Bioacoustic Event Detection with Self-Supervised Contrastive Learning

Peter C. Bermant[1,*], **Leandra Brickson**[1], **and Alexander J. Titus**[1,2]

[1]Colossal Biosciences, Austin, TX, 78701, USA
[2]Form Bio, Austin, TX, 78701, USA
[*]publications@colossal.com

ABSTRACT

While deep learning has revolutionized ecological data analysis, existing strategies often rely on supervised learning, which is subject to limitations on real-world applicability. In this paper, we apply self-supervised deep learning methods to bioacoustic data to enable unsupervised detection of bioacoustic event boundaries. We propose a convolutional deep neural network that operates on the raw waveform directly and is trained in accordance with the Noise Contrastive Estimation principle, which enables the system to detect spectral changes in the input acoustic stream. The model learns a representation of the input audio sampled at low frequency that encodes information regarding dissimilarity between sequential acoustic windows. During inference, we use a peak finding algorithm to search for regions of high dissimilarity in order to identify temporal boundaries of bioacoustic events. We report results using these techniques to detect sperm whale (*Physeter macrocephalus*) coda clicks in real-world recordings, and we demonstrate the viability of analyzing the vocalizations of other species (e.g. Bengalese finch syllable segmentation) in addition to other data modalities (e.g. animal behavioral dynamics, embryo development and tracking). We find that the self-supervised deep representation learning-based technique outperforms established threshold-based baseline methods without requiring manual annotation of acoustic datasets. Quantitatively, our approach yields a maximal R-value and F1-score of 0.887 and 0.876, respectively, and an area under the Precision-Recall curve (PR-AUC) of 0.917, while a baseline threshold detector acting on signal energy amplitude returns a maximal R-value and F1-score of 0.620 and 0.576, respectively, and a PR-AUC of 0.571. We also compare with a threshold detector using preprocessed (e.g. denoised) acoustic input. The findings of this paper establish the validity of unsupervised bioacoustic event detection using deep neural networks and self-supervised contrastive learning as an effective alternative to conventional techniques that leverage supervised methods for signal presence indication. Providing a means for highly accurate unsupervised detection, this paper serves as an important step towards developing a fully automated system for real-time acoustic monitoring of bioacoustic signals in real-world acoustic data. All code and data used in this study are available online.

1 Introduction

As conservation and management strategies encompass progressively more extreme solutions, ranging from species de-extinction to translation of non-human communication[1], novel computational techniques can provide improved methods for processing large quantities of multimodal ecological data. In recent years, advances in machine learning (ML) and deep learning (DL), in particular, have revolutionized the analysis of bioacoustic data, facilitating the development of automated pipelines for processing data and contributing to enhanced conservation tactics across diverse taxa. To date, applications of DL to bioacoustics tend to rely on heavily supervised methods demanding large manually labeled datasets[2], and these approaches often treat detection as a binary classification task[3] that is more reflective of presence indication than event detection[2]. In this paper, we propose a self-supervised representation learning-based approach to bioacoustic event detection and segmentation that can predict temporal onsets and offsets of bioacoustic signals in a fully unsupervised regime.

Recent developments in observation hardware and methods now offer ecologists and biologists immense quantities of data captured via high-end cameras, static microphones, biologging devices, and satellites, among others[2]. The unprecedented amounts of high-quality data supersede the conventional approach of manual annotation and demand the application of novel computational methods to discover patterns of animal behavior with important ecological implications. To this end, contemporary methods of data analysis–and bioacoustic data analysis, in particular–often leverage DL and deep neural networks (DNNs).

In general, implementations of ML to ecological data analysis heavily depend on supervised techniques in which sufficiently large training datasets have been curated with accurate information, often corresponding to labels such as presence, location, or other identifying features of a species, object, sound, etc. For example, early research studies applied DL to sperm whale

bioacoustics[4], using supervised learning to train DNNs to perform detection and classification tasks of input spectrogram feature representations, requiring manually labeled datasets annotated according to relevant targets including but not limited to signal presence, clan membership, individual identity, and call type. Large-scale studies often employ Convolutional Neural Network (CNN) architectures to carry out supervised detection or classification tasks across diverse species. These include detection of humpback whale vocalizations in passive acoustic monitoring datasets[5]; detection, classification, and censusing of blue whale sounds[6]; avian species monitoring based on a CNN trained to predict the species class label given input audio[7]; presence indication of Hainanese gibbons using a ResNet-based CNN[8]; and, recently, detection and classification of marine sound sources using an image-based approach to spectrogram classification[9]. However, such supervised learning approaches remain limited in their scope, hindering their capacity to be deployed in real-time data processing pipelines. A salient limitation remains the dependency on large amounts of high-quality manually annotated datasets[2], a problem compounded by its reliance on trained experts, the laborious nature (i.e. extensive time and resources) involved in manual labeling, and the potential for bias, variability, or uncertainty during the labeling process associated with variable human confidence and perception, especially across cohorts of individual annotators[10,11]. Further, manual data annotation remains expensive in terms of time[2], expertise, and financial cost[12], which has motivated citizen science-based crowdsourcing efforts to annotate large datasets at scale, though these approaches are similarly accompanied by error and limitations[13,14]. Reliance on manual annotations represents a key constraint on supervised DL methods for bioacoustic data analysis, making unsupervised approaches an attractive alternative with the potential to provide greater insight into animal communication and behavior.

More fundamentally, however, the foundational principles upon which DNN classifier-based detectors are constructed limits their applicability to real-world ecological data analysis. DNN-based bioacoustic detectors often pose the detection problem as a binary classification problem[5] in which a neural network learns discriminative feature representations of input acoustic features (e.g. raw waveforms, handcrafted acoustic parameters, spectrograms, miscellaneous time-frequency representations), enabling the model to predict a binary class label denoting signal or non-signal (i.e. background). While attempts to address this shortcoming exist, often by employing smaller window (i.e. acoustic segment) sizes to artificially improve temporal resolution[15], these approaches continue to treat detection as a presence indication task, unable to localize the signal directly within the window. Further, this workaround method of smaller windows is accompanied by its own disadvantages, mainly an impaired computational efficiency due to the need for overlapping windows to account for signals that could occur at the transition boundary between adjacent spectrograms[2]. The inability to precisely locate signal in a continuous recording renders it challenging to address downstream questions of ecological and/or animal behavioral importance. This is especially true in unsupervised regimes when labels (individual identity, call type, etc.) may not be readily accessible. For instance, unsupervised clustering-based techniques for call-type classification[16] often expect manually segmented calls, which remain a challenge to extract from DNN-based presence indicators.

As an alternative approach, we treat the detection problem as a segmentation problem in which we aim to predict temporal onsets and/or offsets of bioacoustic event signals given a continuous stream of acoustic data. While previous studies have attempted both supervised[17–19] and unsupervised[20,21] bioacoustic sound event detection (SED) and segmentation, modern DNN-based–and self-supervised–methods remain underexplored[3]. We take inspiration from recent advances in representation learning and self-supervised contrastive learning[22–25], especially those applied to human phoneme boundary detection and segmentation[26,27]. We propose a fully unsupervised DNN-based system for bioacoustic signal event detection that leverages self-supervised contrastive learning to map input audio to a learnable feature representation that encodes information regarding spectral boundaries, transitions, and changes that distinguish non-signal background from signal events.

