The impacts of transfer learning, phylogenetic distance, and sample size on big-data bioacoustics

Kaiya L. Provost[1*], Jiaying Yang[1], Bryan C. Carstens[1]

1: Department of Evolution, Ecology and Organismal Biology, The Ohio State University. 318 W. 12th Ave. 300 Aronoff Laboratory Columbus, OH 43210

*: Corresponding author: Kaiya L. Provost, provost.27@osu.edu, Department of Evolution, Ecology and Organismal Biology, The Ohio State University. 318 W. 12th Ave. 300 Aronoff Laboratory Columbus, OH 43210

Running Title: Machine learning in bioacoustics

Abstract:

1. Vocalizations in animals, particularly birds, are critically important behaviors that influence their reproductive fitness, but automatically extracting vocalization data from existing large databases has only recently gained traction and has yet to be evaluated with respect to accuracy of different approaches.
2. Here, we use a recently-published machine learning framework to extract syllables from six bird species ranging in their phylogenetic relatedness from 1–85 million years, comparing how phylogenetic relatedness impacts accuracy as well as the utility of applying trained models to novel species.
3. Model performance is best on conspecifics, with accuracy progressively decreasing as phylogenetic distance increases between taxa; however, using models trained on multiple distantly related species can recover the lost accuracy.
4. When planning big-data bioacoustics studies, care must be taken in sample design to maximize sample size and minimize human labor without sacrificing accuracy.

Keywords: bioacoustics, birds, machine learning, phylogenetics, transfer learning

Introduction:

Avian bioacoustics

Vocalization is an important form of communication across many taxa. Birds (class Aves) make and respond to vocalizations profusely, for example in territory defense (de Kort et al., 2009), mate choice (Nowicki & Searcy, 2005; O'Loghlen & Rothstein, 1995; West et al., 1981), both conspecific and heterospecific communication (Magrath et al., 2015; Riebel et al., 2019), species discrimination (Gill & Lanyon, 1964; Derryberry, 2011), prey detection (Rice,

1982), and food solicitation via begging calls (Wright, & Leonard 2002; Catchpole & Slater, 1995). Bird vocalizations are generally classified into songs and calls. Significant attention has been paid to song vocalizations, which are thought to mediate sexual selection (MacDougall-Shackleton, 1997; Tinghitella et al., 2018; Burdfield-Steel & Shuker, 2018) and be important in the generation of reproductive isolation in many bird taxa (Brambilla et al., 2008; Jenkins, 1978; Uy et al., 2018; Dingle et al., 2010). Though less prominent, bird calls are also important as they can mediate non-sexual behaviors (Kondo & Watanabe, 2009), for instance between parents and offspring (Jiang et al., 2021) or even intra-specific communication (Keen et al., 2020). Despite evidence that both the biotic and abiotic environment can impact the development and evolution of vocalization, there are few studies that examine how the community soundscape evolves in tandem (but see Mason et al., 2017; Freeman et al., 2022).

The sheer volume of recording data available to modern researchers represents a tremendous asset to avian bioacoustic research. For example, the citizen-science initiative Xeno-Canto (xeno-canto.org) having ~700,000 recordings and the Macaulay Library bioacoustics repository (which includes eBird, Sullivan et al., 2009) having ~1.1 million recordings (birds.cornell.edu/MacaulayLibrary). Even the smaller bioacoustic collections, which represent important repositories of historical recordings, can have tens of thousands of songs to parse (e.g., the Borror Laboratory of Bioacoustics with ~36,000, blb.osu.edu). This volume of data represents a methodological challenge because most bioacoustic workflows require the songs to be segmented (Catchpole & Slater, 1995), either identifying them from background sounds or, for more complex analyses, identifying individual syllables of song (e.g., Du & Troyer, 2009; Alosnso et al., 2017).

Ideally, automated methods would be useful if they could reduce the amount of human interaction without compromising data quality, as song segmentation is extremely time consuming (e.g., Swiston et al., 2009 found manual annotation to take nearly five times as long, even after accounting for needing to check automatic annotations), though less so than the gold-standard methods of performing playback experiments to directly evaluate behavioral responses to different vocalization treatments (e.g., Provost et al., 2018 performed ~4,600 minutes of playback experiments to investigate a single species).

Machine Learning

Machine learning, which can be either unsupervised or supervised, is a useful tool for data processing. Unsupervised machine learning is typically used to cluster data (Reynolds & Rose, 1995, Greenspan et al., 2006, Edgington et al., 2006), while supervised learning is frequently used to make predictions (Heileman & Myler, 1990). Many well-known algorithms in ecology and evolution are derived from machine learning models in some capacity, for example linear regression, k-means clustering (Lloyd, 1982), and MaxEnt niche modeling (Phillips et al., 2006, Phillips et al., 2017). However, many of these techniques require inputs that are specifically curated, such that the presence or absence of features and variables must be determined beforehand, which can lead to a slew of problems.

More recently, deep learning via artificial neural networks (ANNs) has become popular. ANNs are made from many layers of nodes which construct weighted mathematical functions to perform tasks, typically from some kind of input data with diagnostic features (LeCun et al., 2015). Deep learning methods are methods where the features used do not need to be pre-trained. Because ANNs can iteratively learn their own features, use large amounts of data, and are less

constrained by assumptions about that data, they are extremely flexible and can handle many kinds of tasks (e.g., Hopfield. 1982; Ripley. 1996, Sprengel et al., 2016; Zell, 1994; O'Mahony et al., 2020; LeCun et al., 2015; Alom et al., 2018). These methods have been used on a variety of topics including image processing, video segmentation, and speech recognition (Hemanth & Estrela, 2017; Valipour & Wang, 2017; Kamath et al., 2019). Although challenges remain with respect to scalability, computational efficiency, and how to handle depauperate data (Alom et al., 2018), deep learning is one of the most powerful analytical tools in the modern researcher's toolbox, particularly when human knowledge is lacking, or datasets are too large to be workable by traditional means. In the context of ecology and evolutionary biology, there have been many recent applications of both shallow and deep machine learning, including population genetics and phylogeography (e.g., Schrider & Kern, 2018; Provost et al., 2021), bioacoustics (e.g., Nicholson, 2016; Zhong et al., 2020; Cohen et al., bioRxiv), species classification (e.g., Gupta et al., 2021), phylogenetics (e.g., Halgaswaththa et al., 2012; Wang et al., 2016), sequencing and genomics (e.g., Adrion et al., 2020; Boža et al., 2017), and phenotypic analyses and morphometrics (e.g., Devine et al., 2020; Lürig et al., 2021). Neural networks and support vector machines tend to be the most applied algorithms towards these analyses.

Though there are many ways to categorize machine learning algorithms, but for our purposes dividing these into evaluative vs extractive algorithms is convenient. Evaluative algorithms are typically used to make predictions from data, for example by training a model to distinguish between simulated evolutionary scenarios (Pudlo et al., 2016; Sheehan & Song, 2016; Schrider & Kern, 2018; Gehara et al., bioRxiv; Provost et al., 2021). Extractive algorithms, on the other hand, are designed to process data in some way. With respect to bioacoustics, an extractive algorithm could, for example, segment out syllables within

vocalizations. A recently developed application, TweetyNet, was released to perform just this task (Cohen et al., bioRxiv) using deep learning via ANNs. Specifically, TweetyNet uses convolutional and recurrent ANNs. As such, it is one extractive algorithm that could be useful in reducing the human workload in bioacoustics. However, though potentially much faster than hand-processing data, ANNs like TweetyNet can still take prohibitively long times to train, though they often only need to be trained once.

Minimizing training time can be done by co-opting existing trained ANNs to perform similar tasks, known as transfer learning. This can be used to take a neural network trained on one batch of data, and then applying it to new, related data (Alom et al., 2018; Weiss et al., 2016). For example, a neural network that is trained to identify the leaves of oak trees from images may be able to be co-opted to identify leaves of maple trees from images without having to re-train the model to recognize a generic leaf shape. As an example, which is more applicable for bioacoustics, a neural network that is trained to identify the song of one species against background noise may be able to be applied to a different species.

Here we compare different treatments of bioacoustic data to optimize the overall performance of machine learning models trained to segment syllables of avian song. We investigate the utility of these methods across a subset of bird species, including birds from different clades ranging in divergence times from ~1 to ~85 million years. We use those taxa to evaluate how well models perform on closely vs distantly related taxa for use in maximizing computational efficiency. Finally, we provide suggested practices for moving forward with neural networks in avian bioacoustics.

Materials and Methods:

Unless otherwise noted, all analyses were performed using a MacBook Pro running macOS Catalina (10.15.7) with a 2.8 GHz i7 core and 16 GB of RAM.

