapped the BirdNET feature embeddings for each avian dataset to two dimensions using a principal component initialized t-SNE fit for 5000 iterations with a learning rate of 10. We considered four perplexity values—perplexity balances t-SNE's relative optimization on local and global representations and approximates the number of neighbours each point has—ranging from 3 to 50 to examine the stability of the t-SNE mapping (Figure S1; Wattenberg et al., 2016).

## 3 | RESULTS

We found that training custom acoustic classifiers with as few as two positive examples improves classifier performance relative to BirdNET's off-the-shelf predictions (Experiment 2; Figure 2). For all three avian datasets, average relative performance increased asymptotically as the number of training source examples increased, while the variance of the average relative performance among training replicates decreased (Experiment 4; Figure 3; Table S4). Average performance of the classifiers trained with soundscape simulation saturated quickly, with both $\text{AUC}_{\text{macro}}$ and mAP reaching 95% of the maximum performance for each avian dataset with just four training samples. Of the four approaches to training data construction, the simulated clips approach consistently performed the best and yielded improvements relative to the BirdNET baseline. The other three training data strategies often failed to improve upon the BirdNET baseline, except for the Blue Mountains dataset, where all four strategies resulted in higher $\text{AUC}_{\text{macro}}$ and mAP relative to the baseline. We observed a plateau in model performance gains, similar to other transfer learning applications, where increases in data can show diminishing relative performance improvements (Ghani et al., 2023; Kath et al., 2024). These diminishing gains may reflect inherent limitations to the information content of the

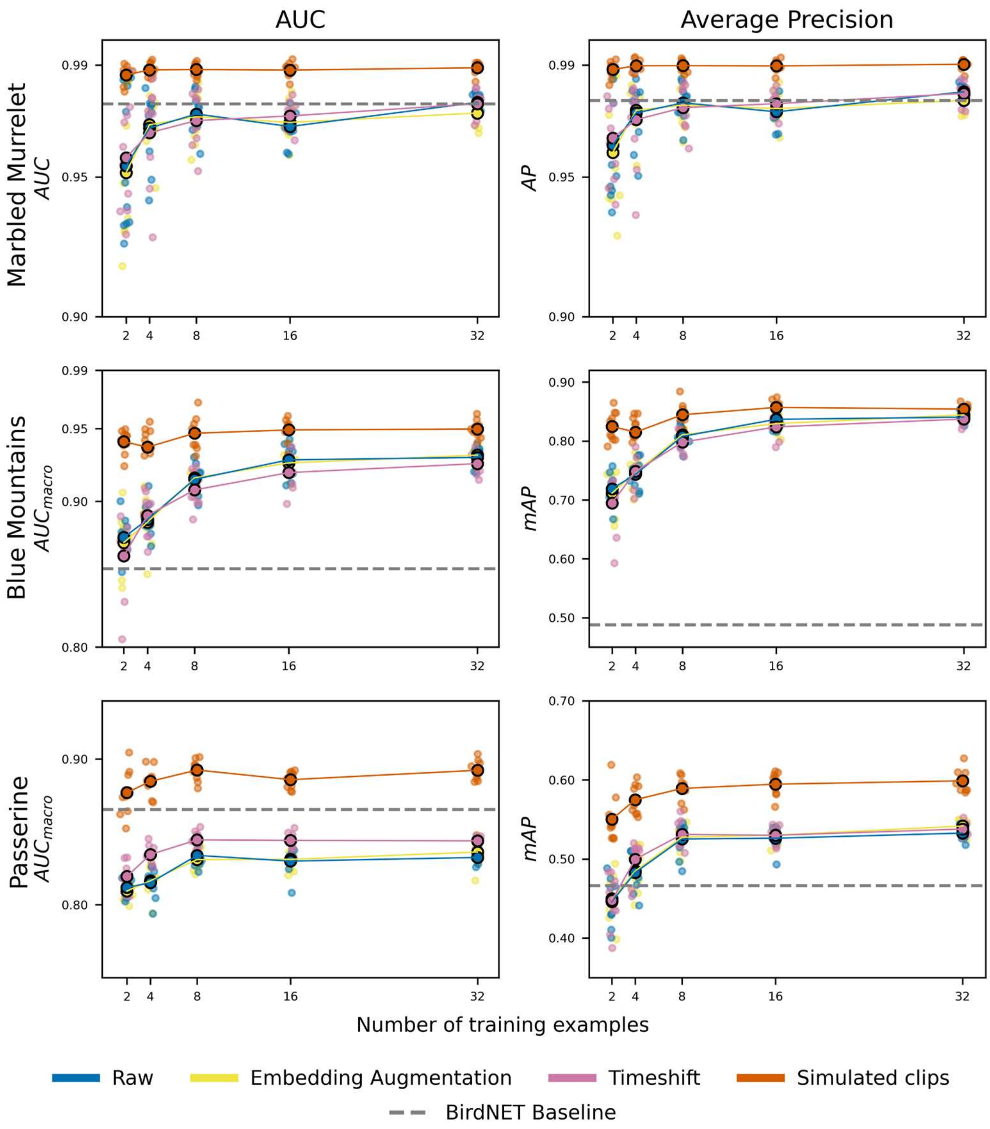

**FIGURE 2** Relative performance of four training data strategies for developing acoustic classification algorithms using transfer learning under data scarcity. The simulated clips data augmentation approach consistently performed the best and yielded improvements relative to the BirdNET baseline. The number of raw examples varied, and the dataset construction and training were replicated 10 times for 2, 4, 8, 16 and 32 raw examples. Dark points and lines show the average performance of the four training approaches. Light points indicate replicate-level performance. Area under the receiver operating characteristic (ROC) curve (AUC) measures the probability that a randomly selected true positive example is scored higher than a randomly selected true negative example. Average precision (AP) is the weighted mean of precision across all thresholds. $\text{AUC}_{\text{macro}}$ and mAP are the mean of all class-specific metrics. of precision across all thresholds. $\text{AUC}_{\text{macro}}$ and mAP are the mean of all class-specific metrics.