Finally, we adjust our methods to account for unique challenges incurred by processing real-world bioacoustic data recorded in the native environment. In particular, a key challenge remains the presence of significant background noise in real-world bioacoustic recordings and the associated difficulties DNNs have in processing noisy data–often requiring modifications to DL-based methods to accommodate non-zero noise[28]–and distinguishing signal from non-signal[29], especially in unsupervised schemes. As in previous work[30], we account for environmental noise that would otherwise interfere with the contrastive learning objective by integrating on-the-fly noise reduction layers into the model architecture. In this paper, we provide an unsupervised framework for bioacoustic event detection based on contrastive representation learning, and we design our methods to operate directly on real-world acoustic data.

2 Materials and Methods

Motivated by recent advances in acoustic and visual representation learning involving self-supervised contrastive training objectives, we apply representation learning techniques to bioacoustic data to enable unsupervised detection and segmentation of bioacoustic events. In particular, given a continuous stream of unlabeled acoustic data, we implement a self-supervised learning algorithm that exploits an auxiliary proxy task with pseudo labels inferred from unlabeled data[25] and aims to train a model, f, to encode the raw input waveform to a spectral representation that emphasizes spectral boundaries in the signal. The method relies on the Noise Contrastive Estimation principle[31], which involves optimizing a probabalistic contrastive loss

function that allows the model, f, to distinguish between sequential acoustic windows and randomly sampled distractor windows. For the model, f, we make use of a deep neural network to discover hidden patterns and complex relationships[32] and to extract features relevant for optimizing the contrastive learning objective. During inference, a peak-finding algorithm detects regions of high dissimilarity between the encoded features of adjacent acoustic windows, which correspond to temporal boundaries of bioacoustic events.

2.1 Dataset

We apply our methods to a sperm whale (*Physeter macrocephalus*) click dataset. We process the raw acoustic data using the 'Best Of' cuts from the William A. Watkins Collection of Marine Mammal Sound Recordings database from Woods Hole Oceanographic Institution (https://cis.whoi.edu/science/B/whalesounds/index.cfm), retrieving 71 wav files amounting to 1.5 hours of acoustic data. Of these, we manually label with high confidence the clicks present in 38 files using Raven Pro 1.6.1, yielding a dataset containing 1738 annotated transients. All audio files were resampled to 48 kHz.

2.2 Self-Supervised Feature Extraction

Operating on the raw acoustic waveform directly, the model maps the input audio to a representation encoding discriminative features between sequential acoustic windows. Feature extraction involves training a neural network, $f : \mathbb{R}^N \mapsto \mathbb{R}^{T\times F}$, that encodes the N-sample acoustic waveform to a T-element temporal sequence of F-dimensional spectral features. While the model can process variable-length input waveforms, during training, the model f is given a real-valued continuous vector with fixed length N, i.e. $\mathbf{x} \in \mathbb{R}^N$, that represents the input audio signal and outputs an encoded spectral representation $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_T) \in \mathbb{R}^{T\times F}$ consisting of a sequence of feature vectors sampled at low frequency. To train f, we follow the paradigm of Kreuk et al., 2020[26]. We optimize the encoding function f in accordance with the Noise Contrastive Estimation principle such that the network learns to maximize the similarity between adjacent (i.e. sequential) windows $\mathbf{z}_i, \mathbf{z}_{i+1}$ for $i \in [1, T]$ in the learned representation while minimizing the similarity between randomly sampled distractor windows $\mathbf{z}_i, \mathbf{z}_j$ for $\mathbf{z}_j \in D(\mathbf{z}_i)$, the set of nonadjacent windows to $\mathbf{z}_i$, i.e. $D(\mathbf{z}_i) = \{\mathbf{z}_j : |i-j| > 1, \mathbf{z}_j \in \mathbf{z}\}$.

Following Kreuk et al., 2020[26], we use the cosine similarity between two feature vectors as our similarity metric

$$sim(\mathbf{z}_1, \mathbf{z}_2) = \frac{\mathbf{z}_1^T \mathbf{z}_2}{\|\mathbf{z}_1\| \|\mathbf{z}_2\|}$$

For the contrastive loss function, we implement a cross-entropy loss

$$\hat{\mathscr{L}}(\mathbf{z}_i, D_k(\mathbf{z}_i)) = -\log \frac{\exp^{sim(\mathbf{z}_i, \mathbf{z}_{i+1})}}{\sum_{\mathbf{z}_j \in \{\mathbf{z}_{i+1}\} \cup \{D_k(\mathbf{z}_i)\}} \exp^{sim(\mathbf{z}_i, \mathbf{z}_j)}}$$

This means that given a batch of n training samples $S = \{\mathbf{x}_i\}_{i=1}^n$, the total loss is given by[26]

$$\mathscr{L} = \sum_{\mathbf{x} \in S} \sum_{\mathbf{z} \in f(\mathbf{x})} \hat{\mathscr{L}}(\mathbf{z}_i, D_k(\mathbf{z}_i))$$

Given an acoustic waveform as input, this training criterion seeks to yield a representation in which the dissimilarity between sequential encoded feature vectors of (i.e. adjacent) audio windows is minimized.

2.3 The Model

We use a Convolutional Neural Network (CNN) comprised of one-dimensional convolutional layers (Conv1d) that operate on the raw waveform directly. Importantly, we also include on-the-fly data preprocessing layers to remove acoustic interferences.

In particular, the first layer of the model f consists of a high-pass filter layer constructed using a nontrainable Conv1d layer with frozen weights determined by a windowed *sinc* function[30,33,34]. Given the broadband nature of sperm whale clicks[35], we select a cutoff frequency of 3 kHZ and a transition bandwidth 0.08 to remove low-frequency environmental background noise while preserving the spectral structure of the clicks.

Following the denoising layer is the convolutional model, consisting of Conv1d layers with batch normalization (BatchNorm1d) and LeakyReLU activation. We use a 4-layer network with kernel sizes [8, 6, 4, 4] and strides [4, 3, 2, 2], corresponding to a hop length of 48 samples (i.e. 1ms) and a receptive field of 136 samples (i.e. $\sim$3ms). For all Conv1d layers, we use 128 output channels. Lastly, we include a fully-connected linear projection layer with Tanh activation to output the spectral representation. With this architecture, the model f encodes input audio to a learned representation comprised of a sequence of 32-dimensional feature vectors sampled at 1 kHz.

To train the model, we optimize the contrastive learning objective using Stochastic Gradient Descent (SGD). We use a learning rate 1e-4, momentum 0.9, batch size 64 for 50 epochs. Given the stochasticity of training, we serialize the model weights after each epoch and select the top-performing (in terms of the learning objective) model for inference.

Model	PR-AUC	F1-Score	R-Value
SSL Detection	0.917	0.876	0.887
Energy Amplitude Detection	0.571	0.576	0.620
Energy Amplitude + HPF Detection	0.706	0.639	0.643

Table 1 – Summary of experiments implemented in this study. We compare our approach based on self-supervised learning (SSL) with baseline energy amplitude threshold detectors with and without high pass filter (HPF) preprocessing. We report PR-AUC, maximal F1-Score, and maximal R-value.