Acquiring and filtering song data

We downloaded bioacoustics data from two song databases: Xeno-Canto (xeno-canto.org), and the Borror Laboratory of Bioacoustics (blb.osu.edu). We used the Xeno-Canto database as it has a large amount of data of varying quality, due to it being a citizen science initiative, and a convenient R package exists for accessing and filtering the files. The Borror Lab of Bioacoustics database was used because it is full of historical data and based in North America, the region we wished to focus on, in addition to being the lab where the authors are based (Figure 1).

For the Xeno-Canto songs, we targeted songs from *Cardinalis sinuatus*, *Cardinalis cardinalis*, *Zonotrichia leucophrys*, *Empidonax virescens*, and *Calypte anna*. We chose these taxa because they were distributed across the phylogenetic tree and had relatively high numbers of recordings available. We also targeted songs from *Melozone fusca* which we used to evaluate the performance on a novel species. We restricted the data downloaded to all of those in North America that were of Quality "A" according to the Xeno-Canto database using the package warbleR version 1.1.26 (Araya-Salas & Smith-Vidaurre, 2017) in R version 4.0.3 (R Core Team 2020). Xeno-Canto recordings are available in the MP3 format, but since our downstream pipeline requires WAV format, we converted all of them using the package tuneR version 1.3.3 (Ligges et al., 2018). For the Borror Laboratory of Bioacoustics songs, we downloaded all available data from all species, which were already available as WAV format. Stereo recordings were converted to mono using the 'mono' function in tuneR. We then retained recordings that

had a sample rate of 48,000 Hz, which was the most common sample rate in our dataset, because downstream analyses require recordings of the same sample rate and will not work on stereo recordings.

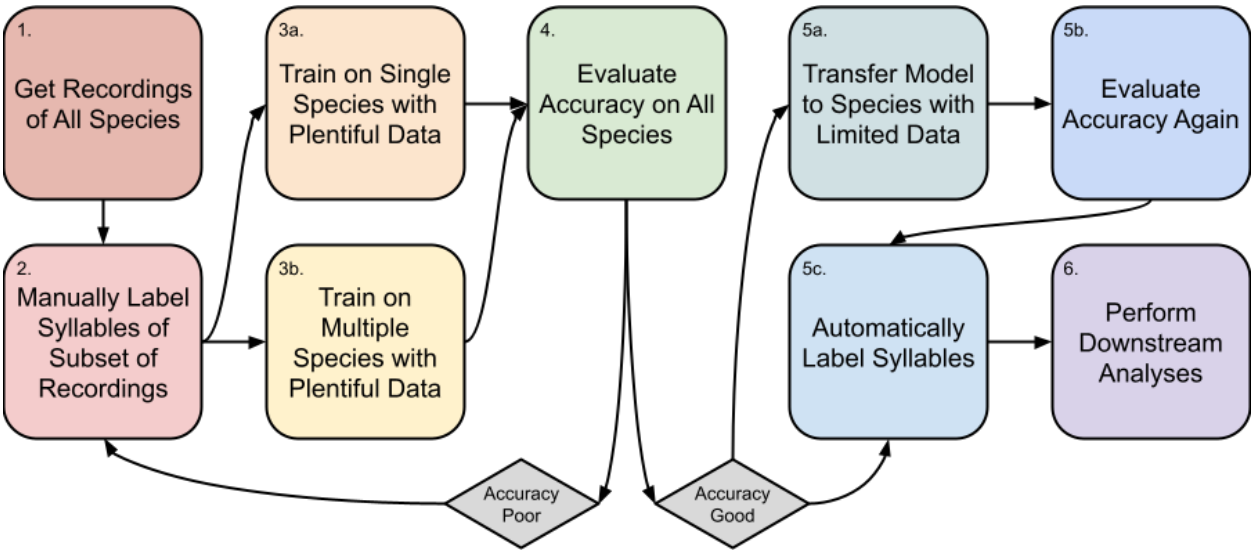

Figure 1: Flowchart describing the procedure from acquiring song recordings to having segmented syllables for downstream analysis. Boxes show general steps, arrows show flow of steps, and diamonds show decision point for data quality.

Identification of syllables via manual and automatic annotation of song data

We arbitrarily chose a subset of recordings to annotate from each species, starting with *Zonotrichia leucophrys* which had the largest dataset, and moving down to *Calypte anna* which had the smallest dataset. As few as three minutes of recording is sufficient for the TweetyNet algorithm to perform accurately (Cohen et al., bioRxiv), so we ensured that all species had at least 180 seconds worth of annotated syllables. We then added more recordings for species with more data to investigate the influence of sample size. For *Melozone fusca*, we deliberately annotated (and trained on) a disproportionately small number of songs to generate artificial scarcity of data. In both the Borror Lab of Bioacoustics and the Xeno-Canto database, many

species have a small number of recordings. For the former, which primarily sampled North American taxa, nearly 9,400 bird species lack any data, with an additional ~400 species with fewer than 10 recordings. For the latter, ~600 species lack any data, with an additional ~2,500 with fewer than 10 recordings (Figure 2). As such, having methods which can handle species with small numbers of recordings is critical.

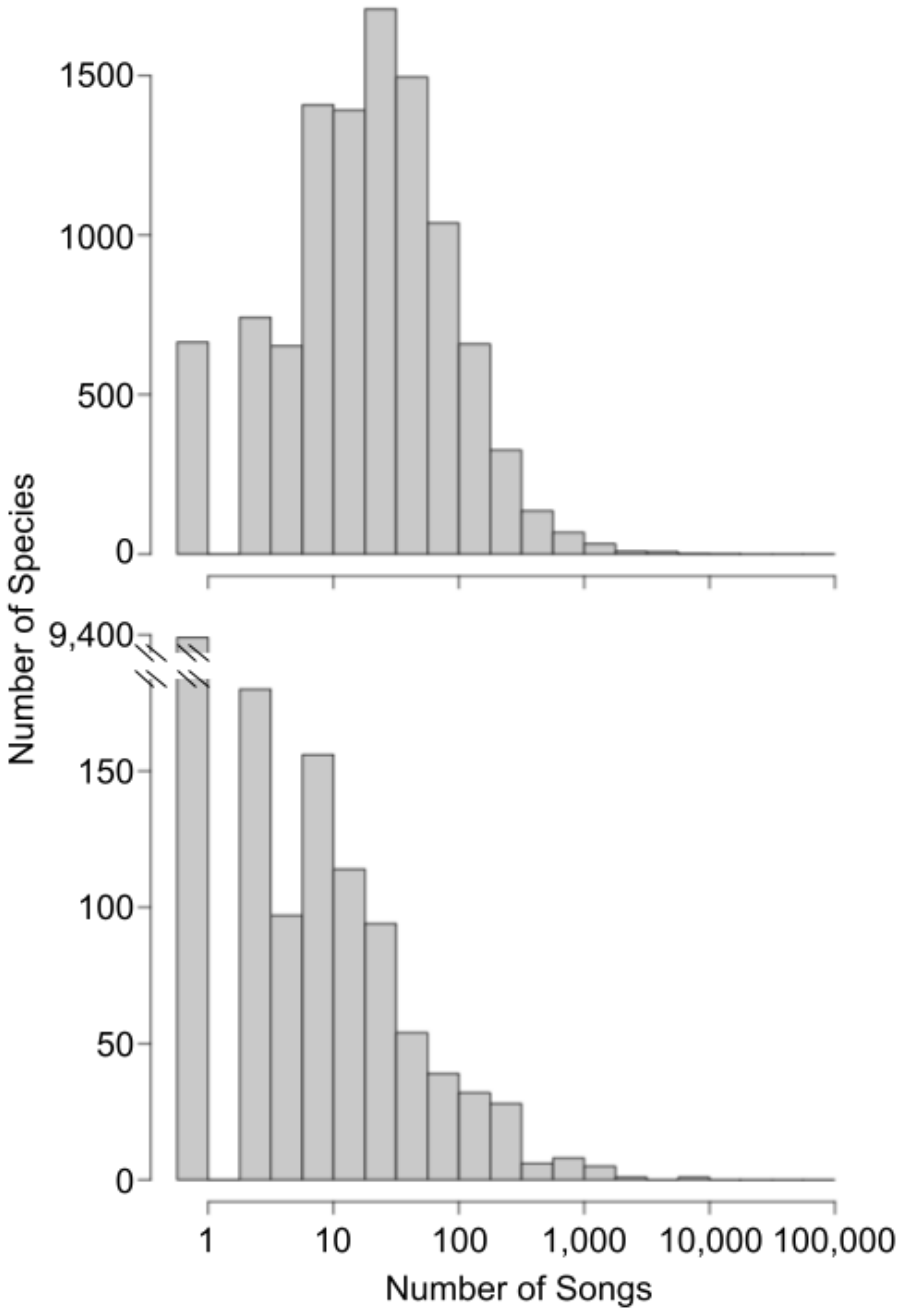

Figure 2: Most species of bird have little bioacoustic data. X-axis shows, on a log scale, the number of songs present per species. Y-axis shows the number of species per song bin. Top panel: Xeno-Canto data. Bottom panel: Borror Laboratory of Bioacoustics (BLB) data; note the x-axis has a gap to show an exceedingly high number of species with no data, as BLB is North America focused.