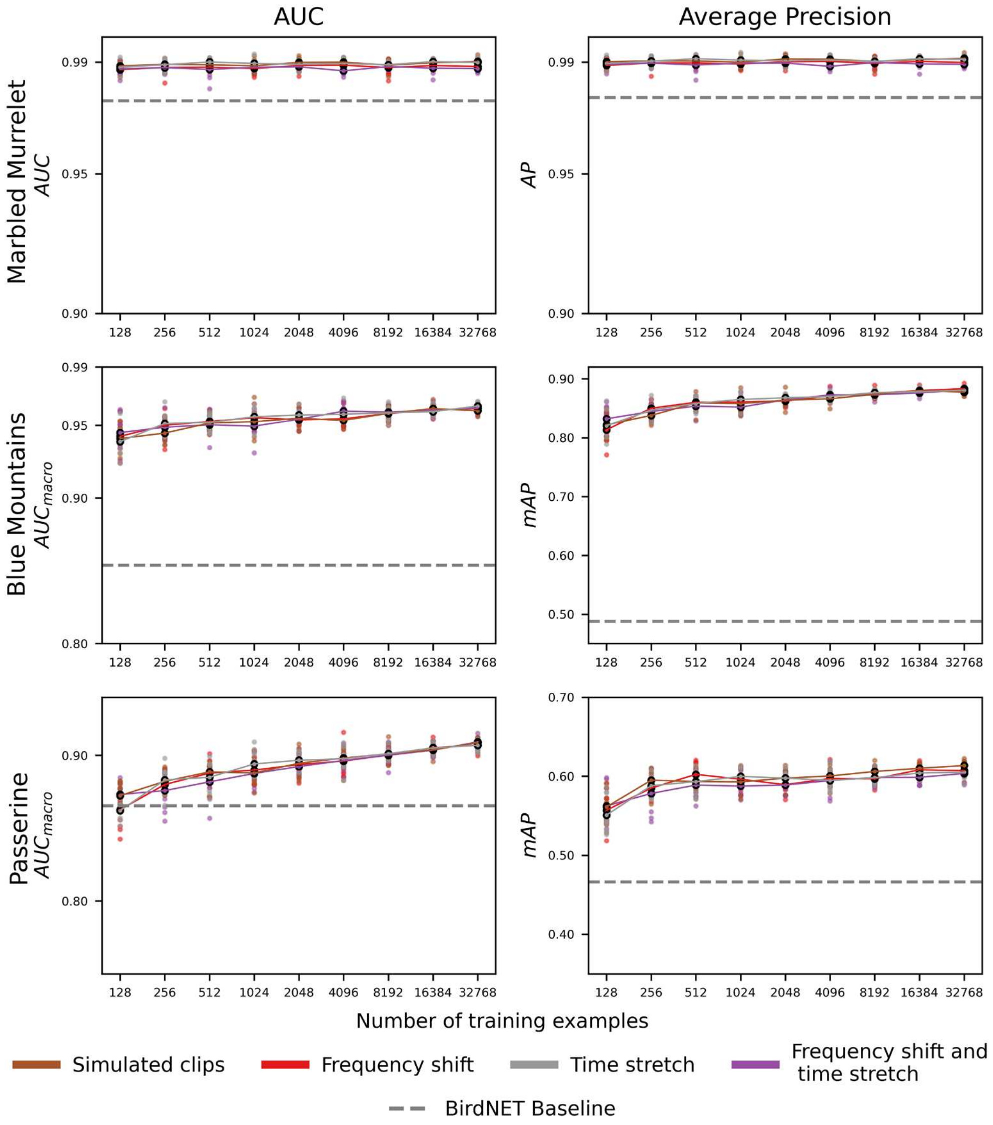

**FIGURE 3** Relative effects of increasing the number of simulated clips and three acoustic data augmentation techniques relative to soundscape simulation without acoustic augmentations for acoustic classification algorithms using transfer learning under data scarcity. Increasing the number of simulated soundscapes consistently improved relative model performance, with no evidence of overfitting. The number of simulated examples varied, and the dataset construction and training were replicated 10 times. Dark points and lines indicate the average performance of the acoustic augmentations.

embeddings, such as limited discriminatory power to differentiate among different sounds produced by the same species (Figure 4), or differences among training and evaluation datasets.

Average classifier performance varied little among the eight classifier architectures for classifiers with up to 10 classes (Experiment 3; Table 3). All four base architectures were competitive, and future studies facing more complex classification problems should evaluate classifier architectures with higher relative capacity. Adding a hidden layer decreased classifier performance relative to the linear classifier, likely due to overfitting. This overfitting was evidenced by continued decreases in training loss over batches of simulated data, even after generalization performance on the evaluation datasets plateaued. However, applying dropout before prediction partially mitigated this decrease in performance, preventing overfitting to the simulated training data and restoring two-layer model performance closer to the level of the linear classifier.

Increasing the number of simulated soundscapes increased relative model performance, with no evidence for overfitting (Experiment 4; Figure 3). However, the rate of increase in performance was slow at greater than 1000 simulated soundscapes (Figure 3). Adding acoustic augmentations, such as pitch shifting and time stretching, to the positive training examples during the simulations did not increase average relative performance (Figure 3).

Each hyperparameter value was included in the top 10 average hyperparameter combinations at least once, and the relative performance varied across replicate linear classifier fits with the same hyperparameter combination (Experiment 1; Table 2). We found slight evidence of overfitting for the marbled murrelet and Blue Mountains