The baseline energy amplitude detector model (which is functionally equivalent to a signal-to-noise (SNR)-based threshold detector[36]) requires no trainable parameters and involves using a peak-finding algorithm to detect peaks in the energy of the input waveform. We use the raw audio as input. We repeat the baseline detector with and without data preprocessing (e.g. high pass filter denoising). In this case, all suprathreshold detections are considered to be clicks.

2.4 Inference and Peak-Finding

During inference, we obtain the spectral representation $\mathbf{z}$ given the acoustic waveform $\mathbf{x}$ using the trained model (i.e. $\mathbf{z} = f(\mathbf{x})$). We compute the distance between sequential windows using a bounded L2 dissimilarity metric:

$$d(\mathbf{z}_1, \mathbf{z}_2) = \sqrt{(\mathbf{z}_1 - \mathbf{z}_2)^2}$$

$$dissim(\mathbf{z}_1, \mathbf{z}_2) = \tanh(\alpha \cdot d(\mathbf{z}_1, \mathbf{z}_2) + \beta)$$

where α and β are fixed hyperparameters obtained by exhaustive grid search. Unlike the cosine similarity metric, which neglects normalization, the Euclidean distance-based metric preserves information regarding scaling, which is an important consideration for signals in the presence of low-amplitude nonstationary noise. We evaluate the dissimilarity between the feature representations of adjacent windows $dissim(\mathbf{z}_i, \mathbf{z}_{i+1})$ for $i \in [1, T-1]$. As the contrastive learning objective aims to suppress dissimilarity between successive windows, a relatively high dissimilarity metric indicates a spectral boundary, suggesting a transition from background to bioacoustic signal (or vice versa) in the input audio. In the case of sperm whale clicks, we smooth the dissimilarity metric over the temporal axis by convolving *dissim* with a box function of width σ, operating under the assumption that we are interested in single click times as opposed to boundaries corresponding to onsets and offsets of the transient clicks.

Finally, using the `scikit-learn` package, we employ a peak-finding algorithm to search for peaks in the dissimilarity. We search over peak prominences δ and attribute all suprathreshold peak detections to sperm whale clicks.

2.5 Evaluation Metrics

Motivated by work on phoneme boundary detection[26], we evaluate model performance using precision (P), recall (R) and F1-score with a tolerance level τ. As in Kreuk et al., 2020[26], we also include the R-value, which serves as a more robust metric than F1-score that is less sensitive to oversegmentation. In the case of sperm whale clicks we choose $\tau = 20$ ms, a value that exceeds observed inter-pulse intervals (IPIs)[37] but is less than typical inter-click intervals (ICIs)[38], even in the case of buzzes which can exhibit click rates 1-2 orders of magnitude larger[39]. This emphasizes the objective of resolving acoustic structures on the temporal order of individual clicks. Finally, we report the area under the Precision-Recall curve (PR-AUC).

3 Results

In Table 1, we report F1-score, R-value, and PR-AUC for the proposed self-supervised detection model as well as the energy threshold detector baseline with and without high pass filter (HPF) preprocessing. Our method achieves a maximal F1-score of 0.876, an R-value of 0.887, and a PR-AUC of 0.917. The baseline threshold-based energy detection method operating on raw input audio with no data preprocessing (e.g. denoising) yields an F1-score of 0.576, an R-value of 0.620, and a PR-AUC of 0.571. Inclusion of denoising in the threshold-based energy detection method yields modest performance benefits by eliminating high-amplitude low frequency environmental noise, boosting the F1-score, R-value, and PR-AUC to 0.639, 0.643, and 0.706, respectively. These results suggest that the unsupervised deep learning-based approaches to bioacoustic signal detection implemented in this paper can outperform established baseline techniques for detecting events in acoustic recordings. PR curves are presented in Fig. 3.

In Fig. 1, we demonstrate the self-supervised detection pipeline and results applied to an acoustic recording containing a sequence of clicks. Fig 1 **a)** represents the acoustic waveform, on which the model f operates directly. Fig 1 **b)** presents a spectrogram representation to visualize the spectrotemporal features of the signal. The learned representation $\mathbf{z}$ resulting from