Annotations were performed in Raven Pro version 1.6.1 using a spectral window size (i.e., Fast Fourier Transform size) of 512 samples, which for a sample rate of 48,000 Hz was ~9.97 milliseconds. We drew boxes representing frequency and time boundaries for every syllable in the recording of the focal species. These annotations were exported as text files using Raven's built-in selection format. After annotating the songs, we then subsetted the recordings such that individual recordings had approximately equal amounts of annotated song and silence. Preliminary results showed that when TweetyNet was trained on full files, which had a high proportion of non-annotated "silence", TweetyNet would fail to identify the songs and would only predict silence, but with high accuracy. As such, we wrote a script using tuneR in R which automatically sliced WAV files and their corresponding annotation files for training (see Data Availability). Annotations that were separated by at least 1.0 second of non-annotated sound were split. Splits were made such that at least 50% of the resulting file was annotated: in cases where this was not possible, a larger percentage of the file may have been annotated. We then converted the annotation files into XML files that TweetyNet could use following the "Koumura" format (following Koumura & Okanoya, 2016). All annotations were given the label "1".

Training of machine learning model using TweetyNet

When training TweetyNet, we set our spectrogram parameters as follows: Fast Fourier Transform size of 512, step size of 32 samples, frequency cutoffs of 500 and 10,000 Hz, a log transformed spectrogram, and a minimum power log threshold of 6.25. We also used a window size of 88 spectral windows and normalized spectrograms while training. Unless otherwise noted, we used an 80:10:10 split for training, validation, and test data, and tracked the total

training time for each model. These splits were generated manually for each species. Each model used a batch size of three and was validated every 50 steps. Checkpoints were taken every 200 steps. We ran for 10 epochs (i.e., 10 passes the machine learning algorithm made over the full training dataset) but stopped training prematurely if the model had gone 50 checkpoint steps without an improvement in accuracy. We then used the maximum accuracy checkpoint as our final trained model. Some models were restarted if they were killed before completing 50 checkpoints without improvement. Models that failed this process multiple times were restarted and ran as needed. Models were run sequentially (i.e., one at a time).

We trained three sets of models: single-species, multi-species, and transfer learning. For the single-species models, we trained a model in TweetyNet on a given species to 5-step and 50-step checkpoints. We did this once for each of our five species with sufficient data, resulting in ten total single-species models. For the multi-species models, we trained a model in TweetyNet as above but using a combination of the five species with sufficient available data. We did this for two models: an unbalanced and a balanced model. For the unbalanced, we included all the training data from each of the five species irrespective of the sample sizes for each. For the balanced model, we subset the training and validation data used such that each species had the same amount of training and validation data (in seconds) as our species with the smallest sample size, to mitigate any effects of biased sampling. Each was trained to 5-step and 50-step checkpoints, resulting in four total multi-species models.

To incorporate transfer learning for use in the transfer models, we wrote a custom module for TweetyNet ("train_checkpoint") which was able to continue training from an existing checkpoint file from a previously completed trained model using new training data from a different taxon. At the time of writing, this functionality has been planned but not executed in

TweetyNet's source code. After training both multi-species models as above, we then used that finished checkpoint to train on data from *Melozone fusca*, which we manipulated to use a relatively small amount of data (aka "few-shot learning"; Wang et al., 2019). Instead of training for 50 checkpoints which would likely lead to overfitting, we trained for 5 checkpoints only. Though we had a comparable amount of data for this species compared to the other five, we generated artificial scarcity by only using ~30% of this as training and validation data, with the remainder acting as test data. This resulted in four total transfer-learning models depending on whether we trained from the balanced or unbalanced multi-species model, and whether we trained from the 5-step or 50-step checkpoint of those models.

We tested the single-species and multi-species models on annotated data from each of the five species we trained on with sufficient available data (*Zonotrichia*, *Cardinalis*, *Empidonax*, and *Calypte* species; either zero-shot or normal training) as well as the species with artificially insufficient available data (*Melozone fusca*; zero-shot training). For the transfer models, we only tested on *Melozone fusca* (few-shot training). In all cases, we did this testing to assess the accuracy of the model. We evaluated accuracy across species with respect to the predicted genetic distance. In addition, we evaluated the performance with respect to different transfer methods, the number of species present in the training dataset, the balancing approach for multi-species models, and the impact of zero-shot vs few-shot vs normal training.

For some analyses we examined the phylogenetic distance between trained species and test species. We mined these data from timetree.org (Kumar et al., 2017), taking the estimated time listed as the time of divergence in millions of years (Mya). Intraspecific comparisons were given an estimated time of 0.0 Mya; we also did this for the transfer learning done with *Melozone fusca*. For the multiple-species models, we calculated the weighted average estimated

divergence time based on the sample sizes for each species, again taking conspecifics as 0.0 Mya.

Song diversity

We evaluated song diversity to determine whether this impacted our machine learning model accuracy. Song diversity was evaluated in two ways: first, we consulted an analysis that focused on Passeriformes which measured song complexity (Medina & Francis, 2012). Three of the species had complexity analyzed in that study (*Zonotrichia leucophrys*, *Cardinalis cardinalis*, *Cardinalis sinuatus*) which were calculated by performing a PCA on song properties (e.g., number of notes, frequency bandwidth). We directly used those values as our complexity measure. Our second metric of song diversity was done by estimating the hypervolume complexity. To do this, we took the 5 longest WAV files for each species that had been manually annotated. We then put them through the SoundShape pipeline (Rocha & Romano, 2021). SoundShape aligns and normalizes syllables of sound to make them directly comparable with respect to their size and shape. After aligned and normalized syllables were converted to TPS format, we randomly selected 50 syllables per species and calculated a principal components analysis on their values, retaining the first three principal components (PCs). We then used the three PCs to calculate a hypervolume per species, as a percentage of the total hypervolume occupied by all species. We repeated this 50 times and then calculated the mean percentage hypervolume occupied per species as our complexity metric. Species with more diverse syllable types, and therefore more song diversity, should have larger hypervolumes.

Impact of various metrics on accuracy and time of training

To investigate what aspects of our models determined accuracy and training time, we used generalized linear regressions as well as ANOVA and Tukey's Honest Significant Differences tests (Chambers et al., 1992; Miller, 1981; Yandell, 1997) using the stats package in R. For all models, we took the log values of training time and sample size, as these varied by orders of magnitude across taxa and models. Unless otherwise specified, we set our alpha level to 0.05 to determine significance, and we applied linear models for 5-step and 50-step checkpoint runs independently.

We tested the relationship between the size of the training dataset and the amount of training time using a linear regression. We did this both with and without transfer learning models. We also tested the relationship between accuracy and phylogenetic distance, accuracy and song complexity, and/or accuracy and hypervolume diversity as linear regressions both for single-species models only and for all models. The linear regressions involving song complexity and hypervolume diversity were run calculating these metrics for the testing species as well as the training species. Finally, we used an ANOVA to determine whether individual species differed in accuracy as well.

Results:

Song data summary

Our dataset included 6,563 *Zonotrichia leucophrys* songs, 918 *Cardinalis cardinalis* songs, 36 *Cardinalis sinuatus* songs, 128 *Empidonax virescens* songs, 102 *Calypte anna* songs, and 72 *Melozone fusca* songs. We filtered out songs that did not have a sample rate of 48,000 Hz and then annotated 197 *Zonotrichia leucophrys* songs, 61 *Cardinalis cardinalis* songs, 8 *Cardinalis sinuatus* songs, 22 *Empidonax virescens* songs, 15 *Calypte anna* songs, and 12

Melozone fusca songs for each species. After subsetting files to balance the relative amount of annotated sound to silence, this resulted in total datasets of 4,843 seconds for *Zonotrichia leucophrys*, 1,083 seconds for *Cardinalis cardinalis*, 401 seconds for *Cardinalis sinuatus*, 291 seconds for *Empidonax virescens*, 316 seconds for *Calypte anna*, and 272 seconds for *Melozone fusca* (see Supplementary Table S1).

Table 1: Models trained in TweetyNet across multiple species vary in training time, dataset size, and type of mode. “5 Species” models were trained on all data from the “Single species” type models. “Balanced” models were trained on equal amounts of data from each of the “Single species” type models.