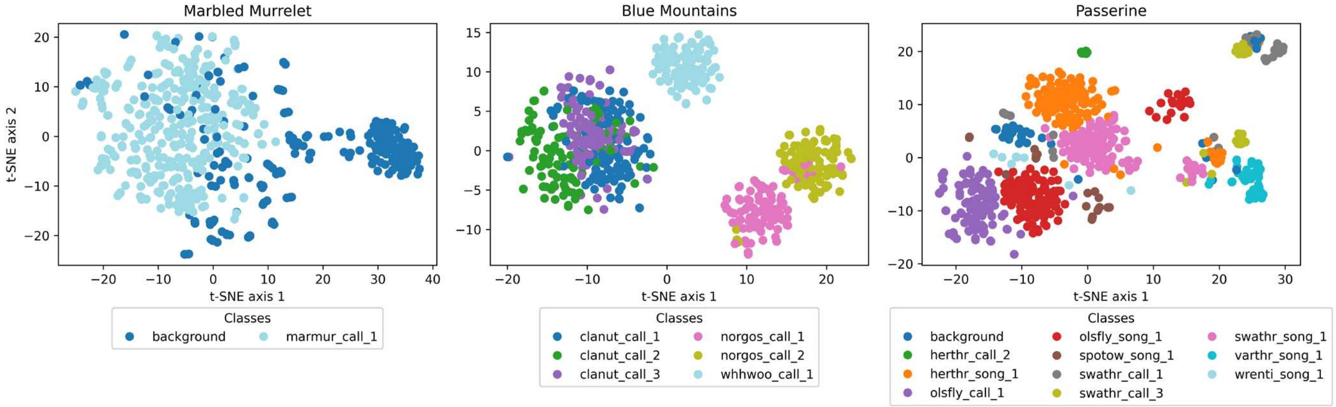

FIGURE 4 t-distributed stochastic neighbour embedding (t-SNE) plots of the 3-s avian evaluation datasets. The t-SNE embeddings were fit for 5000 iterations using a learning rate 10 and a perplexity value 50. Each point on the plots is the 2-dimensional t-SNE projection of the 1024-dimensional BirdNET feature embedding for a 3-s annotated audio clip. The t-SNE visualization illustrates the variability among acoustic clips based on BirdNET feature embeddings. Different colours indicate groups of distinct sound types (sonotypes). See Table 1 for descriptions of sonotypes and corresponding species.

TABLE 2 Hyperparameter ablation for simulation-based transfer learning classifiers built over BirdNET embeddings for the three avian datasets.

| | | Marbled murrelet | | Blue mountains | | Passerine | |
|---|---|---|---|---|---|---|---|
| Hyperparameter | Value | AUC | AP | $\text{AUC}_{\text{macro}}$ | mAP | $\text{AUC}_{\text{macro}}$ | mAP |
| Batch size | 16 | 2 | 2 | 2 | 0 | 2 | 0 |
| | 32 | 1 | 2 | 2 | 3 | 2 | 3 |
| | 64 | 3 | ***3*** | ***3*** | 3 | 2 | 2 |
| | 128 | **4** | ***3*** | ***3*** | **4** | **4** | **5** |
| Learning rate | 0.001 | 2 | 0 | ***4*** | 0 | 0 | 3 |
| | 0.01 | **7** | **8** | ***4*** | 3 | 2 | **7** |
| | 0.1 | 1 | 2 | 2 | **7** | **8** | 0 |
| Training steps | 100 | 2 | 1 | ***5*** | **5** | **5** | 0 |
| | 500 | 0 | 0 | ***5*** | 4 | 1 | 2 |
| | 1000 | ***4*** | 4 | 0 | 1 | 2 | 3 |
| | 2000 | ***4*** | **5** | 0 | 0 | 2 | **5** |

*Note*: We varied batch size, learning rate and the number of gradient descent steps. The value indicates the number of times the average performance in terms of area under the curve (AUC) and average precision (AP) for a specific hyperparameter value was in the top 10 average hyperparameter combinations. Area under the ROC curve (AUC) measures the probability that a randomly selected true positive example is scored higher than a randomly selected true negative example. Average precision (AP) is the weighted mean of precision across all thresholds. $\text{AUC}_{\text{macro}}$ and mAP are the mean of all class-specific metrics. Bold values indicate the top-performing value for a specific hyperparameter, metric and avian dataset. Bold and italic values indicate shared top-performing values for a specific hyperparameter.

classifiers because model performance estimates across the set of hyperparameter combinations were within a range of 0.05 (Table S3). The hyperparameter evaluation provided weak evidence that larger batch sizes and intermediate learning rates marginally improved relative performance. We adopted the following hyperparameter values for the remainder of our analyses: batch size: 128, learning rate: 0.01 and 500 gradient descent steps. Our choice for the number of gradient descent steps represents a compromise between reducing the tendency for overfitting while training long enough for the model to be exposed to all the data at least once.

Overall, BirdNET and PNW-Cnet performed well on the avian evaluation datasets. PNW-Cnet AUC and AP scores were higher than BirdNET for 80% of the shared acoustic classes ($n=10$; Table 4). The average class-specific performance of the simulation-based classifiers trained using four known examples was higher than BirdNET for all classes except for two cases where the AP of the simulation-based classifiers did not improve upon BirdNET's performance (Table 4). However, the maximum class-specific performance of the simulation-based classifiers was higher than BirdNET's score for all classes. The average class-specific performance of the

TABLE 3 Model structure ablation for the three avian datasets.

| Structure | No. of units | Dropout | Marbled murrelet | | Blue mountains | | Passerine | |
|---|---|---|---|---|---|---|---|---|
| | | | $\overline{\text{AUC}}$ | $\overline{\text{AP}}$ | $\overline{\text{AUC}}_{\text{macro}}$ | mAP | $\overline{\text{AUC}}_{\text{macro}}$ | mAP |
| LP | | N | 0.990 | 0.991 | 0.955 | 0.866 | 0.898 | 0.602 |
| | | Y | 0.990 | 0.991 | 0.958 | 0.872 | 0.90 | 0.609 |
| MLP | 1024 | N | 0.934 | 0.934 | 0.952 | 0.861 | 0.893 | 0.591 |
| | | Y | 0.953 | 0.938 | 0.953 | 0.860 | 0.897 | 0.602 |
| | 2048 | N | 0.933 | 0.929 | 0.953 | 0.863 | 0.893 | 0.591 |
| | | Y | 0.933 | 0.913 | 0.954 | 0.862 | 0.897 | 0.601 |
| | 512 | N | 0.966 | 0.958 | 0.951 | 0.859 | 0.893 | 0.590 |
| | | Y | 0.974 | 0.962 | 0.955 | 0.865 | 0.897 | 0.600 |

*Note*: We report the average performance for eight classifier model architectures across 10 replicate model trainings. LP refers to a single-layer linear probe. Three two-layer perceptron (MLP) architectures vary in the number of units included in a single hidden layer. Area under the ROC curve ($\overline{\text{AUC}}$) measures the probability that a randomly selected true positive example is scored higher than a randomly selected true negative example. Average precision ($\overline{\text{AP}}$) is the weighted mean of precision across all thresholds. $\overline{\text{AUC}}_{\text{macro}}$ and $\overline{\text{mAP}}$ are the mean of all class-specific metrics.