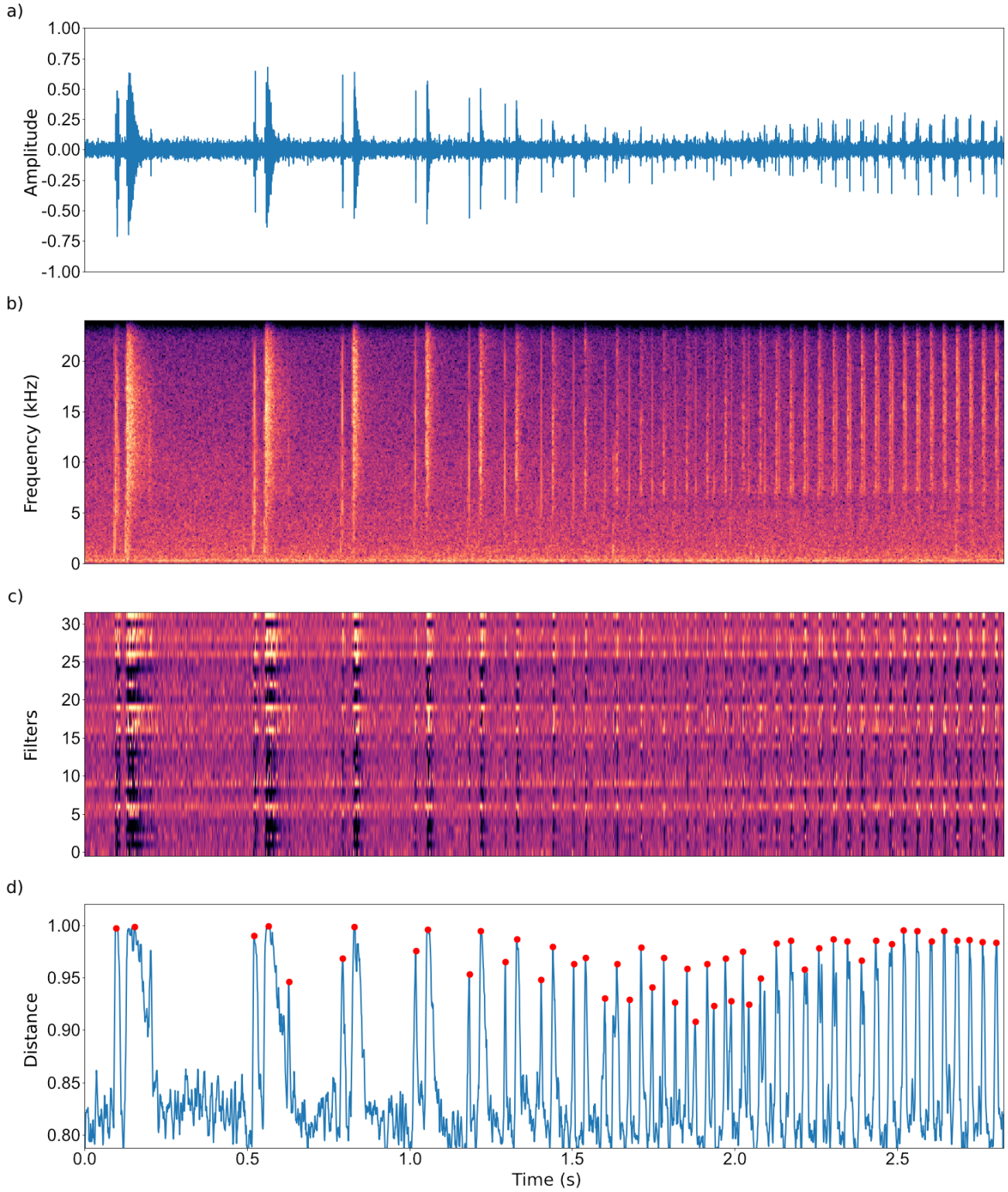


Figure 1 – Application of self-supervised contrastive learning to sperm whale click detection. (**a**) Input acoustic waveform. (**b**) Spectrogram representation. (**c**) The learned representation. (**d**) The distance metric *dissim* demonstrating boundaries of high dissimilarity between acoustic windows. Peaks were detected with $\alpha = 0.1$, $\beta = 0.9$, $\sigma = 5$ms, $\delta = 0.1$

the optimization of the contrastive learning objective is shown in Fig 1 **c)**. Finally, we compute the bounded L2 dissimilarity metric using $\alpha = 0.1$ and $\beta = 0.9$ and plot the result in Fig 1 **d)**. Employing a peak-finding algorithm, we search over prominences δ to obtain the PR-AUC, observing maximal F1-score with $\delta = 0.1$, and we superimpose the detected peaks in Fig 1 **d)**. Together, the quantitative metrics and visualizations in Fig 1 demonstrate the efficacy of applying self-supervised contrastive learning in the context of bioacoustics event detection.

In Fig. 2, we provide initial exploratory results applying self-supervised contrastive learning to other datasets and data modalities. In Fig. 2 **a)**, we segment Bengalese finch (*Lonchura striata domestica*) birdsong[40] into predictions for sequential vocal units, or syllables. In Fig. 2 **b)**, we use the contrastive detection framework to predict behavioral transitions in green sea turtle (*Chelonia mydas*) behavioral dynamic data[41]. In Fig. 2 **c)**, we investigate unsupervised embryo tracking by identifying transitions characteristic of cell divisions in a mouse embryo development dataset[42]. We defer a more thorough analysis to future studies.

4 Discussion

While state-of-the-art computational methods for bioacoustic event detection tend to treat detection as a binary classification and/or presence indication problem and rely on supervised learning techniques, we propose a novel self-supervised approach to detect spectral changes associated with bioacoustic signals. Importantly, our unsupervised method requires no manual annotation during training, shifting the paradigm from existing techniques that expect large quantities of labeled training data. We demonstrate that unsupervised DNN-based approaches to bioacoustic detection and segmentation can outperform established baseline methods, yielding improved results in terms of precision, recall, F1-score, R-value, and PR-AUC.

4.1 Vision for a DL-Based Pipeline for Bioacoustic Monitoring

In line with existing research[21,29], we envision a complete pipeline for ecological data analysis–particularly in the context of bioacoustics–involving discrete DNN-based modules to address ecologically important questions regarding animal communication and behavior. Further, we advocate the development and implementation of *un*supervised techniques to minimize the role of the human observer and circumvent the reliance on–and drawbacks of–manual annotation[2]. Explicitly, we propose a framework encompassing detection, classification, localization, and tracking (DCLT)[36]. The multi-step pipeline involves (1) signal processing; (2) event boundary detection; (3) source separation, localization, and tracking; and (4) clustering and classification. This study serves as a proof-of-concept to demonstrate the application of unsupervised DNN-based methods to solve Steps (1) and (2). We encourage future experimental setups to expand on our results.

Step (1) (i.e. signal processing) requires an emphasis on denoising given the infeasibility of obtaining non-noisy acoustic data in real-world environmental recordings; bioacoustic denoising is an active area of research[43], and to our knowledge, there exist no DNN-based techniques for bioacoustic noise reduction, a regime that differs from the related subfield of human speech enhancement in that human speech enhancement typically relies on supervised methods to separate noise and signal[44]. While we integrate fixed denoising into our model architecture using a high-pass filter to eliminate low-frequency environmental noise, our methods could benefit from enhanced DL-based denoising to remove noise that spans the entire spectral zone of support for the given signal of interest.

The focus of our paper remains Step (2) (i.e. boundary detection). While our results indicate that DNN-based approaches can significantly outperform threshold-based methods (which were used in Koumura & Okanoya, 2016), more recent DL techniques could be employed to further boost performance. For instance, vector-quantized contrastive predictive coding (VQ-CPC) has been used to advance phoneme segmentation and acoustic unit discovery[27].

Step (3) (i.e. bioacoustic source separation) has seen major inroads in recent years[30,45]. This includes the development of unsupervised source separation methods[45], an important consideration that can improve downstream classification performance. While mixture invariant training (MixIT) algorithm[45] provides an effective means for separating sources in bioacoustic recordings, bioacoustic source separation could further benefit from alternative unsupervised methods. The related problems of source localization and source tracking involve extracting signals from continuous acoustic streams, even when the sources responsible for generating the event are traveling in space and time; while DL-based approaches to localization have indicated promising results[3,46], together, separation, localization, and tracking remain underexplored in the literature but represent important considerations when processing ecological data to assess animal behavior and communication[36].

Step (4) (i.e. clustering and classification) can provide important insight into questions concerning animal behavior, communication, and ecology. Analyzing vocal repertoires, classifying vocalizations according to type, identifying unique individuals, and characterizing animal behavioral dynamics represent central classification-based objectives with significant ecological and conservation implications. While existing methods have explored both supervised[4] and unsupervised[16] approaches for classifying input data according to target label (i.e. call type, animal identity, etc.), DNN-based advances could offer improved performance and reduced reliance on the human observer. DNNs are powerful instruments for discovering hidden and cryptic patterns in large-scale multi-modal datasets[2], and unsupervised DNN-based techniques for classification

tasks, in particular, have shown the potential for enhancing clusterability and feature discrimination[47], enabling the scientific community to answer complex questions regarding ecology and animal behavior.