Species Trained On	**Training Time, 5 checkpoints (sec)**	**Training Time, 50 checkpoints (sec)**	**Training : Val : Test (sec)**	**Type**
Calypte anna	310	4,506	256:29:31	Single species
Cardinalis cardinalis	377	12,464	867:108:108	Single species
Cardinalis sinuatus	468	7,484	320:40:41	Single species
Empidonax virescens	647	7,941	231:31:29	Single species
Zonotrichia leucophrys	1,999	44,694	3878:483:482	Single species
5 Species	4,214	25,863	5553:691:692	Multiple species
5 Species Balanced	557	10,348	1130:150:671	Multiple species
5 Species (5) + *Melozone fusca*	665	n/a	71:9:192	Transfer
5 Species Balanced (5) + *Melozone fusca*	553	n/a	71:9:192	Transfer
5 Species (50) + *Melozone fusca*	341	n/a	71:9:192	Transfer
5 Species Balanced (50) + *Melozone fusca*	246	n/a	71:9:192	Transfer

Table 2: Accuracy of models trained in TweetyNet across species, excluding *Melozone fusca*. Accuracy values are given for models checkpointed at 5-steps and 50-steps. Estimated divergence time for "5 Species" models is a weighted average, proportional to the sample sizes used for each species trained on.

Species Trained On	**Species Tested On**	**Accuracy 5-step checkpoints**	**Accuracy 50-step checkpoints**	**Estimated Divergence (Mya)**
Calypte anna	*Calypte anna*	0.56	0.60	0.0
Calypte anna	*Cardinalis cardinalis*	0.71	0.67	85.0
Calypte anna	*Cardinalis sinuatus*	0.76	0.74	85.0
Calypte anna	*Empidonax virescens*	0.63	0.69	85.0
Calypte anna	*Zonotrichia leucophrys*	0.82	0.82	85.0
Cardinalis cardinalis	*Calypte anna*	0.50	0.52	85.0
Cardinalis cardinalis	*Cardinalis cardinalis*	0.67	0.91	0.0
Cardinalis cardinalis	*Cardinalis sinuatus*	0.87	0.92	6.6
Cardinalis cardinalis	*Empidonax virescens*	0.70	0.88	66.0
Cardinalis cardinalis	*Zonotrichia leucophrys*	0.57	0.87	38.0
Cardinalis sinuatus	*Calypte anna*	0.67	0.62	85.0
Cardinalis sinuatus	*Cardinalis cardinalis*	0.87	0.89	6.6
Cardinalis sinuatus	*Cardinalis sinuatus*	0.92	0.92	0.0
Cardinalis sinuatus	*Empidonax virescens*	0.80	0.79	66.0
Cardinalis sinuatus	*Zonotrichia leucophrys*	0.69	0.81	38.0
Empidonax virescens	*Calypte anna*	0.59	0.55	85.0
Empidonax virescens	*Cardinalis cardinalis*	0.65	0.71	66.0
Empidonax virescens	*Cardinalis sinuatus*	0.73	0.76	66.0
Empidonax virescens	*Empidonax virescens*	0.93	0.97	0.0
Empidonax virescens	*Zonotrichia leucophrys*	0.65	0.73	66.0
Zonotrichia leucophrys	*Calypte anna*	0.52	0.57	85.0
Zonotrichia leucophrys	*Cardinalis cardinalis*	0.82	0.76	38.0
Zonotrichia leucophrys	*Cardinalis sinuatus*	0.82	0.81	38.0
Zonotrichia leucophrys	*Empidonax virescens*	0.86	0.92	66.0
Zonotrichia leucophrys	*Zonotrichia leucophrys*	0.89	0.92	0 .0
5 Species	*Calypte anna*	0.68	0.77	81.1
5 Species	*Cardinalis cardinalis*	0.87	0.89	33.6
5 Species	*Cardinalis sinuatus*	0.87	0.91	34.2
5 Species	*Empidonax virescens*	0.90	0.94	64.1
5 Species	*Zonotrichia leucophrys*	0.90	0.91	14.8
5 Species Balanced	*Calypte anna*	0.69	0.76	68.9
5 Species Balanced	*Cardinalis cardinalis*	0.86	0.90	39.7
5 Species Balanced	*Cardinalis sinuatus*	0.87	0.92	39.9
5 Species Balanced	*Empidonax virescens*	0.86	0.95	57.6
5 Species Balanced	*Zonotrichia leucophrys*	0.81	0.88	41.3

Machine model performance on annotated song

Training time of the single species models to 5 checkpoints ranged from 310 seconds (~5.2 minutes, *Calypte anna*, 256 seconds of song) to 1,999 seconds (~33.3 minutes, *Zonotrichia leucophrys*, 3,878 seconds of song; Table 1). For 50 checkpoints, these times ranged from 4,506 seconds (~75.1 minutes or ~1.3 hours) to 44,694 seconds (~744.9 minutes or ~12.4 hours). The multiple species models took 4,214 seconds (~70.2 minutes or ~1.2 hours) for the unbalanced (5,553 seconds of song) and 557 seconds (~9.3 minutes) for the balanced (1,130 seconds of song) to reach 5 checkpoints. For 50 checkpoints these numbers were 25,863 seconds (~431 minutes or ~7.2 hours) for unbalanced and 10,348 seconds (~172.4 minutes or ~2.9 hours) for balanced. The four transfer learning models, which were all trained for 5 checkpoints each, took between 246 seconds (~4.1 minutes) and 665 seconds (~11.1 minutes) to train on the *Melozone fusca* data (71 seconds). Less additional training time was required when training using the 50-step checkpoints than the 5-step checkpoints, though total training time was higher. The relationship between the size of the training dataset and the amount of training time needed is positive on a log scale for both the 5-step and 50-step checkpoint runs; this holds even after excluding transfer learning models which were only run for 5-step checkpoints (adjusted R^2 range=0.51–0.77, $p<0.021$; Table 1).

Accuracy in the 50-step checkpoint runs was nearly always equal or higher than accuracy in the 5-step checkpoint runs. Across all 42 pairs of single-species and multiple-species models, 32/42 had higher accuracy in the 50-step checkpoint (difference range=1–30%), 3/42 had the same accuracy, and 7/42 had higher accuracy in the 5-step checkpoint (difference range=1–6%).

For models that were trained on single species, accuracy was generally highest when models were tested on the same species that they were trained on (5-step range=89–93%; 50-step

range=91–97%), except for *Cardinalis cardinalis* in the 5-step checkpoints (in which accuracy was highest for its congener *Cardinalis sinuatus*, 87% vs 67% for *Cardinalis cardinalis*) and *Calypte anna* in both 5-step and 50-step checkpoints (where highest accuracy was also for *Cardinalis sinuatus*, 62–67% vs 56–60% for *Calypte anna*).

Accuracy dropped off in a manner that scaled to the predicted phylogenetic distance in the models that were not trained and tested on the same taxa (i.e., zero-shot training), with loss of 0–5% for congeneric comparisons (plus one gain of 20% in *Cardinalis cardinalis*), 4–32% for taxa with the same oscine/suboscine status (plus one gain of 15% in *Cardinalis cardinalis*), 2–24% for comparisons between birds in the same order, and 3–30% between passerines and non-passerines (with gains of 2–11% in *Cardinalis cardinalis* and *Calypte anna*). This relationship is significantly negative with respect to estimated phylogenetic distance for both 5-step and 50-step models (adjusted R^2 range=0.27–0.35, $p<0.0002$; Figure 3).

Most of the multi-species models perform the same or worse at classifying a species than the corresponding single-species model trained on that species (0–8% loss in accuracy across all taxa), except for the *Cardinalis cardinalis* and *Calypte anna* models mentioned above (12–20% gain in accuracy) and one *Zonotrichia leucophrys* 5-step checkpoint model (with a 1% gain). However, multi-species models tend to perform better than zero-shot learning on other single-species models (improvement of 0–33%) with a few exceptions that are worse (by 1–3%). Multiple species models perform better than expected for their mean estimated divergence time: of the 24 tests done across the models, 21 are above the line of best fit and three are below (Figure 3).

Comparing the 12 unbalanced and balanced multi-species model pairs, the differences in performance within species ranged from 4% better for balanced models to 9% better in

unbalanced models, with 8/12 of those pairs differing by less than 1%. *Zonotrichia leucophrys*, the most over-represented species, has the unbalanced models performing much better (by 3–9%).

There is no significant association between sample size of either training or test data and average accuracy across single-species models (log sample sizes, p-value range=0.33–0.53), although when multiple-species and transfer learning models are included, there is a slight positive association with test data sample size and accuracy (p-value range=0.009–0.03, adjusted R^2 range=0.09–0.12). However, *Calypte anna* vocalizations are significantly less likely to be classified accurately than other species (p-value<0.03). Notably, this is the only non-passerine in the dataset and has a very different type of song than the others.