TABLE 4 Comparative class-specific performance of BirdNET, PNW-Cnet and simulation-based transfer learning classifiers built over BirdNET embeddings and four known examples for shared classes in the 12-s avian evaluation datasets.

| Sonotype | $\overline{\text{AUC}}$ | | | | $\overline{\text{AP}}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | BirdNET | PNW-Cnet | $\text{Sim}_{\text{mean}}$ | $\text{Sim}_{\text{max}}$ | BirdNET | PNW-Cnet | $\text{Sim}_{\text{mean}}$ | $\text{Sim}_{\text{max}}$ |
| marmur_call_1 | 0.976 | 0.998 | 0.988 | 0.991 | 0.977 | 0.999 | 0.989 | 0.992 |
| clanut_call_1 | 0.782 | | 0.958 | 0.981 | 0.221 | | 0.847 | 0.915 |
| clanut_call_2 | 0.741 | | 0.934 | 0.961 | 0.084 | | 0.713 | 0.770 |
| clanut_call_3 | 0.897 | 0.843 | 0.939 | 0.969 | 0.736 | 0.611 | 0.872 | 0.945 |
| norgos_call_1 | 0.820 | | 0.953 | 0.969 | 0.320 | | 0.840 | 0.877 |
| norgos_call_2 | 0.920 | | 0.976 | 0.985 | 0.610 | | 0.916 | 0.942 |
| whhwoo_call_1 | 0.964 | | 0.987 | 0.992 | 0.956 | | 0.980 | 0.988 |
| herthr_song_1 | 0.810 | 0.928 | 0.849 | 0.877 | 0.570 | 0.863 | 0.642 | 0.709 |
| herthr_call_2 | 0.875 | | 0.874 | 0.908 | 0.190 | | 0.589 | 0.622 |
| olsfly_song_1 | 0.836 | 0.926 | 0.902 | 0.933 | 0.243 | 0.673 | 0.605 | 0.656 |
| olsfly_call_1 | 0.911 | | 0.924 | 0.959 | 0.541 | | 0.619 | 0.651 |
| spotow_song_1 | 0.888 | 0.803 | 0.922 | 0.951 | 0.606 | 0.120 | 0.540 | 0.633 |
| swathr_song_1 | 0.897 | 0.971 | 0.931 | 0.945 | 0.484 | 0.886 | 0.733 | 0.759 |
| swathr_call_1 | 0.907 | | 0.962 | 0.974 | 0.541 | | 0.745 | 0.822 |
| swathr_call_3 | 0.891 | | 0.896 | 0.920 | 0.512 | | 0.605 | 0.678 |
| varthr_song_1 | 0.833 | 0.924 | 0.854 | 0.871 | 0.607 | 0.836 | 0.656 | 0.697 |
| wrenti_song_1 | 0.806 | 0.922 | 0.847 | 0.874 | 0.373 | 0.637 | 0.364 | 0.426 |

*Note*: We report the average and maximum area under the receiver operator curve ($\overline{\text{AUC}}$) and average precision ($\overline{\text{AP}}$) for each class across 10 replicate datasets and model training steps.

simulation-based classifiers only surpassed PNW-Cnet's scores for classes in which BirdNET also scored higher, but the maximum performance of the simulation-based classifier was competitive or surpassed PNW-Cnet for all classes (Table 4).

For the three non-avian datasets, relative performance decreased with increasing task complexity (Table 5). BirdNET's overall baseline performance for the amphibian dataset was high ($\text{AUC}_{\text{macro}}$: 0.993; mAP: 0.991) with high class-specific performance: the American bullfrog $\text{AUC}_{\text{macro}}$ was 0.995 and mAP was 0.995, the Pacific chorus frog $\text{AUC}_{\text{macro}}$ was 0.991 and mAP was 0.987. AUC and AP scores for our methods were consistent with, but slightly lower than, BirdNET's baseline performance for these species (Table 5). We could not estimate BirdNET's baseline performance for the cricket and small mammal datasets because the species comprising those datasets are not included in BirdNET's training data. For our approach, the overall performance of the cricket dataset was

**TABLE 5** Overall and class-specific performance for three out-of-domain linear classifiers trained with few examples using a simulation-based transfer learning approach.

| Taxa | Species | No. of examples | No. of evaluations | $\overline{\text{AUC}}$ | $\overline{\text{AP}}$ |
|---|---|---|---|---|---|
| Amphibian | Bullfrog | 45 | 965 | 0.996 | 0.995 |
| | Pacific chorus frog | 51 | 912 | 0.994 | 0.992 |
| | Non-target | | 895 | | |
| | Overall | | | 0.995 | 0.994 |
| Cricket—Genus | Archenopterus | 11 | 163 | 0.787 | 0.538 |
| | Bullita | 13 | 132 | 0.831 | 0.346 |
| | Calscirtus | 1 | 221 | 0.814 | 0.076 |
| | Koghiella | 6 | 147 | 0.647 | 0.252 |
| | Notosciobia | 5 | 126 | 0.938 | 0.870 |
| | Pseudotrigonidium | 6 | 29 | 0.598 | 0.200 |
| | Trigonidiinae | 15 | 53 | 0.767 | 0.239 |
| | Non-target | | 509 | | |
| | Overall | | | 0.769 | 0.360 |
| Cricket—Species | *Archenopterus bouensis* | 11 | 163 | 0.787 | 0.540 |
| | *Bullita fusca* | 6 | 88 | 0.825 | 0.344 |
| | *Bullita mouirangensis* | 2 | 20 | 0.811 | 0.078 |
| | *Bullita obscura* | 5 | 24 | 0.664 | 0.260 |
| | *Calscirtus magnus* | 1 | 221 | 0.934 | 0.863 |
| | *Koghiella flammea* | 3 | 43 | 0.598 | 0.201 |
| | *Koghiella nigris* | 3 | 104 | 0.756 | 0.233 |
| | *Notosciobia affnis paranola* | 2 | 105 | 0.810 | 0.499 |
| | *Notosciobia minoris* | 3 | 21 | 0.836 | 0.185 |
| | *Pseudotrigonidium caledonica* | 6 | 29 | 0.846 | 0.179 |
| | Trigonidiinae spp. | 15 | 53 | 0.683 | 0.171 |
| | Non-target | | 509 | | |
| | Overall | | | 0.777 | 0.323 |
| Small mammal | American pika | 60 | 845 | 0.970 | 0.977 |
| | Douglas squirrel: chirp | 30 | 236 | 0.959 | 0.876 |
| | Douglas squirrel: rattle | 30 | 157 | 0.915 | 0.774 |
| | Non-target | | 515 | | |
| | Overall | | | 0.948 | 0.876 |