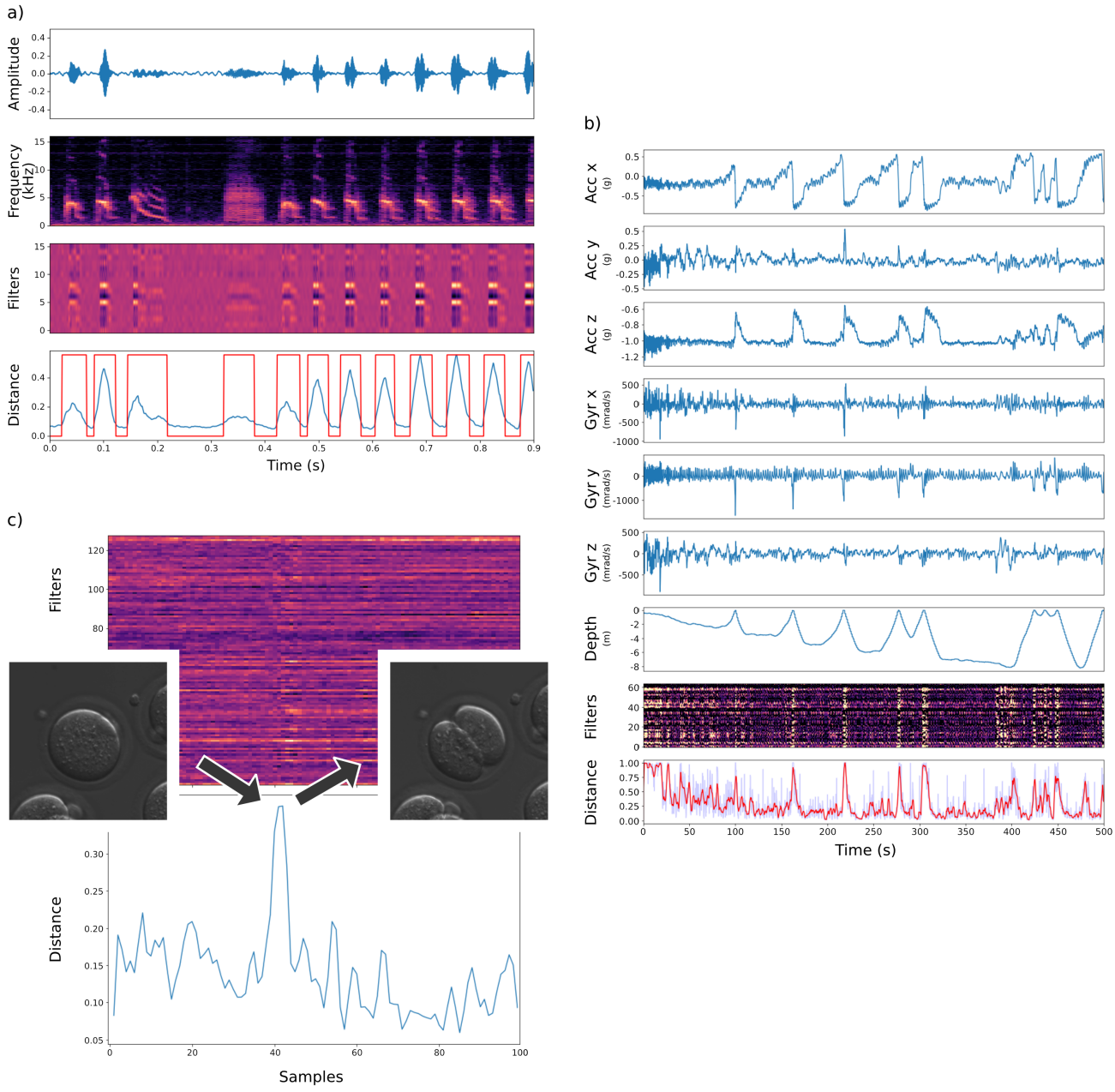


Figure 2 – Schematic experimental applications of self-supervised contrastive learning to other datasets and data modalities. (**a**) Bengalese finch syllable segmentation; vocal units were segmented using a Z-score-based peak detection algorithm; (**b**) Detection of behavioral transitions in free-ranging Green sea turtles according to animal-borne multi-channel accelerometry and gyroscope data. (**c**) Tracking of mouse embryo development.

4.2 Applications to Other Datasets and Data Modalities

Finally, in Fig. 2 we schematically demonstrate that these methods can generalize to other datasets and data modalities beyond bioacoustic click detection. In Fig. 2 **a**), we apply our methods to Bengalese finch[40] syllable segmentation. Bengalese finch birdsong is comprised of sequential vocal elements, known as syllables. We reformulate the detection problem as a segmentation problem, which involves detecting temporal onsets and offsets of signals and extracting the vocal unit bounded by the detections. As in the case of sperm whales, we use a dataset containing acoustic recordings; in particular, we apply the contrastive learning

objective to a collection of song from four Bengalese finches[40] to segment the continuous audio stream into predictions for vocal units.

In Fig. 2 **b)**, we use self-supervised learning to address green sea turtle behavioral dynamics[41]. The aim is to automatically identify, predict, and monitor various behaviors of free-ranging sea turtles in their natural habitat through the use of animal-borne multi-sensor recorders[48]. The dataset consists of multi-channel time series corresponding to acceleration, gyroscope, and depth recorded using animal-borne sensors from 13 immature green sea turtles. We show that self-supervised learning has the potential to reveal transition boundaries between behaviors, providing for automatic segmentation of animal behavior data. Future studies can consider downstream classification and/or clustering[16] to predict the class label of the segmented behavior.

And in Fig. 2 **c)**, we show that self-supervised contrastive learning can be leveraged to address questions regarding embryo tracking and development[42]. We use a mouse embryo tracking database containing 100 samples of embryos progressing to the 8-cell stage[42]. A DNN trained by optimizing the contrastive learning objective encodes embryo image input to a representation emphasizing boundaries and discontinuities in embryo development, which is used to predict transitions between developmental stages.

While we demonstrate the potential for additional implementations, we suggest further studies exploring ecological applications of our methods by employing self-supervised contrastive learning to detect discontinuities, transitions, and events in additional bioacoustics datasets as well as animal behavioral dynamics datasets.

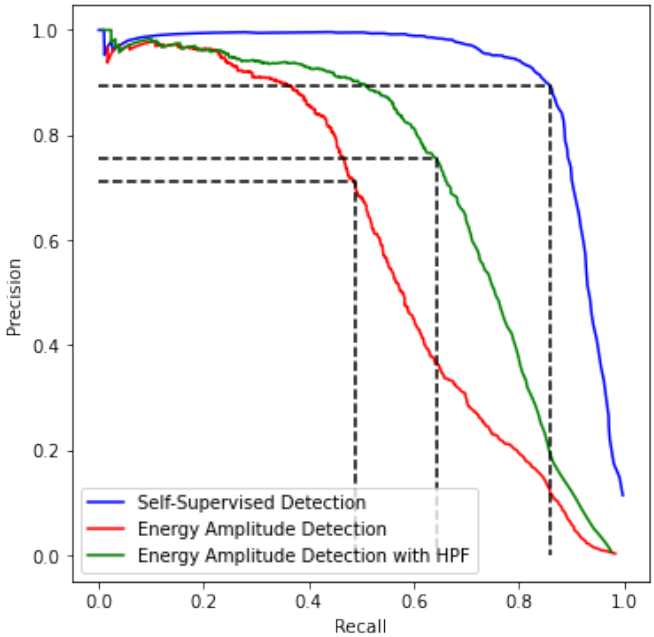


Figure 3 – Precision-Recall (PR) curves for our self-supervised detector (blue), the baseline energy amplitude threshold detector with high pass filter (HPF) preprocessing (green), and the baseline energy amplitude threshold detector with no preprocessing (red). Dashed lines denote PR values corresponding to maximal F1-scores.