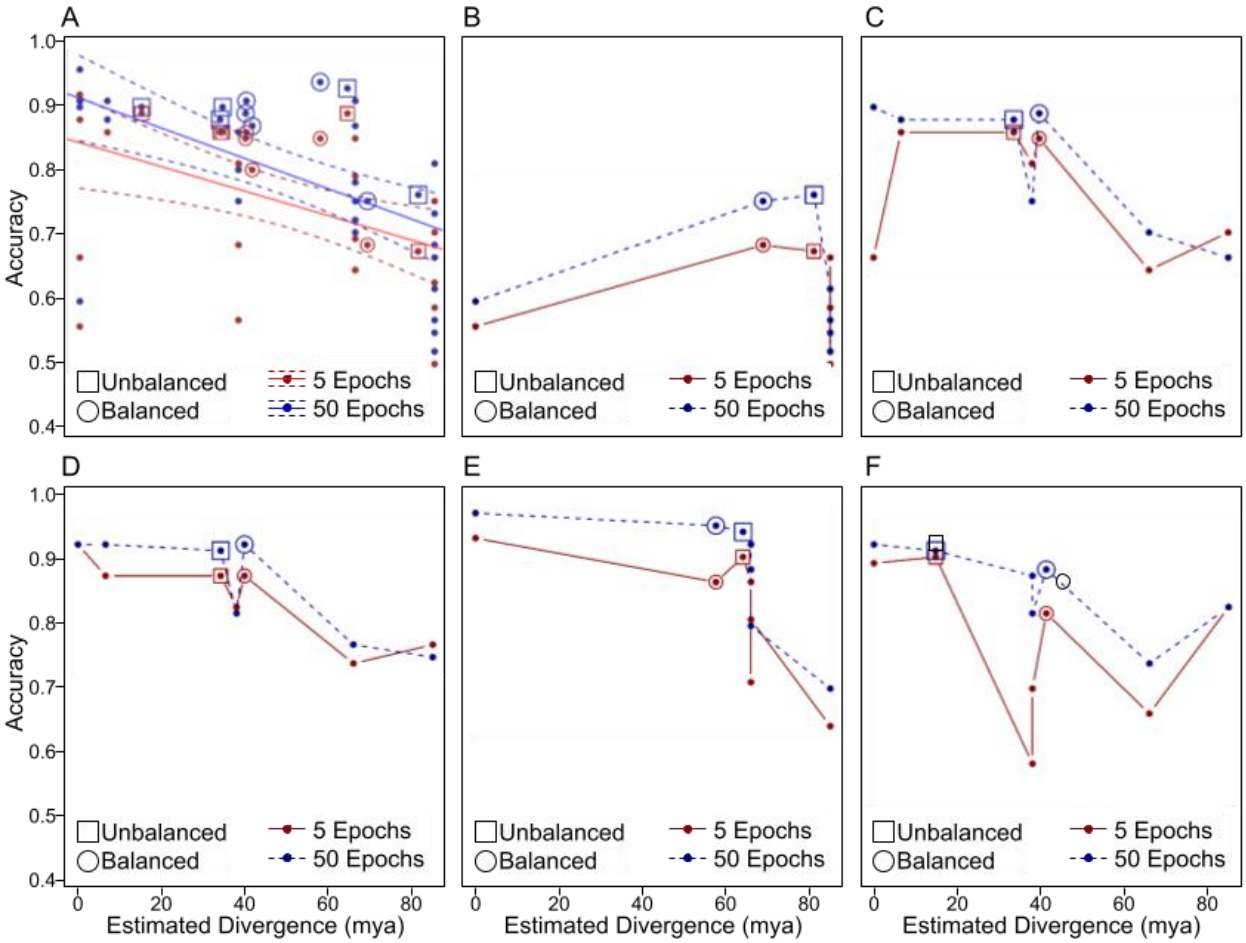

Figure 3: Model performance by species drops with estimated divergence, but training on multiple species performs better than expected given average divergence time. X-axis gives the estimated divergence time. Y-axis gives the accuracy. Lines are drawn to connect points in panels B through F. Points outlined with large squares were trained on the "5 Species" model. Points outlined with large circles were trained on the "5 Species Balanced" model. Shades of red indicate models run for 5-step checkpoints. Shades of blue indicate models run for 50-step checkpoints. A) all data, excluding *Melozone fusca* (see Figure 4). Colored lines show the line of best fit. B) *Calypte anna*. C) *Cardinalis cardinalis*. D) *Cardinalis sinuatus*. E) *Empidonax virescens*. F) *Zonotrichia leucophrys*.

Table 3: Accuracy of models trained in TweetyNet for *Melozone fusca*. Accuracy values are given for models checkpointed at 5-steps and 50-steps. Estimated divergence time for "5 Species" models is a weighted average, proportional to the sample sizes used for each species trained on.

Trained On	**Tested On**	**Accuracy 5-step checkpoints**	**Accuracy 50-step checkpoints**	**Estimated Divergence (Mya)**
Calypte anna	*Melozone fusca*	0.77	0.80	85.0
Cardinalis cardinalis	*Melozone fusca*	0.81	0.91	38.0
Cardinalis sinuatus	*Melozone fusca*	0.91	0.91	38.0
Empidonax virescens	*Melozone fusca*	0.86	0.91	66.0
Zonotrichia leucophrys	*Melozone fusca*	0.88	0.91	8.6
5 Species	*Melozone fusca*	0.91	0.94	20.8
5 Species Balanced	*Melozone fusca*	0.91	0.93	43.8
5 Species (5) + *Melozone fusca*	*Melozone fusca*	0.95	n/a	0.0
5 Species Balanced (5) + *Melozone fusca*	*Melozone fusca*	0.94	n/a	0.0
5 Species (50) + *Melozone fusca*	*Melozone fusca*	0.95	n/a	0.0
5 Species Balanced (50) + *Melozone fusca*	*Melozone fusca*	0.93	n/a	0.0

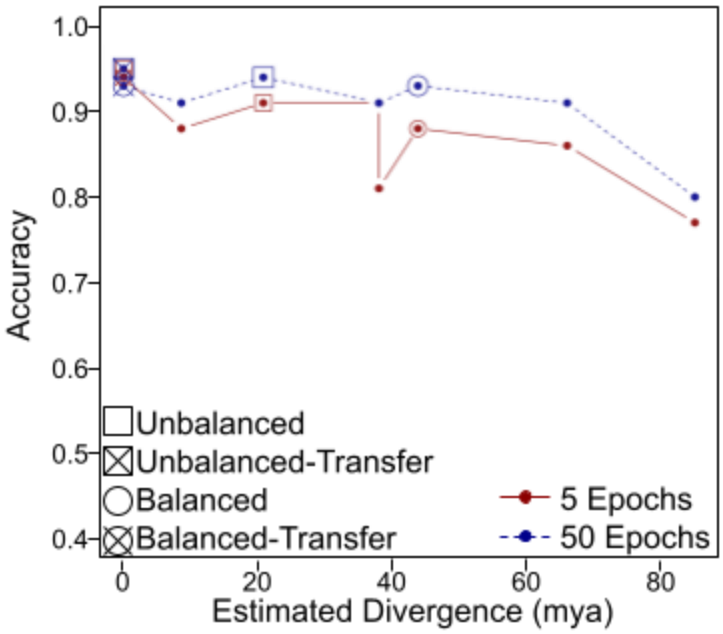

Figure 4: Model performance on *Melozone fusca* drops with estimated divergence, but also shows recovery when trained on multiple species. X-axis gives the estimated divergence time. Y-axis gives the accuracy. Lines are drawn to connect points. Points outlined with large squares were trained on a "5 Species" model. Points outlined with large circles were trained on a "5 Species Balanced" model. Large black squares and circles with cross-marks indicate transfer models. Shades of red indicate models run for 5-step checkpoints. Shades of blue indicate models run for 50-step checkpoints.

Zero-shot vs few-shot learning

With respect to the small *Melozone fusca* dataset, classification of this species' songs is overall highly accurate. For the single-species models, the least accurate was that of the distantly related *Calypte anna* (72–80%). For the other species, the 50-step checkpoint models all perform equivalently (91% accuracy) while in the 5-step checkpoint models, accuracy ranges (81–91% accuracy). For the multi-species models, the unbalanced model (91–94%) performed slightly

better than the balanced model (88–93%), and slightly better than the closest related species *Zonotrichia leucophrys* (88–91%), suggesting that having a model trained on a more diverse subset of sounds improves performance of zero-shot learning. Lastly, the few-shot learning transfer models perform the best of them all with little additional training time (93–95%), suggesting that having a single multi-purpose model rich in data to "seed" the training of species with few data may be beneficial.