*Note*: We report the dataset properties, as well as overall and class-specific average area under the receiver operator curve ($\overline{\text{AUC}}$) and average precision ($\overline{\text{AP}}$) for the amphibian, cricket and small mammal datasets.

moderate ($\text{AUC}_{\text{macro}}$: 0.777; mAP: 0.323) and improved slightly after aggregating species-level classes by genus ($\text{AUC}_{\text{macro}}$: 0.769; mAP: 0.360). For the species-level classifier, overall performance was lowered by the poor performance of *the Bullita obscura* and *Koghiella flammea* classes (Table 5). Overall performance on the mammalian dataset was strong ($\text{AUC}_{\text{macro}}$: 0.948; mAP: 0.876), with all three classes scoring high in terms of AUC and AP (Table 5).

## 4 | DISCUSSION

The recent release of pre-trained avian classification models marks an important advancement for PAM. These models offer ready-to-use acoustic detection and classification for many vocalizing species (Kahl et al., 2021) and strong foundations for developing custom acoustic classifiers using transfer learning. Here, we demonstrate a low-cost, rapid computational workflow that leverages pre-trained models to develop custom acoustic classifiers with as few as two vocalization examples. The performance of our custom acoustic classifiers typically exceeds the off-the-shelf performance of pre-trained models targeting global sets of species and approaches the performance of specialized pre-trained local classifiers that may take years and substantial investment to build (Gibb et al., 2019). This workflow reduces reliance on large annotated datasets, expediting the time it takes to transform acoustic data into ecological insights, potentially increasing stakeholder

TABLE 6 Suggested transfer learning strategy for training custom acoustic classification models.

| Consideration | Experiment | Finding | Suggestion |
|---|---|---|---|
| Hyperparameter values | 1 | There were multiple competitive combinations of hyperparameter values | We suggest starting with large batch sizes, moderate learning rates and short training schedules. However, project-specific hyperparameter ablations may yield marginal relative performance gains |
| Data augmentation | 2 | Soundscape simulation consistently improved performance, while other augmentations had no effect | In data-scarce contexts, use soundscape simulation |
| Model architecture | 3 | Single-layer linear classifiers were surprisingly robust for all three avian datasets. However, applying a dropout prior to prediction marginally improve relative performance | Use a linear classifier directly on pre-trained embeddings. Apply dropout on the embeddings during training |
| Acoustic augmentations | 4 | Additional acoustic augmentations, such as pitch shifting and time stretching, did not increase model performance and slowed down the simulation process | Avoid adding additional acoustic augmentations to the soundscape simulation |
| Increasing the number of simulated clips | 4 | Increasing the number of simulated clips when using soundscape simulation augmentation marginally improved model performance, but the relative gains were slow after 4000 clips | Simulating at least ~4000 clips, increasing the number when practical, especially for multi-label tasks |

*Note*: We report our findings and suggestions for five transfer learning model training considerations.

engagement and the conservation impact (Makiola et al., 2020; Weiskopf et al., 2022).

Transfer learning is a promising approach for adapting pre-trained foundational models to local problems. This study provides a transfer learning strategy for developing custom acoustic classifiers (Table 6). Simple linear classifiers trained on supervised embeddings are a robust approach for developing custom classifiers to improve performance for in-domain sounds, adapt species-level predictions to within-species sound types, or classify novel sounds (Ghani et al., 2023; Kath et al., 2024). We found that the performance of simple linear classifiers trained on raw examples improved by training on simulated soundscapes, while other augmentations included in this study failed to produce consistent improvements over the baseline (Figure 2).

Adopting a transfer learning approach that leverages pre-trained classification models allows new and ongoing PAM programmes to mitigate the risks associated with developing computational processing tools by shortening the time between model training and performance feedback. This shortened feedback loop allows PAM programmes to quickly incorporate new monitoring targets or respond to changing environmental conditions. Additionally, the shorter development cycle facilitates the use of an active learning framework (Zhao et al., 2020). In an active learning framework, users start with a simple linear classifier and iteratively develop an informative local training dataset through model training, prediction and review cycles—training both 'what is' and 'what is not' an acoustic target (Williams et al., 2024). In our analysis, a local classification model had higher class-specific performance than a global model's off-the-shelf predictions for eight of 10 shared classes, revealing room for global models to improve when adapted to local problems. But in all these cases, our soundscape simulation-based transfer learning classifier, trained with a few examples, substantially narrowed the gap in performance between these two pre-trained models, and the transfer learning model will likely continue to improve after exposure to more annotated local data.

Our method is applicable to other pre-trained embedding models, including other wildlife-focused acoustic models like Perch (Ghani et al., 2023) and PNW-Cnet (Ruff et al., 2023), as well as general-purpose acoustic classification (Hershey et al., 2017; Kong et al., 2020). Embedding models map acoustic training datasets to numeric embeddings, with differences in training datasets reflected in the information content of the embeddings (Turian et al., 2022). Consequently, the effectiveness of transfer learning depends on the chosen embedding model, particularly when the target data differ from the training dataset (Williams et al., 2024). For instance, BirdNET's species-level training may cause its feature space to collapse dissimilar acoustic sounds from the same species into similar representations, limiting the utility of its embeddings to distinguish among call types within a species (Figure 4). Transfer learning applications will likely perform better when embedding models are selected based on project-specific factors, such as similarities between the training and sample data domains, alignment of model context window length with vocalization duration or the model's receptive frequency range with the target vocalizations. Nonetheless, other embedding models should be considered for complex problems, as they may yield different performance outcomes. However, special care should be given to ensure that evaluation datasets for transfer learning tasks are independent of both the transfer learning classifier's training data and that of the embedding model.

There is potential for our approach to extend beyond the original training scope of BirdNET (Table 5), enabling PAM programmes to rapidly adapt pre-trained global classifiers to local monitoring or management objectives. For in-domain (amphibian) and close-domain (small mammal) problems, transfer learning over embeddings is expected to achieve high accuracy for many classes (Table 5). However, for tasks diverging further beyond the original pre-trained model training domain (crickets)—where stridulation classification requires fine-scale differentiation of frequencies and repeated syllables—more annotated training data, alternative embedding methods (Evci et al., 2022) or deeper levels of fine-tuning on the embedding model (Dufourq et al., 2022) may be necessary to achieve comparable performance. Our findings underscore the need for further refinement and additional training data to address challenging out-of-domain classification tasks.