5 Conclusion

In this paper, we apply self-supervised representation learning to ecological data analysis with an emphasis on bioacoustic event detection. In particular, we construct a CNN-based model and optimize a contrastive learning objective in accordance with the Noise Contrastive Estimation principle to yield a representation of input audio that encodes features involved in detecting spectral changes and/or boundaries, enabling the model to predict temporal onsets and/or offsets of bioacoustic signals. We compute a dissimilarity metric to compare sequential acoustic windows, and we employ a peak-finding algorithm to detect suprathreshold dissimilarities indicative of a transition from non-signal to signal and vice versa. In the case of sperm whale clicks, we present quantitative performance metrics in the form of F1-scores, R-values, and PR-AUCs, concluding that the unsupervised DL approach based on contrastive representation learning outperforms baseline methods such as energy amplitude detectors. Interestingly, we observe that omitting the smoothing operation σ may enable the resolution of finer-temporal-scale structures such as IPIs, which could allow biologists to identify individual sperm whales[49] using an unsupervised computational

technique; we encourage additional studies to further explore this observation. We also consider applications of the methods to other datasets and data modalities, including bioacoustic data produced by other taxa, behavioral dynamic data, and imaging data, demonstrating that the contrastive learning objective can have wide-ranging implications for ecological data analysis.

This paper serves as an important step towards the realization of a fully automated system for processing bioacoustic data while minimizing the conventional role of the trained expert human observer. Our methods and proposals pave the way for future studies that should aim to construct a complete framework for ecological data analysis in order to elucidate and understand animal behavior and, subsequently, to design better-informed strategies and approaches for species conservation at large.

6 Data Availability

The sperm whale click data that support the findings of this study are publicly available in the 'Best Of' cuts from the Watkins Marine Mammal Sound Database, Woods Hole Oceanographic Institution, and the New Bedford Whaling Museum (https://cis.whoi.edu/science/B/whalesounds/index.cfm). The bengalese finch data[40] are available from *figshare* with identifier https://doi.org/10.6084/M9.figshare.4805749.V5, and the green sea turtle data[41] used in this study are available from *Dryad* with identifier https://doi.org/10.5061/dryad.hhmgqnkd9. The embryo development data[42] that support the results are publicly available online http://celltracking.bio.nyu.edu/. The custom-written code (Python 3.8.3) is available at our GitHub https://github.com/colossal-compsci/SSLUnsupDet.

References

1. Andreas, J. *et al.* Toward understanding the communication in sperm whales. *iScience* **25**, 104393, DOI: https://doi.org/10.1016/j.isci.2022.104393 (2022).
2. Goodwin, M. *et al.* Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. *ICES J. Mar. Sci.* **79**, 319–336, DOI: https://doi.org/10.1093/icesjms/fsab255 (2022).
3. Stowell, D. Computational bioacoustics with deep learning: A review and roadmap. *PeerJ* **10**, DOI: https://doi.org/10.7717/peerj.13152 (2022).
4. Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S. & Gruber, D. F. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Sci. Reports* **9**, DOI: https://doi.org/10.1038/s41598-019-48909-4 (2019).
5. Allen, A. N. *et al.* A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. *Front. Mar. Sci.* **8**, DOI: https://doi.org/10.3389/fmars.2021.607321 (2021).
6. Zhong, M. *et al.* Detecting, classifying, and counting blue whale calls with siamese neural networks. *The J. Acoust. Soc. Am.* **149**, 3086–3094, DOI: https://doi.org/10.1121/10.0004828 (2021).
7. Kahl, S., Wood, C., Eibl, M. & Klinck, H. Birdnet: A deep learning solution for avian diversity monitoring. *Ecol. Informatics* **61**, 101236, DOI: https://doi.org/10.1016/j.ecoinf.2021.101236 (2021).
8. Ruan, W., Wu, K., Chen, Q. & Zhang, C. Resnet-based bio-acoustics presence detection technology of hainan gibbon calls. *Appl. Acoust.* **198**, 108939, DOI: https://doi.org/10.1016/j.apacoust.2022.108939 (2022).
9. White, E. L. *et al.* More than a whistle: Automated detection of marine sound sources with a convolutional neural network. *Front. Mar. Sci.* **9**, DOI: https://10.3389/fmars.2022.879145 (2022).
10. Nguyen Hong Duc, P. *et al.* Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. *Ecol. Informatics* **61**, 101185, DOI: https://doi.org/10.1016/j.ecoinf.2020.101185 (2021).
11. Leroy, E. C., Thomisch, K., Royer, J.-Y., Boebel, O. & Van Opzeeland, I. On the reliability of acoustic annotations and automatic detections of antarctic blue whale calls under different acoustic conditions. *The J. Acoust. Soc. Am.* **144**, 740–754, DOI: https://doi.org/10.1121/1.5049803 (2018).
12. Cronkite, D., Malin, B., Aberdeen, J., Hirschman, L. & Carrell, D. Is the juice worth the squeeze? costs and benefits of multiple human annotators for clinical text de-identification. *Methods Inf. Medicine* **55**, 356–364, DOI: https://doi.org/10.3414/me15-01-0122 (2016).
13. Cartwright, M., Dove, G., Méndez, A. E. M., Bello, J. P. & Nov, O. Crowdsourcing multi-label audio annotation tasks with citizen scientists. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, DOI: https://doi.org/10.1145/3290605.3300522 (ACM, 2019).

14. Cartwright, M. *et al.* Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proc. ACM on Human-Computer Interact.* **1**, 1–21, DOI: https://doi.org/10.1145/3134664 (2017).

15. Bergler, C. *et al.* Orca-spot: An automatic killer whale sound detection toolkit using deep learning. *Sci. Reports* **9**, 10997, DOI: https://doi.org/10.1038/s41598-019-47335-w (2019).

16. Thomas, M. *et al.* A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *J. Animal Ecol.* **91**, 1567–1581, DOI: https://doi.org/10.1111/1365-2656.13754 (2022).

17. Xie, J., Hu, K., Zhu, M. & Guo, Y. Bioacoustic signal classification in continuous recordings: Syllable-segmentation vs sliding-window. *Expert. Syst. with Appl.* **152**, 113390, DOI: https://doi.org/10.1016/j.eswa.2020.113390 (2020).

18. Steinfath, E., Palacios-Muñoz, A., Rottschäfer, J. R., Yuezak, D. & Clemens, J. Fast and accurate annotation of acoustic signals with deep neural networks. *eLife* **10**, e68837, DOI: https://doi.org/10.7554/eLife.68837 (2021).

19. Briggs, F., Raich, R. & Fern, X. A supervised approach for segmentation of bioacoustics audio recordings. *The J. Acoust. Soc. Am.* **133**, 3310–3310, DOI: https://doi.org/10.1121/1.4805500 (2013).

20. Roger, V., Bartcus, M., Chamroukhi, F. & Glotin, H. *Unsupervised Bioacoustic Segmentation by Hierarchical Dirichlet Process Hidden Markov Model*, 113–130 (Springer International Publishing, 2018).

21. Koumura, T. & Okanoya, K. Automatic recognition of element classes and boundaries in the birdsong with variable sequences. *PLOS ONE* **11**, 1–24, DOI: https://doi.org/10.1371/journal.pone.0159188 (2016).