Song diversity results

Song complexity values (from Medina & Francis, 2012) ranged such that *Cardinalis cardinalis* had the lowest complexity, followed by *Cardinalis sinuatus*, and then *Zonotrichia leucophrys* with the highest complexity. In contrast, our calculated hypervolume diversity metric finds that instead *Cardinalis sinuatus* has the smallest hypervolume, followed by *Empidonax virescens*, *Melozone fusca*, *Cardinalis cardinalis*, *Zonotrichia leucophrys*, and finally *Calypte anna* with the largest hypervolume. The latter two species occupy much higher percentages of hypervolumes compared to the remaining four species. With such a small number of points (N=3 for complexity from Medina & Francis, 2012; N=6 for hypervolume estimates) we do not have the power to detect any correlations between the two values, but we expect that complexity and diversity should be correlated.

Neither complexity (Medina & Francis, 2012) nor mean hypervolume of the species used for training are significant predictors of accuracy ($p>0.11$). Likewise, song complexity (Medina & Francis, 2012) of the species used for testing is not a significant predictor of accuracy ($p>0.45$). However, the mean hypervolume of the species used for testing is negatively

associated with accuracy, such that the more diverse a species' song is, the less accurate the model is at predicting it ($p<0.0002$, adjusted $R^2>0.28$).

Table 4: Species vary with respect to song complexity and hypervolume diversity. Complexity scores range from lower complexity (more negative) to higher complexity (more positive). Hypervolume diversity is a percentage of total hypervolume across all six species, with means and standard deviations from 50 independent runs of 50 random syllables per species.

Species	Complexity[1]	Hypervolume Mean ± SD
Calypte anna	n/a	0.5000 ± 0.1600
Zonotrichia leucophrys	-0.695	0.2400 ± 0.1100
Cardinalis cardinalis	-1.798	0.0140 ± 0.0100
Melozone fusca	n/a	0.0055 ± 0.0040
Empidonax virescens	n/a	0.0041 ± 0.0032
Cardinalis sinuatus	-1.086	0.0028 ± 0.0037

[1]from Medina and Francis 2012

Discussion:

Numerous initiatives have been undertaken over the past decades to accumulate and archive biological information into searchable and accessible databases. Such efforts include databases of morphological characteristics (MorphoNet; Leggio et al., 2019), localities for use in modeling abundance and niche (GBIF; gbif.org), genetic, genomic, and protein sequences (GenBank; Clark et al., 2016), and of course bioacoustics. Taking full advantage of all these datasets requires increases in efficiency, including development of automated retrieval of data as well as data extraction.

Bioacoustic data such as those we have used here have traditionally been parsed manually, since spectrograms were developed in the 1950s (Borror & Reese, 1953). These methods have required large amounts of time, both to extract the usable data but also to verify and validate the homologous aspects of songs for comparisons within and between taxa. Songs must be searched for the actual sounds, and then researchers must decide which parts of the sound are considered unique syllables, before even beginning to classify them by which type of syllable is being produced. Often this necessitates deep understanding of the species in question,

and while this approach is inherently valuable and has provided the foundation for avian bioacoustics, it may preclude more broad understanding of vocalizations outside of those focal taxa. Automated methods promise to improve the efficiency of bioacoustics by identifying data more readily and by automatically calculating differences between sounds (e.g., Rocha & Romano, 2021). Despite these improvements, the attributes and limitations of these methods are not as well understood in comparison to more traditional manual analyses. One aspect that has been investigated is the amount of time spent: processing songs in automated ways can be up to ~5 times faster compared to manual methods (Swiston et al., 2009).

In terms of performance, training a model on a species and applying it to that same species appears to be the most accurate method of extracting syllables, provided that the sample size is sufficiently high. This is irrespective of whether the training dataset includes only that species or also includes other taxa (e.g., the 5-species models). However, the range in accuracy varies highly across taxa. Species with more simple vocalizations (e.g., *Empidonax*) appear more likely to be classified with high accuracy than species with more complex vocalizations (e.g., *Zonotrichia*). We speculate this is due to a combination of factors. Taxa with more complex vocalizations must optimize the detection algorithm for multiple different kinds of syllables at once, which is expected to be more difficult than categorizing fewer kinds of syllables. Our models were run under the same parameters and for the same training times, though they were allowed to exit early if they reached an accuracy plateau; given this, it is likely that the convolutional neural network can solve a “simpler” problem in the same amount of time more efficiently. Another consideration, however, is sample size given to the model. Although we did not find an explicit correlation between sample size and accuracy, it is nevertheless possible that our high accuracy on these species is due to over-fitting, rather than any biological aspect of the

song. However, we do not attribute these differences to learning vs. non-learning behavior; although *Empidonax* are the only suboscine (non-learning) Passeriform bird in our dataset, there is not a relationship between song complexity and oscine/suboscine status in Passeriformes (Pearse et al., 2018).

We find that model accuracy drops with phylogenetic distance between the training data and tested data. This suggests that there is commonality between closely related species in terms of their song; indeed, evidence across passerines suggests that song is phylogenetically constrained, in addition to being influenced by morphology and ecology (Pearse et al., 2018; Mikula et al., 2021). From these results we would predict that any similar learning algorithm should behave in this fashion with traits that are associated with phylogeny. This pattern may also extend to species with convergent song features as well, though this remains to be tested.

Though the sample size does not appear to impact accuracy, it is an important consideration when trying to decide how to best extract data from species. With sufficiently large sample sizes (TweetyNet suggests 3 minutes of annotation, Cohen et al., bioRxiv), techniques like convolutional neural networks will work effectively to learn and parse the diversity of syllables and sounds present in these data. However, for some taxa the available data are not sufficient for these types of learning. In this case, we have two recommendations depending on the study system. If a study only concerns one single species without a large amount of data, it would likely be best to manually annotate all the data available, rather than try to fit a computationally expensive model to it. On the other hand, if a study is focused on multiple species at once, especially if some of those species have a large sample size, we recommend training a model on multiple different species, optionally then using transfer learning techniques to fine-tune the model. The caveat with this latter approach is that the benefit of this learning

may decrease as phylogenetic distance (and song dissimilarity) increases with respect to your focal taxa.

We find that transfer learning from multiple species models substantially increases accuracy for taxa without a large amount of data. However, this might also be partially due to overfitting of very small datasets. Transfer efficiency seems to be just as good when balanced vs unbalanced. We recommend balancing, however, because it appears to perform just as well with significantly less data and thus less training time. In terms of computational efficiency, we find that training a model with multiple different species takes longer than training a model on each species separately. We suspect this is because the algorithm needs to optimize segmentation of many disparate kinds of syllables, much like with species that have more complex songs compared to species with fewer songs. However, we also find that models trained on multiple species perform better than models trained on single species when applied to a never-before-seen taxon, even after considering average phylogenetic distance between training species and testing species. Despite the increase in computation time, having a more general-purpose, multi-species model trained can be worthwhile for researchers looking to maximize the number of taxa they can segment syllables for, while minimizing the amount of manual annotation and training time they need.

Downstream applications and implications

The approach we develop here is fast, flexible, and can be run on a standard laptop, though performance will benefit from parallel processing. As such, our application can supplement the deep avian bioacoustics literature with the breadth afforded by efficient data generation. In addition, our overall framework will likely be applicable to other machine learning

algorithms: TweetyNet could be replaced by other convolutional/recurrent neural network methods with these guidelines still being relevant. It remains to be seen whether other methods, including random forests or evolving neural networks, will perform similarly (Knapp et al., 2017; Juang & Chen, 2007; Ranjard & Ross, 2008; Neal et al., 2011).

Automatic methods can be used to validate and supplement manual segmentation analyses, as are traditional in avian bioacoustics. From the most basic, this approach can be used with sparser forms of data, for instance continuous recordings where the species of interest only vocalizes for a short amount of time. Segmentation with this method and downstream scripts we have developed here can reduce the amount of recording that needs to be searched by quickly eliminating large portions of the files. This segmentation method can also be used to standardize projects among individuals who are all working concurrently. Human biases in segmentation are a known phenomenon (Tachibana et al., 2014; Ranjard et al., 2015; Milinski, 1997), but applying a single model to begin from would potentially alleviate some of those issues. Further, when working from automatically segmented data, the ability to recognize incorrect or otherwise anomalous syllables where models and manual labeling do not match would make data cleaning swifter and more reliable. Lastly, the use of these models can help assist in training younger scientists without potentially compromising data quality; with a standard template to work from and the ability for the machine learning algorithm to partially check classifications, students could spend more time developing other research skills than simple song annotation.