Our approach could be particularly impactful for regions where data are extremely scarce and off-the-shelf pre-trained model predictions are unavailable because the local species are not included in the pre-trained model datasets. Notably, many parts of the world with the highest biodiversity, which are often under the greatest threat (Betts et al., 2017; Cui et al., 2023), lack extensive annotated datasets (van Merriënboer et al., 2024). In these biodiverse yet data-poor regions, our transfer learning strategy, which utilizes a minimal number of vocalization examples and simulated soundscapes, offers a viable method for developing effective acoustic classifiers. This approach equips managers and policymakers with the necessary tools to quickly develop monitoring systems for understanding, detecting and responding to emerging biodiversity threats, and facilitates the monitoring of otherwise overlooked species due to the lack of pre-existing data, thereby supporting conservation efforts in some of the most ecologically critical areas on the planet.

Despite the potential of simulated soundscapes and transfer learning to improve acoustic classification models, several limitations remain. First, the relative performance gains of our approach may vary depending on the specific ecological context, the quality of the initial training data and the characteristics of the target species' vocalizations. For instance, species with long, highly variable or low-amplitude calls may still pose challenges for accurate classification, even with advanced augmentation techniques (Zhao et al., 2023). Furthermore, we view strong classification performance on relatively simple classification tasks (e.g. Amphibian, Marbled Murrelet and Small Mammal) as an indication of success in favourable contexts and not as evidence of robustness in all classification contexts. Second, the generalizability of our methods to different ecosystems and taxa requires further validation, particularly when the ecosystem or focal taxa are novel relative to an embedding model's training dataset (Table 5). Lastly, while our approach offers strong classification performance and significant efficiency gains in the short term, it does not replace the need for high-quality, manually annotated data. For example, in data-scarce contexts, classifier performance can vary widely across training runs (Figure 3), and it can be challenging to assess model performance. In these situations, investing resources in iterative classifier training in pursuit of a highly performant classifier could be tempting, which would likely result in a classifier over-optimized for a specific and likely small evaluation dataset. Instead, leveraging the trained classifiers to identify and annotate additional data in model-guided data review will likely result in more substantial increases in classifier performance and more relevant insights into its overall performance.

Our study demonstrates the potential for simulated soundscapes to improve the performance of acoustic classification models in contexts with limited training data. By leveraging transfer learning and our simulation-based augmentation approach, we offer an effective and efficient workflow that improves the performance of acoustic classification models and reduces the need for extensive manual data labelling. To support the application of our methods to novel classification tasks, we include a general-purpose Python script (11_new_applications.py) in the manuscript Zenodo archive and a vignette describing the application of this script to develop an acoustic classifier for golden-crowned kinglet (*Regulus satrapa*) songs using two vocalization examples extracted from a XenoCanto recording. Our findings have practical implications for PAM programmes and other domains of bioacoustic research, both enabling the rapid development of classifiers for data-deficient, rare or understudied species and facilitating fine-grained classification tasks, including vocalization-associated behaviours and spatiotemporal variation in vocalizations.

## AUTHOR CONTRIBUTIONS

Matthew J. Weldy: Conceptualization; methodology; software; validation; formal analysis; data curation; writing—original draft; visualization. Damon B. Lesmeister: Validation; resources; data curation; writing—review and editing; supervision; project administration; funding acquisition. Tom Denton: Methodology; validation; writing—original draft. Adam Duarte: Validation; resources; data curation; writing—review and editing; project administration; funding acquisition. Ben J. Vernasco and Jennifer C. Rowe: Validation; data curation; writing—review and editing. Amandine Gasc and Michael J. Adams: Validation; resources; data curation; writing—review and editing; project administration. Matthew G. Betts: Validation; resources; data curation; writing—review and editing; supervision; project administration; funding acquisition.

## ACKNOWLEDGEMENTS

Funding for this project was provided by the USDA Forest Service Pacific Northwest Research Station, Oregon State University Department of Forest Ecosystems and Society, and the USDA Forest Service Collaborative Forest Landscape Restoration Program (CFLRP). Additional support for data collection was provided by the Oregon Department of Fish and Wildlife and USDA Forest Service Wallowa-Whitman National Forest, and the National Science Foundation LTER8 DEB-2025755. We are grateful for data collection and annotation support from A. Kohlberg and J. Crawford, as well as USFS staff N. Rugg, Z. Ruff and J. Jenkins. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA Forest Service determination or policy. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the US Government. This is Contribution Number 942 of the US Geological Survey Amphibian Research and Monitoring Initiative (ARMI).

## CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

## PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70089.

## DATA AVAILABILITY STATEMENT

Evaluation data are available via https://doi.org/10.5281/zenodo.15578980 (Weldy et al., 2025). Python scripts, an environment.yml, Dockerfile and compressed docker image used to execute the analyses are also included in a compressed '.zip' folder via Zenodo at the same link (Weldy et al., 2025). The environment.yml file describes the required Python package dependencies, and the Dockerfile preserves the computational environment. We also include a general-purpose Python script (11_new_applications.py) and a vignette describing its use to develop novel classifiers beyond the experimental case-studies.

## ORCID

*Matthew J. Weldy* https://orcid.org/0000-0002-8627-2466
*Damon B. Lesmeister* https://orcid.org/0000-0003-1102-0122
*Tom Denton* https://orcid.org/0000-0003-3866-0031
*Adam Duarte* https://orcid.org/0000-0003-0034-1764
*Ben J. Vernasco* https://orcid.org/0000-0002-5561-7273
*Amandine Gasc* https://orcid.org/0000-0001-8369-4930
*Jennifer C. Rowe* https://orcid.org/0000-0002-5253-2223
*Michael J. Adams* https://orcid.org/0000-0001-8844-042X
*Matthew G. Betts* https://orcid.org/0000-0002-7100-2551

## REFERENCES

Altman, B., & Bresson, B. (2017). Conservation of landbirds and associated habitats and ecosystems in the Northern Rocky Mountains of Oregon and Washington. Version 2.0. Oregon-Washington Partners in Flight (www.orwapif.org) and American Bird Conservancy and U.S. Forest Service/Bureau of Land Management.

