Marine Mammal Science

ARTICLE OPEN ACCESS

Special Issue—Machine Learning and Artificial Intelligence in Marine Mammal Research

Adaptive Acoustic Monitoring for Endangered Cook Inlet Beluga Whales in Complex Soundscapes

Manuel Castellote[1] | Rahul Dodhia[1] | Daniela Ruiz[1] | Zhongqi Miao[1] | Pablo Arbelaez[2] | Verena Gill[3] | Lori Polasek[4] | Juan M. Lavista Ferres[1]

[1]Microsoft AI for Good Research Lab, Redmond, Washington, USA | [2]Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Bogota, Colombia | [3]Protected Resources Division, NOAA Fisheries, Anchorage, Alaska, USA | [4]Marine Mammal Program, Alaska Department of Fish & Game, Juneau, Alaska, USA

Correspondence: Manuel Castellote (v-manoloc@microsoft.com)

Received: 1 January 2026 | **Revised:** 22 May 2026 | **Accepted:** 26 May 2026

Keywords: cook inlet beluga whale | deep learning | domain generalization | endangered species conservation | multi-species classification | passive acoustic monitoring

ABSTRACT

Effective conservation of the endangered Cook Inlet beluga whale (*Delphinapterus leucas*) requires comprehensive spatiotemporal data, yet monitoring efforts remain spatially biased, underrepresenting important southern habitats. Passive acoustic monitoring (PAM) provides the necessary broad-scale coverage, but its expansion introduces substantial computational challenges, including high soundscape variability, rarity of target species' signals relative to background noise, and multi-species signal interference that can compromise classifier performance. We present an open-source deep learning framework designed to improve robustness, adaptability, and domain generalization of PAM analyses for this population. Building on a beluga-focused binary classifier, we implemented a dual-stage model that separates signal detection from species classification and expanded the framework to a multi-species context that includes killer whales (*Orcinus orca*) and humpback whales (*Megaptera novaeangliae*). Contrastive audio-language models were used to efficiently increase annotation coverage for previously underrepresented species, while active learning enabled iterative refinement of model performance on new data. The framework was applied to PAM datasets from management and ecologically significant regions of lower Cook Inlet. Results demonstrated improved detection of rare species occurrences and increased confidence in daily beluga presence estimates, strengthening the role of PAM in informing recovery efforts and management decisions. This transferable workflow supports continued advancement of large-scale, long-term marine mammal monitoring programs.

1 | Introduction

1.1 | Cook Inlet Beluga Population

Recent research on the reason(s) for the lack of recovery for the endangered Cook Inlet beluga whale (*Delphinapterus leucas*) has been focused primarily on samples or data collection in their northern range during spring to fall (hereafter summer habitat) (e.g., Himes Boor et al. 2022; McGuire et al. 2021; McGuire, Shelden, et al. 2020; McGuire et al. 2021; McHuron et al. 2023; Warlick et al. 2024). During this period, the population of 331 whales (Goetz et al. 2023) gathers in habitat hot spots where anadromous fish concentrate or become more accessible to be preyed upon during their spawning runs (Hobbs et al. 2005; McGuire, Himes Boor, et al. 2020). This facilitates research as the spatial distribution of whales is predictable and concentrated, and favorable weather

Manuel Castellote and Rahul Dodhia contributed equally to this work.

conditions and absence of sea ice make field work easier. Although causes for the lack of recovery are not yet clear, important information is being obtained on habitat use, fecundity, demography and mortality, effects of noise, diet, and population trends (e.g., Himes Boor et al. 2022; McGuire, Stephens, et al. 2020; McHuron et al. 2023; Small et al. 2017; Warlick et al. 2024). Previous work has focused on the upper inlet, while the southern portion of their current range that they predominantly occupy from fall to spring (hereafter winter habitat) remains underexplored. So far, remote sensing such as aerial surveys for abundance estimates and distribution and passive acoustic monitoring (PAM) for seasonality and noise characterization have been the only recent efforts in this wide range of their critical habitat, encompassing the mid and southern Cook Inlet (Castellote, Small, et al. 2020; Shelden et al. 2016). This large area is used by belugas when adverse weather, short daylight, and sea ice presence makes visual methods challenging, however, PAM remains a cost-effective approach unaffected by the adverse conditions of the environment (Zimmer 2011).

The identification of winter grounds is vital for actions aimed at recovering the Cook Inlet beluga population. As their historic summer range has contracted, predominantly focusing on the upper inlet (Rugh et al. 2010), the description of winter habitat use provides new opportunities for targeted conservation measures. For example, some areas of Cook Inlet beluga critical habitat are within or adjacent to ongoing human activities that include oil and gas exploration and development, port construction and marine traffic, mining, and renewable energy (Castellote et al. 2024). Understanding how beluga and other protected species use this habitat would greatly inform mitigations to aid in the recovery of these at-risk populations. Therefore, further application of PAM to this region of Cook Inlet is important.