Automatic segmentation can be useful not just to reduce the workload of people, but also to set up for downstream applications. First, though not implemented here, TweetyNet can not only identify the presence of syllables, but also individual types of syllables; indeed, this is TweetyNet's original goal, and it does so with high accuracy. However, this requires knowledge

of what types of syllables exist in the dataset, and in taxa that have not been studied before, it is hard to know *a priori* what those sounds are. However, an algorithm that can broadly detect sounds can thus lead into automatic classification of syllables via clustering analyses, which exist already in easily implemented fashions (e.g., K-means clustering, Lloyd, 1982). It is beyond the scope of this manuscript to discuss clustering methods but being able to automatically characterize the repertoire of a new species is within the realm of possibility. Further, any methods that can segment the syllables as well as classify types of syllables allow for the fine detection variation within those syllable classes, as well as changes in the syntax of bird song (i.e., the arrangement of sounds into specific sequences). This contrasts with changes in other features of song, like frequency, duration, or bandwidth, which are independent of the order of specific syllables. Regardless, the ability to categorize these sounds will be particularly useful in getting a more thorough assessment of song complexity within and between taxa.

Segmentation of sound is a critical part of analysis of song in birds, as well as other taxa that communicate with sound including some insects, frogs, and bats. Given the sheer number of species that this encompasses, having a standardized and automatic approach will allow for fast performance of traditional bioacoustic analysis. Our method forms a starting point for this work. Given enough time, we foresee that these and other models can be further refined and used to segment all recordings available for all species, which would likely comprise multiple lifetimes of work if done manually and be intractable with the current speed at which new recordings are added. Further, it is not outlandish to think that as machine learning methods become more sophisticated, software could be developed to auto-segment sounds as they are cataloged into these databases and repositories, or even to be integrated into the recorders and microphones that are commonly used today.

Conclusions

In this paper, we investigate the influence of phylogeny on machine learning methods to extract syllables from recordings of bird song, finding that the accuracy of these methods depends in part on the phylogenetic relatedness of the taxa being trained on and tested on. However, the loss in accuracy associated with distantly related taxa can be ameliorated by including multiple species. Further, species with smaller sample sizes can be classified with high accuracy by tuning previously trained models. We suggest the following best practices for using machine learning in bioacoustics, depending on the sample sizes of species of interest: for very small amounts of total data, it is best to hand-annotate syllables to segment them irrespective of the number of species. With larger amounts of data, hand-annotation of select species that span the phylogenetic (and bioacoustic) variation in the dataset, then training a machine learning model to segment syllables from it will reduce computational load without a major disruption to accuracy. This framework will be broadly applicable to many regions of the world and many taxa, allowing us to achieve a global perspective on vocalizations. Ultimately, this innovation will change how biologists analyze songs by expanding the scope of what is possible from depth to breadth.

Author Contributions:

KLP and BCC conceived the study; KLP and JY collected the data; KLP performed the analyses, wrote custom code, and wrote the initial manuscript. All authors edited the manuscript.

Acknowledgements:

We thank all the individuals throughout the years who have contributed to the bioacoustics repositories, including the Borror Laboratory of Bioacoustics and Xeno-Canto. For helpful feedback on these analyses, we thank D Kelly, T Berger-Wolf, G Carter, R Adams, and S Gaunt. Thoughtful reviews come from T Yuri, S Decker, F Mol Lanna, D Parsons, J Wieringa, D Duckett, S Moshier, EM Da Fonesca, and MTC Thome. KLP and BCC were supported by NSF DEB-2016189.

Conflict of Interest Statement:

The authors declare no conflicts of interest.

Data Availability:

Data used will be archived on Dryad. Bioacoustic data are publicly available from Xeno-Canto and the Borror Laboratory of Bioacoustics. Code is available on GitHub at github.com/kaiyaprovost/bioacoustics.

References:

Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the landscape of recombination using deep learning. *Molecular Biology and Evolution, 37(6)*, 1790-1808.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Brian C. Van Esesn, B. C, Awwal, A.S.A., & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164.

Alonso, J. B., Cabrera, J., Shyamnani, R., Travieso, C. M., Bolaños, F., García, A., Villegas, A., & Wainwright, M. (2017). Automatic anuran identification using noise removal and audio activity detection. *Expert Systems with Applications, 72*, 83-92.

Araya-Salas, M., & Smith-Vidaurre, G. (2017). warbleR: an R package to streamline analysis of animal acoustic signals. *Methods in Ecology and Evolution, 8(2)*, 184-191.

Borror, D. J., & Reese, C. R. (1953). The analysis of bird songs by means of a vibralyzer. *The Wilson Bulletin*, 271-276.

Boža, V., Brejová, B., & Vinař, T. (2017). DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE, 12(6)*, e0178751.

Brambilla, M., Janni, O., Guidali, F., & Sorace, A. (2008). Song perception among incipient species as a mechanism for reproductive isolation. *Journal of Evolutionary Biology, 21(3)*, 651-657.

Burdfield-Steel, E. R., & Shuker, D. M. (2018). Divergence is not speciation, or why we need females: a comment on Tinghitella et al. *Behavioral Ecology, 29(4)*, 801-801.

Catchpole, C. K., & Slater, P. J. (1995). Bird song: biological themes and variations. Cambridge University Press.

Chambers, J. M., Freeny, A. E., Heiberger, R. M. (1992). Analysis of variance. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical Models in S* (pp. 145-194). Wadsworth and Brooks/Cole.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research, 44(D1)*, D67-D72.

Cohen, Y., Nicholson, D. A., & Gardner, T. J. (2020). TweetyNet: A neural network that enables high-throughput, automated annotation of birdsong. *bioRxiv*.

de Kort, S. R., Eldermire, E. R., Cramer, E. R., & Vehrencamp, S. L. (2009). The deterrent effect of bird song in territory defense. *Behavioral Ecology, 20(1)*, 200-206.

Derryberry, E. P. (2011). Male response to historical and geographical variation in bird song. *Biology Letters, 7(1)*, 57-59.

Devine, J., Aponte, J. D., Katz, D. C., Liu, W., Vercio, L. D. L., Forkert, N. D., Marcucio, R., Percival, C. J., & Hallgrímsson, B. (2020). A registration and deep learning approach to automated landmark detection for geometric morphometrics. *Evolutionary Biology, 47(3)*, 246-259.

Dingle, C., Poelstra, J. W., Halfwerk, W., Brinkhuizen, D. M., & Slabbekoorn, H. (2010). Asymmetric response patterns to subspecies-specific song differences in allopatry and parapatry in the gray-breasted wood-wren. *Evolution: International Journal of Organic Evolution, 64(12)*, 3537-3548.

Du, P., & Troyer, T. W. (2006). A segmentation algorithm for zebra finch song at the note level. *Neurocomputing, 69(10-12)*, 1375-1379.

Edgington, D. R., Cline, D. E., Davis, D., Kerkez, I., & Mariette, J. (2006). Detecting, tracking and classifying animals in underwater video. *IEEE OCEANS 2006*, 1-5.

Freeman, B. G., Rolland, J., Montgomery, G. A., & Schluter, D. (2022). Faster evolution of a premating reproductive barrier is not associated with faster speciation rates in New World passerine birds. *Proceedings of the Royal Society B, 289(1966)*, 20211514.

Gehara, M., Mazzochinni, G. G., & Burbrink, F. (2020). PipeMaster: inferring population divergence and demographic history with approximate Bayesian computation and supervised machine-learning in R. *bioRxiv*.

Gill, F. B., & Lanyon, W. E. (1964). Experiments on species discrimination in Blue-winged Warblers. *The Auk, 81(1)*, 53-64.

Greenspan, H., Ruf, A., & Goldberger, J. (2006). Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Transactions on Medical Imaging, 25(9)*, 1233-1245.

Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., & Ferres, J. L. (2021). Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific Reports, 11(1)*, 1-12.

Halgaswaththa, T., Atukorale, A. S., Jayawardena, M., & Weerasena, J. (2012). Neural network based phylogenetic analysis. *IEEE 2012 International Conference on Biomedical Engineering*, 155-160.

Heileman, G., Myler, H., & Georgiopoulos, M. (1990). An object-oriented approach to the simulation of artificial neural networks. In G. W. Zobrist, J. V. Leonard (Eds.), *Progress in Simulation* (pp. 126). Ablex Publishing Corp.

Hemanth, D. J., & Estrela, V. V. (2017). Deep learning for image processing applications. IOS Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79(8)*, 2554-2558.

Jenkins, P. F. (1978). Cultural transmission of song patterns and dialect development in a free-living bird population. Animal Behaviour, 26, 50-78.

Jiang, Y., Han, J., Zhang, Z., Chen, X., & Yang, C. (2021). Parent-offspring and inter-offspring responses to conspecific vs heterospecific distress calls in two sympatric birds. *Current Zoology*, zoab103.