22. Papapanagiotou, V., Diou, C. & Delopoulos, A. Self-supervised feature learning of 1d convolutional neural networks with contrastive loss for eating detection using an in-ear microphone. *2021 43rd Annu. Int. Conf. IEEE Eng. Medicine & Biol. Soc. (EMBC)* 7186–7189, DOI: https://doi.org/10.1109/EMBC46164.2021.9630399 (2021).

23. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, no. 149 in ICML'20, 11, DOI: https://doi.org/10.5555/3524938.3525087 (JMLR.org, 2020).

24. van den Ooord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://arxiv.org/abs/1807.03748 (2018).

25. Fonseca, E., Ortego, D., McGuinness, K., O'Connor, N. E. & Serra, X. Unsupervised contrastive learning of sound event representations. Preprint at https://arxiv.org/abs/2011.07616 (2020).

26. Kreuk, F., Keshet, J. & Adi, Y. Self-supervised contrastive learning for unsupervised phoneme segmentation. Preprint at https://arxiv.org/abs/2007.13465 (2020).

27. van Niekerk, B., Nortje, L. & Kamper, H. Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. Preprint at https://arxiv.org/abs/2005.09409 (2020).

28. Denton, T. *et al.* Handling background noise in neural speech generation. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, 667–671, DOI: 10.1109/IEEECONF51394.2020.9443535 (2020).

29. Sainburg, T. & Gentner, T. Q. Toward a computational neuroethology of vocal communication: From bioacoustics to neurophysiology, emerging tools and future directions. *Frontiers* DOI: https://doi.org/10.3389/fnbeh.2021.811737/full (2021).

30. Bermant, P. C. Biocppnet: automatic bioacoustic source separation with deep neural networks. *Sci. Reports* **11**, 23502, DOI: https://doi.org/10.1038/s41598-021-02790-2 (2021).

31. Gutmann, M. & Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. W. & Titterington, M. (eds.) *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9 of *Proceedings of Machine Learning Research*, 297–304 (2010).

32. Lee, B. D. *et al.* Ten quick tips for deep learning in biology. *PLOS Comput. Biol.* **18**, 1–20, DOI: https://doi.org/10.1371/journal.pcbi.1009803 (2022).

33. Smith, S. *The Scientist and Engineer's Guide to Digital Signal Processing*, chap. Windowed-Sinc Filters (California Technical Publishing, 1999).

34. Ravanelli, M. & Bengio, Y. Interpretable convolutional filters with sincnet. Preprint at https://arxiv.org/abs/1811.09725 (2019).

35. Lopatka, M., Adam, O., Laplanche, C., Motsch, J.-F. & Zarzycki, J. Sperm whale click analysis using a recursive time-variant lattice filter. *Appl. Acoust.* **67**, 1118–1133, DOI: https://doi.org/10.1016/j.apacoust.2006.05.011 (2006). Detection and localization of marine mamals using passive acoustics.

36. Zimmer, W. M. *Passive Acoustic Monitoring of Cetaceans* (Cambridge University Press, 2011).

37. Bøttcher, A., Gero, S., Beedholm, K., Whitehead, H. & Madsen, P. T. Variability of the inter-pulse interval in sperm whale clicks with implications for size estimation and individual identification. *The J. Acoust. Soc. Am.* **144**, 365–374, DOI: https://doi.org/10.1121/1.5047657 (2018).

38. Whitehead, H. & Weilgart, L. Click rates from sperm whales. *The J. Acoust. Soc. Am.* **87**, 1798–1806, DOI: https://doi.org/10.1121/1.399376 (1990).

39. Fais, A., Johnson, M., M and. Wilson, Aguilar Soto, N. & Madsen, P. Sperm whale predator-prey interactions involve chasing and buzzing, but no acoustic stunning. *Sci. Reports* **6**, 28562, DOI: https://doi.org/10.1038/srep28562 (2016).

40. Nicholson, D., Queen, J. E. & Sober, S. J. Bengalese finch song repository, DOI: https://doi.org/10.6084/M9.figshare.4805749.V5 (2017).

41. Jeantet, L. *et al.* Raw acceleration, gyroscope and depth profiles associated with the observed behaviours of free-ranging immature green turtles in martinique, DOI: https://doi.org/10.5061/dryad.hhmgqnkd9 (2020).

42. Cicconet, M., Gutwein, M., Gunsalus, K. C. & Geiger, D. Label free cell-tracking and division detection based on 2d time-lapse images for lineage analysis of early embryo development (2014).

43. Xie, J., Colonna, J. G. & Zhang, J. Bioacoustic signal denoising: a review. *Artif. Intell. Rev.* **54**, 3575–3597, DOI: https://doi.org/10.1007/s10462-020-09932-4 (2021).

44. Saleem, N. & Khattak, M. I. A review of supervised learning algorithms for single channel speech enhancement. *Int. J. Speech Technol.* **22**, 1051–1075, DOI: https://doi.org/10.1007/s10772-019-09645-2 (2019).

45. Denton, T., Wisdom, S. & Hershey, J. R. Improving bird classification with unsupervised sound separation. *ICASSP 2022 - 2022 IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP)* 636–640 (2022).

46. Francl, A. & McDermott, J. H. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nat. Hum. Behav.* **6**, 111–133, DOI: https://doi.org/10.1038/s41562-021-01244-z (2022).

47. Károly, A. I., Fullér, R. & Galambos, P. Unsupervised clustering for deep learning: A tutorial survey. *Acta Polytech. Hungarica* **15** (2018).

48. Jeantet, L. *et al.* Behavioural inference from signal processing using animal-borne multi-sensor loggers: a novel solution to extend the knowledge of sea turtle ecology. *Royal Soc. Open Sci.* **7**, 200139, DOI: https://doi.org/10.1098/rsos.200139 (2020).

49. Hirotsu, R., Ura, T., Bahl, R. & Yanagisawa, M. Analysis of sperm whale click by music algorithm. In *OCEANS 2006 - Asia Pacific*, 1–6, DOI: https://10.1109/OCEANSAP.2006.4393900 (2006).