Betts, M. G., Wolf, C., Ripple, W. J., Phalan, B., Millers, K. A., Duarte, A., Butchart, S. H. M., & Levi, T. (2017). Global forest loss disproportionately erodes biodiversity in intact landscapes. *Nature*, *547*(7664), 441–444. https://doi.org/10.1038/nature23285

Bielinski, N., Pajda-De La, O. J., Gorniak, A., & Wise, D. (2020). Improving automated detection of frog calls in noisy urban habitats using narrow-banded recognizers. *Herpetological Conservation and Biology*, *15*(1), 1–15.

Cole, J. S., Michel, N. L., Emerson, S. A., & Siegel, R. B. (2022). Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. *Ornithological Applications*, *124*(2), duac003. https://doi.org/10.1093/ornithapp/duac003

Cui, Y., Carmona, C. P., & Wang, Z. (2023). Identifying global conservation priorities for terrestrial vertebrates based on multidimensions of biodiversity. *Conservation Biology*, *38*, cobi.14205. https://doi.org/10.1111/cobi.14205

Denton, T., Wisdom, S., & Hershey, J. R. (2022). Improving bird classification with unsupervised sound separation. *ICASSP*, *2022*, 636–640. https://doi.org/10.1109/ICASSP43922.2022.9747202

Duarte, A., Vernasco, B., Weldy, M. J., Spaan, R. S., & Ratliff, J. (2024). *Northern Blue Mountains wildlife monitoring, 2022–2023*. USDA Forest Service, Pacific Northwest Research Station.

Duarte, A., Weldy, M. J., Lesmeister, D. B., Ruff, Z. J., Jenkins, J. M. A., Valente, J. J., & Betts, M. G. (2024). Passive acoustic monitoring and convolutional neural networks facilitate high-resolution and broadscale monitoring of a threatened species. *Ecological Indicators*, *162*, 112016. https://doi.org/10.1016/j.ecolind.2024.112016

Dufourq, E., Batist, C., Foquet, R., & Durbach, I. (2022). Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics*, *70*, 101688. https://doi.org/10.1016/j.ecoinf.2022.101688

Evci, U., Dumoulin, V., Larochelle, H., & Mozer, M. C. (2022). Head2Toe: Utilizing intermediate representations for better transfer learning. *Proceedings of the 39th International Conference on Machine Learning*, *162*, 6009–6033.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gasc, A., Anso, J., Sueur, J., Jourdan, H., & Desutter-Grandcolas, L. (2018). Cricket calling communities as an indicator of the invasive ant Wasmannia auropunctata in an insular biodiversity hotspot. *Biological Invasions*, *20*(5), 1099–1111. https://doi.org/10.1007/s10530-017-1612-0

Gaylord, M., Duarte, A., McComb, B., & Ratliff, J. (2023). Passive acoustic recorders increase White-headed Woodpecker detectability in the Blue Mountains. *Journal of Field Ornithology*, *94*(4), Article 1. https://doi.org/10.5751/jfo-00330-940401

Ghani, B., Denton, T., Kahl, S., & Klinck, H. (2023). Feature embeddings from large-scale acoustic bird classifiers enable few-shot transfer learning. *Scientific Reports*, *13*(1), 22876. https://doi.org/10.1038/s41598-023-49989-z

Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, *10*(2), 169–185. https://doi.org/10.1111/2041-210X.13101

Hartig, F., Abrego, N., Bush, A., Chase, J. M., Guillera-Arroita, G., Leibold, M. A., Ovaskainen, O., Pellissier, L., Pichler, M., Poggiato, G., Pollock, L., Si-Moussi, S., Thuiller, W., Viana, D. S., Warton, D. I., Zurell, D., & Yu, D. W. (2023). Novel community data in ecology-properties and prospects. *Trends in Ecology & Evolution*, *39*, 280–293. https://doi.org/10.1016/j.tree.2023.09.017

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131–135). IEEE. http://arxiv.org/abs/1609.09430

Hill, J. E., DeVault, T. L., & Belant, J. L. (2019). Cause-specific mortality of the world's terrestrial vertebrates. *Global Ecology and Biogeography*, *28*(5), 680–689. https://doi.org/10.1111/geb.12881

Hinton, G., & Roweis, S. (2002). Stochastic neighbor embedding. *Proceedings of the 15th International Conference on Neural Information Processing Systems*, 857–864. https://doi.org/10.5555/2968618.2968725

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning* 97, 2790–2799.

Incze, A., Jancso, H.-B., Szilagyi, Z., Farkas, A., & Sulyok, C. (2018). Bird sound recognition using a convolutional neural network. *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, 000295–000300. https://doi.org/10.1109/SISY.2018.8524677

Jourdan, H., Sadlier, R. A., & Bauer, A. M. (2001). Little fire ant invasion (*Wasmannia auropunctata*) as a threat to New Caledonian lizards:

Evidence from a sclerophyll forest (Hymenoptera: Formicidae). *Sociobiology*, *38*, 283–299.

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Kath, H., Serafini, P. P., Campos, I. B., Gouvea, T. S., & Sonntag, D. (2024). Leveraging transfer learning and active learning for sound event detection in passive acoustic monitoring of wildlife. *Ecological Informatics*, *82*, 102710.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR 2015. Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA. http://arxiv.org/abs/1412.6980

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *28*, 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better ImageNet models transfer better? In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2656–2666). IEEE. https://doi.org/10.1109/CVPR.2019.00277

Lesmeister, D. B., & Jenkins, J. M. A. (2022). Integrating new technologies to broaden the scope of northern spotted owl monitoring and linkage with USDA forest inventory data. *Frontiers in Forests and Global Change*, *5*, 966978. https://doi.org/10.3389/ffgc.2022.966978

Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., Brennan, G., Bush, A., Canard, E., Cordier, T., Creer, S., Curry, R. A., David, P., Dumbrell, A. J., Gravel, D., Hajibabaei, M., Hayden, B., Van Der Hoorn, B., Jarne, P., … Bohan, D. A. (2020). Key questions for next-generation biomonitoring. *Frontiers in Environmental Science*, *7*, 197. https://doi.org/10.3389/fenvs.2019.00197

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., … Pimenta, W. (2023). *librosa/librosa: 0.10.0.post1 (0.10.0.post1)* [computer software]. https://doi.org/10.5281/zenodo.7741801