Juang, C. F., & Chen, T. M. (2007). Birdsong recognition using prediction-based recurrent neural fuzzy networks. *Neurocomputing, 71(1-3)*, 121-130.

Kamath, U., Liu, J., & Whitaker, J. (2019). Deep learning for NLP and speech recognition. Springer.

Keen, S. C., Cole, E. F., Sheehan, M. J., & Sheldon, B. C. (2020). Social learning of acoustic anti-predator cues occurs between wild bird species. P*roceedings of the Royal Society B, 287(1920)*, 20192513.

Knapp, J., Qu, G., & Zhang, F. (2016). Automatic Species Recognition Based on Improved Birdsong Analysis. *2016 15th IEEE International Conference on Machine Learning and Applications*, 176-181.

Kondo, N., & Watanabe, S. (2009). Contact calls: information and social function. *Japanese Psychological Research, 51(3)*, 197-208.

Koumura, T., & Okanoya, K. (2016). Automatic recognition of element classes and boundaries in the birdsong with variable sequences. *PLOS ONE, 11(7)*, e0159188.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution, 34(7)*, 1812-1819.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521(7553)*, 436-444.

Leggio, B., Laussu, J., Carlier, A., Godin, C., Lemaire, P., & Faure, E. (2019). MorphoNet: an interactive online morphological browser to explore complex multi-scale data. *Nature Communications, 10(1)*, 1-8.

Ligges, U., Preusser, A., Thieler, A., Mielke, J., & Weihs, C. (2018). Package 'tuneR'.

Lloyd, S. (1982). Least squares quantization in PCM. I*EEE transactions on information theory, 28(2)*, 129-137.

Lürig, M. D., Donoughe, S., Svensson, E. I., Porto, A., & Tsuboi, M. (2021). Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. *Frontiers in Ecology and Evolution, 9*, 148.

Macdougall-Shackleton, S. A. (1997). Sexual selection and the evolution of song repertoires. In V. Nolan, E. D. Ketterson, C. F. Thompson (Eds.), *Current Ornithology* (pp. 81-124). Springer.

Magrath, R. D., Haff, T. M., Fallow, P. M., & Radford, A. N. (2015). Eavesdropping on heterospecific alarm calls: from mechanisms to consequences. *Biological Reviews, 90(2)*, 560-586.

Mason, N. A., Burns, K. J., Tobias, J. A., Claramunt, S., Seddon, N., & Derryberry, E. P. (2017). Song evolution, speciation, and vocal learning in passerine birds. *Evolution, 71(3)*, 786-796.

Medina, I., & Francis, C. D. (2012). Environmental variability and acoustic signals: a multi-level approach in songbirds. *Biology Letters, 8(6)*, 928-931.

Mikula, P., Valcu, M., Brumm, H., Bulla, M., Forstmeier, W., Petrusková, T., Kempenaers, B., & Albrecht, T. (2021). A global analysis of song frequency in passerines provides no support for the acoustic adaptation hypothesis but suggests a role for sexual selection. *Ecology Letters, 24(3)*, 477-486.

Milinski, M. (1997) How To Avoid Seven Deadly Sins in the Study of Behavior. In P. J. B. Slater, J. S. Rosenblatt, C. T. Snowdon, M. Millinski (Eds.), *Advances in the Study of Behavior* (pp. 159–180). Academic Press.

Miller, R. G. 1981. Simultaneous Statistical Inference. Springer.

Neal, L., Briggs, F., Raich, R., & Fern, X. Z. (2011). Time-frequency segmentation of bird song in noisy acoustic environments. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012-2015.

Nicholson, D. (2016). Comparison of machine learning methods applied to birdsong element classification. *Proceedings of the 15th Python in Science Conference*, 57–61.

Nowicki, S., & Searcy, W. A. (2005). Song and mate choice in birds: how the development of behavior helps us understand function. *The Auk, 122(1)*, 1-14.

O'Loghlen, A. L., & Rothstein, S. I. (1995). Culturally correct song dialects are correlated with male age and female song preferences in wild populations of brown-headed cowbirds. *Behavioral Ecology and Sociobiology, 36(4)*, 251-259.

O'Mahony, G. D., Harris, P. J., & Murphy, C. C. (2020). Detecting interference in wireless sensor network received samples: A machine learning approach. *2020 IEEE 6th World Forum on Internet of Things*, 1-6.

Pearse, W. D., Morales-Castilla, I., James, L. S., Farrell, M., Boivin, F., & Davies, T. J. (2018). Global macroevolution and macroecology of passerine song. *Evolution, 72(4)*, 944-960.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190(3-4)*, 231-259.

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography, 40(7)*, 887-893.

Provost, K. L., Mauck III, W. M., & Smith, B. T. (2018). Genomic divergence in allopatric Northern Cardinals of the North American warm deserts is linked to behavioral differentiation. *Ecology and Evolution, 8(24)*, 12456-12478.

Provost, K. L., Myers, E. A., & Smith, B. T. (2021). Community phylogeographic patterns reveal how a barrier filters and structures taxa in North American warm deserts. *Journal of Biogeography, 48(6)*, 1267-1283.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics, 32(6)*, 859-866.

R Core Team. (2020). R: a language and environment for statistical computing, version 4.0.3. R Foundation for Statistical Computing.

Ranjard, L., & Ross, H. A. (2008). Unsupervised bird song syllable classification using evolving neural networks. *The Journal of the Acoustical Society of America, 123(6)*, 4358-4368.

Ranjard, L., Withers, S. J., Brunton, D. H., Ross, H. A., & Parsons, S. (2015). Integration over song classification replicates: Song variant analysis in the hihi. *The Journal of the Acoustical Society of America, 137(5)*, 2542-2551.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing, 3(1)*, 72-83.

Rice, W. R. (1982). Acoustical location of prey by the marsh hawk: adaptation to concealed prey. *The Auk, 99(3)*, 403-413.

Riebel, K., Odom, K. J., Langmore, N. E., & Hall, M. L. (2019). New insights from female bird song: towards an integrated approach to studying male and female communication roles. *Biology Letters, 15(4)*, 20190059.

Ripley, B. D. (1996). Pattern recognition and neural networks. Cambridge University Press.

Rocha, P. C., & Romano, P. S. (2021). The shape of sound: A new R package that crosses the bridge between Bioacoustics and Geometric Morphometrics. *Methods in Ecology and Evolution, 12(6)*, 1115-1121.

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics, 34(4)*, 301-312.

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLOS Computational Biology, 12(3)*, e1004845.

Sprengel, E., Jaggi, M., Kilcher, Y., & Hofmann, T. (2016). Audio based bird species identification using deep learning techniques. *CLEF*, 547-559.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation, 142(10)*, 2282-2292.

Swiston, K. A., & Mennill, D. J. (2009). Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology, 80(1)*, 42-50.

Tachibana, R. O., Oosugi, N., & Okanoya, K. (2014). Semi-automatic classification of birdsong elements using a linear support vector machine. *PLOS ONE, 9(3)*, e92584.

Tinghitella, R. M., Lackey, A. C., Martin, M., Dijkstra, P. D., Drury, J. P., Heathcote, R., Keagy, J., Scordato, E. S. C., & Tyers, A. M. (2018). On the role of male competition in speciation: a review and research agenda. *Behavioral Ecology, 29(4)*, 783-797.

Uy, J. A. C., Irwin, D. E., & Webster, M. S. (2018). Behavioral isolation and incipient speciation in birds. *Annual Review of Ecology, Evolution, and Systematics, 49*, 1-24.

Valipour, M., & Wang, Y. (2017). Building semantic hierarchies of formal concepts by deep cognitive machine learning. *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing*, 51-58.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys, 53(3)*, 1-34.

Wang, Y., He, K., Hopcroft, J. E., & Sun, Y. (2016). Deep Neural Networks and the Tree of Life. 2017 International Conference on Learning Representations.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3(1)*, 1-40.

West, M. J., King, A. P., & Eastzer, D. H. (1981). Validating the female bioassay of cowbird song: relating differences in song potency to mating success. *Animal Behaviour, 29(2)*, 490-501.

Wright, J., & Leonard, M. L. (Eds.). (2007). The evolution of begging: competition, cooperation and communication. Springer.

Yandell, B. S. 1997. Practical data analysis for designed experiments. Chapman & Hall/CRC.

Zell, A., Mache, N., Hübner, R., Mamier, G., Vogt, M., Schmalzl, M., & Herrmann, K. U. (1994). SNNS (stuttgart neural network simulator). In J. Skryzpek (Eds.), Neural network simulation environments (pp. 165-186). Springer.

Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., & Brewer, A. (2020). Beluga whale acoustic signal classification using deep learning neural network models. *The Journal of the Acoustical Society of America, 147(3)*, 1834-1841.