7 Author Information

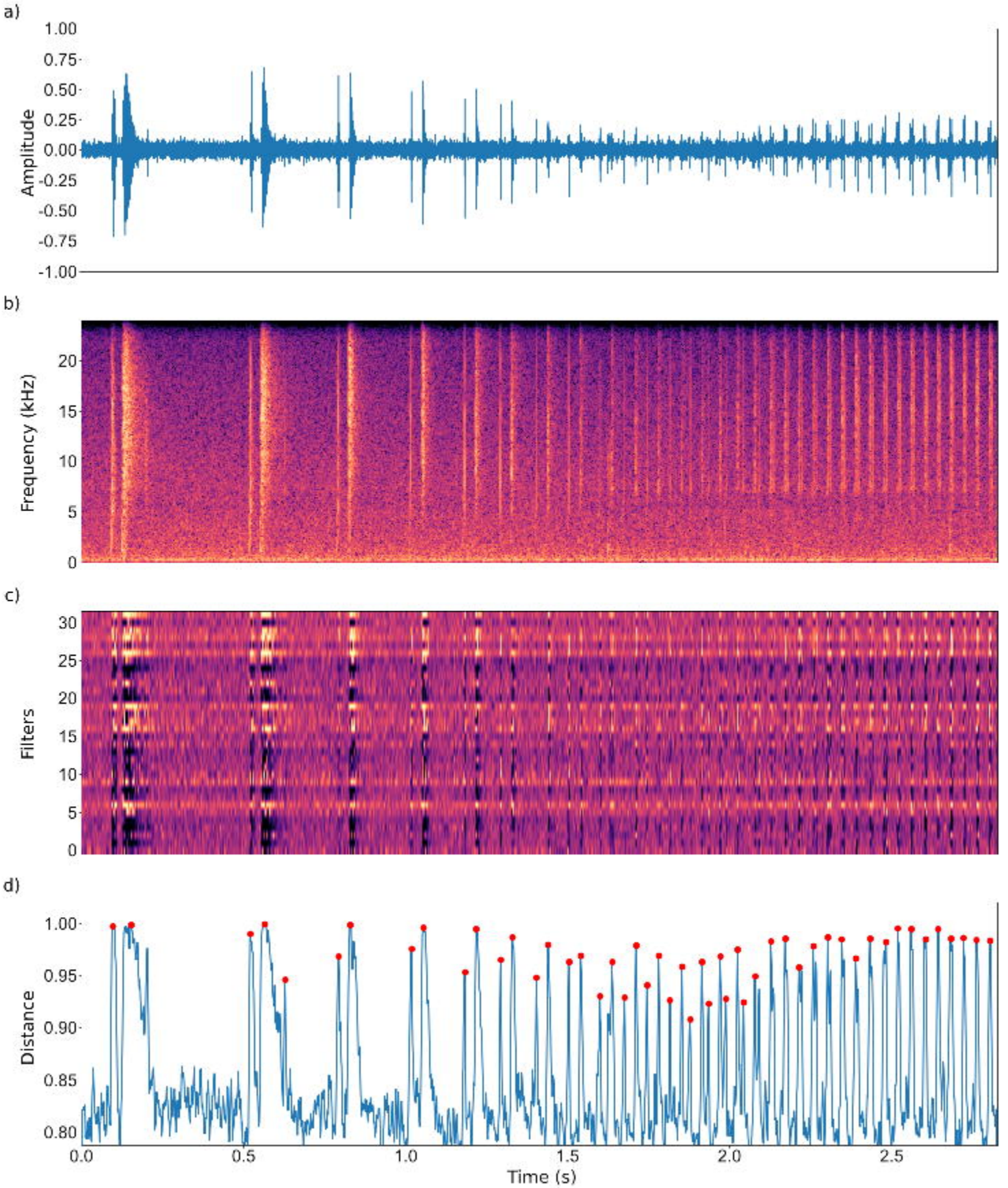
Affiliations
Colossal Biosciences, Austin, TX, USA
Form Bio, Austin, TX, USA

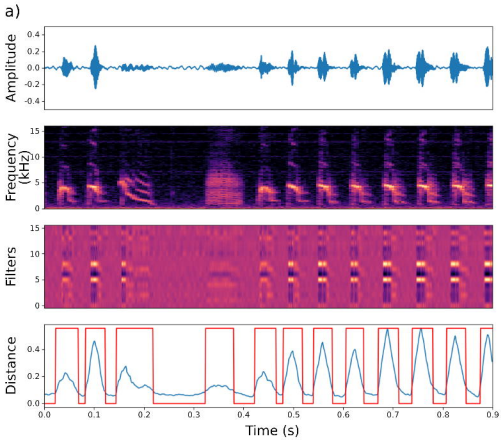
Contributions
P.C.B performed task setting, data processing, machine learning, article writing, figure making, L.B. assisted in experimental design, A.J.T. supervised the analysis, and all authors wrote and reviewed the manuscript.

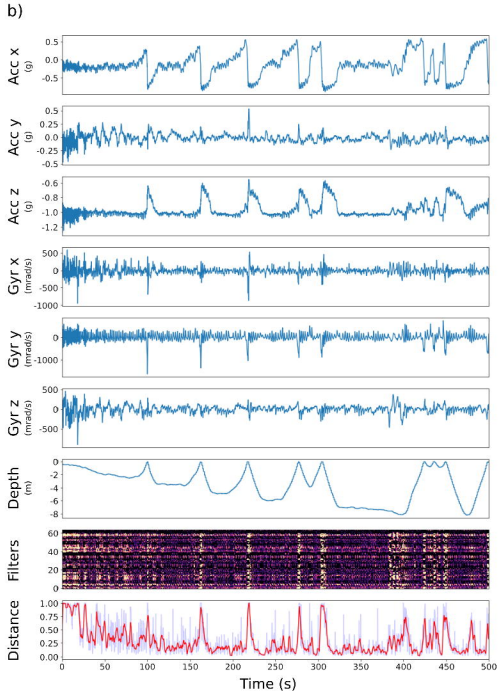
Corresponding Author
Correspondence to P.C. Bermant

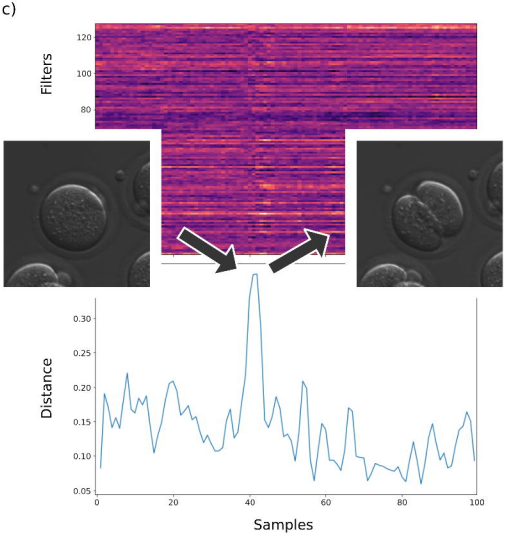
8 Competing Interests

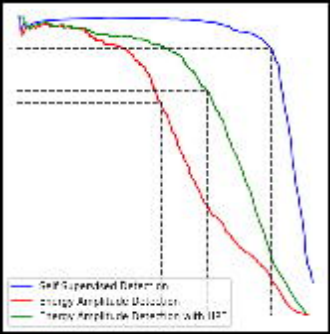
The authors declare no competing interests.


a)
Amplitude
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
b)
Frequency (kHz)
20
15
10
5
0
c)
Filters
30
25
20
15
10
5
0
d)
Distance
1.00
0.95
0.90
0.85
0.80
0.0
0.5
1.0
1.5
2.0
2.5
Time (s)


a)
Amplitude
Frequency (kHz)
Filters
Distance
Time (s)


b)
Acc x (g)
Acc y (g)
Acc z (g)
Gyr x (mrad/s)
Gyr y (mrad/s)
Gyr z (mrad/s)
Depth (m)
Filters
Distance
Time (s)


c)
Filters
Distance
Samples


Self Supervised Detection
Energy Amplitude Detection
Energy Amplitude Detection with UPT