Nafis, G. (2021). *2000-2020 California herps—A guide to the amphibians and reptiles of California*. [Acoustic recordings]. http://www.californiaherps.com

Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I., Jensen, F. H., Morford, J., Emmerson, M. G., Versace, E., Grout, E., Liu, H., Ghani, B., & Stowell, D. (2023). Learning to detect an animal sound from five examples. *Ecological Informatics*, *77*, 102258. https://doi.org/10.1016/j.ecoinf.2023.102258

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1717–1724). IEEE. https://doi.org/10.1109/CVPR.2014.222

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Ruff, Z. J., Lesmeister, D. B., Jenkins, J. M. A., & Sullivan, C. M. (2023). PNW-Cnet v4: Automated species identification for passive acoustic monitoring. *SoftwareX*, *23*, 101473. https://doi.org/10.1016/j.softx.2023.101473

Salamon, J., MacConnell, D., Cartwright, M., Li, P., & Bello, J. P. (2017). Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 344–348). IEEE. https://doi.org/10.1109/WASPAA.2017.8170052

Shonfield, J., & Bayne, E. M. (2017). Autonomous recording units in avian ecological research: Current use and future applications. *Avian Conservation and Ecology*, *12*(1), art14. https://doi.org/10.5751/ACE-00974-120114

Stowell, D. (2022). Computational bioacoustics with deep learning: A review and roadmap. *PeerJ*, *10*, e13152.

Turian, J., Shier, J., Khan, H. R., Raj, B., Schuller, B. W., Steinmetz, C. J., Malloy, C., Tzanetakis, G., Velarde, G., McNally, K., Henry, M., Pinto, N., Noufi, C., Clough, C., Herremans, D., Fonseca, E., Engel, J., Salamon, J., Esling, P., … Bisk, Y. (2022). HEAR: Holistic evaluation of audio representations. *arXiv*, arXiv:2203.03022. http://arxiv.org/abs/2203.03022

van der Maaten, L., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

van Merriënboer, B., Hamer, J., Dumoulin, V., Triantafillou, E., & Denton, T. (2024). Birds, bats and beyond: Evaluating generalization in bioacoustics models. *Frontiers in Bird Science*, *3*. https://doi.org/10.3389/fbirs.2024.1369756

Vellinga, W.-P., & Planque, R. (2015). The Xeno-canto collection and its relation to sound recognition and classification. *2015* CLEF, 10.

Wattenberg, M., Viegas, F., & Johnson, I. (2016). How to use t-sne effectively. *Distill*, *1*(10), e2.

Weiskopf, S. R., Harmáčková, Z. V., Johnson, C. G., Londoño-Murcia, M. C., Miller, B. W., Myers, B. J. E., Pereira, L., Arce-Plata, M. I., Blanchard, J. L., Ferrier, S., Fulton, E. A., Harfoot, M., Isbell, F., Johnson, J. A., Mori, A. S., Weng, E., & Rosa, I. M. D. (2022). Increasing the uptake of ecological model results in policy decisions to improve biodiversity outcomes. *Environmental Modelling & Software*, *149*, 105318. https://doi.org/10.1016/j.envsoft.2022.105318

Weldy, M. J., Denton, T., Fleishman, A. B., Tolchin, J., McKown, M., Spaan, R. S., Ruff, Z. J., Jenkins, J. M. A., Betts, M. G., & Lesmeister, D. B. (2024). Audio tagging of avian dawn chorus recordings in California, Oregon and Washington. *Biodiversity Data Journal*, *12*, e118315. https://doi.org/10.3897/BDJ.12.e118315

Weldy, M. J., Lesmeister, D. M., Denton, T., Duarte, A., Varnasco, B. J., Gasc, A., Rowe, J. C., Adams, M. J., Anso, J., Desutter-Grandcolas, L., Hervé, J., & Betts, M. G. (2025). Simulated soundscapes and transfer learning boost the performance of acoustic classifiers under data scarcity [Datasets]. *Zenodo*, https://doi.org/10.5281/zenodo.15578980

Williams, B., van Merriënboer, B., Dumoulin, V., Hamer, J., Triantafillou, E., Fleishman, A. B., McKown, M., Munger, J. E., Rice, A. N., Lillis, A., White, C. E., Hobbs, C. A. D., Razak, T. B., Jones, K. E., & Denton, T. (2024). Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. *arXiv*. https://doi.org/10.48550/arXiv.2404.16436

Wood, C. M., & Kahl, S. (2024). Guidelines for appropriate use of BirdNET scores and other detector outputs. *Journal of Ornithology*, *165*, 777–782. https://doi.org/10.1007/s10336-024-02144-5

Zhao, S., Heittola, T., & Virtanen, T. (2020). Active learning for sound event detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *28*, 2895–2905. https://doi.org/10.1109/TASLP.2020.3029652

Zhao, Z., Yang, L., Ju, R., Chen, L., & Xu, Z. (2023). Acoustic bird species classification under low SNR and small-scale dataset conditions. *Applied Acoustics*, *214*, 109670. https://doi.org/10.1016/j.apacoust.2023.109670

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**Table S1:** Description of the discrete audio examples from the training and evaluation datasets, including passive acoustic recordings (PAM) and simulated soundscapes (sim).

**Table S2:** Description of BirdNET's embedded values of the training and evaluation datasets, including passive acoustic recordings (PAM) and simulated soundscapes (sim).

**Table S3:** Hyperparameter ablation for simulation-based transfer learning classifiers built over BirdNET embeddings for the three avian datasets.

**Table S4:** Standard deviation ($\sigma$) of macro averaged Area Under the ROC curve and mean Average Precision among ten replicate trainings for four training data strategies for developing acoustic classification algorithms using transfer learning under data scarcity.

**Figure S1:** t-distributed stochastic neighbor embedding (t-SNE) plots of the 3-s avian evaluation datasets demonstrating the effect of increasing the perplexity value.

**How to cite this article:** Weldy, M. J., Lesmeister, D. B., Denton, T., Duarte, A., Vernasco, B. J., Gasc, A., Rowe, J. C., Adams, M. J., & Betts, M. G. (2025). Simulated soundscapes and transfer learning boost the performance of acoustic classifiers under data scarcity. *Methods in Ecology and Evolution*, *00*, 1–17. https://doi.org/10.1111/2041-210X.70089