1.2 | AI for Cook Inlet Beluga

The development of AI tools for bioacoustics, in particular the application of deep learning models for detecting and classifying animal signals, has revolutionized this field of science (Stowell 2022). However, there is still much work to be done for these sophisticated signal processing tools to become more user-friendly for the non-computer scientists, and to be more effectively integrated into typical data analysis workflows in conservation research programs. The Cook Inlet Beluga Acoustics (CIBA) program is an example of where this has been done. In 2019, the analysis methods switched from semi-automated tonal detectors requiring time-intensive manual validation, to an ensembled deep learning convolutional neural network model, reducing the labor-intensive and time-consuming analysis process and increasing the accuracy of the overall detection results (Zhong et al. 2020). As the acoustic monitoring effort became more efficient, sampling effort expanded spatially, and consequentially, more spatially inclusive annotation datasets were required for model fine-tuning purposes. Generalization became a challenge for the deep learning model used in this monitoring program. This catch-up between spatial monitoring expansion and the model fine-tuning required to maintain beluga detection accuracy reduced the initial savings in time and labor.

Most long-term programs (i.e., decades) experience changes in sensors and sampling tools led by technological development, adding the complication of domain shifting. Changes in hydrophone sensitivity, recording gain, sampling rate, duty cycle, and so forth cause unavoidable differences in data collection that could introduce sampling bias. The underperforming machine learning analytical process required additional tools and improvements to keep up with the CIBA expanding research program. Additionally, expanding acoustic monitoring to the winter habitats of Cook Inlet belugas introduced the analytical challenge of distinguishing between new cetacean signals within the data. Humpback whales (*Megaptera novaeangliae*) and killer whales (*Orcinus orca*) are rarely present in the northern region of Cook Inlet due to the high currents, turbidity, and shallow nature of this habitat, but are frequently present in the southern end of the inlet (Saulitis et al. 2015; Shelden et al. 2003). Beluga vocal behavior is very diverse and shares enough acoustic traits with killer whale and humpback whale signals that even a fine-tuned binary model can generate high-confidence false detections of belugas. Therefore, a multi-species classifier in the analysis pipeline was required.

The two-stage modeling architecture adopted in this study followed a sequential learning logic in which the outputs of a first model were used as inputs to a second model to refine decisions. Clarfeld et al. (2025) provide a particularly clear demonstration of this principle, showing that secondary logistic regression models trained on primary model output features achieved 84%–90% accuracy in separating true and false positive detections of Ruffed Grouse (*Bonasa umbellus*) and argue that the approach is broadly transferable across classification models and taxa. The two-stage architecture in the CIBA framework extends this approach but with a specific operational objective: rather than using a secondary model purely to filter false positives from a single-species primary classifier, we used the first stage to detect any cetacean acoustic energy broadly and the second stage to resolve species identity among beluga, killer whale, and humpback whale vocalizations. This is the logic that underlies ANIMAL-SPOT (Bergler et al. 2022), where a first model performs binary signal detection and the resulting detected segments are then passed to a second model for multi-class species or call-type classification.

The two-stage design choice in this work was motivated directly by the challenges of Cook Inlet's complex soundscape and the conservation management consequences of identification in a critically endangered population monitoring context. In long-term PAM datasets, most recordings consist of environmental noise, with comparatively few containing target vocalizations. We hypothesized that framing the problem as a single multi-class classifier (noise plus multiple species) requires the model to learn highly imbalanced decision boundaries simultaneously, whereas, by separating the stages, the first stage can be optimized for high sensitivity under extreme imbalance, and the second stage can operate on a substantially enriched subset of candidate signals where inter-species discrimination becomes the primary objective.

2 | Objectives

The overarching aim of this study was to advance the robustness, accuracy, and scalability of automated PAM for the

CIBA program. We transitioned from a binary beluga detection model to an adaptive, end-to-end AI pipeline capable of multi-species classification across heterogeneous acoustic environments.

The specific objectives of this work were to

- Expand annotation coverage for underrepresented species by integrating a zero-shot contrastive audio-language model, mitigating the manual labeling bottleneck for new target classes.
- Implement multi-species classification via a two-stage deep neural network (DNN) that decouples broadband signal detection from species-level identification, enabling the simultaneous discrimination of beluga, humpback, and killer whale vocalizations.
- Achieve temporal shift invariance through an overlapping inference sliding-window strategy. This approach provides a natural form of data augmentation by generating time-shifted views of acoustic events, ensuring the model identifies vocalizations regardless of their position within a data segment while preserving physical signal integrity.
- Improve environmental generalization via an active learning loop, which enables iterative pipeline refinement and resistance to domain shift. By strategically incorporating novel, previously unseen data by the model as the monitoring program evolves, the framework maintains accuracy across shifting acoustic environments and hardware configurations.
- Quantify performance gains and resistance to domain shift by evaluating the framework against two distinct CIBA datasets a decade apart and collected with very different recording systems, comparing its multi-species accuracy against prior binary analytical methods.
- Provide an open-source, transferable framework with a modular architecture that can be adapted by other long-term PAM programs targeting different species assemblages and recording systems.

3 | Methods

3.1 | Study Area

The southwestern region of Cook Inlet, Alaska, is characterized by a complex coastal landscape dominated by fjord estuaries located within designated critical habitat for the endangered Cook Inlet beluga (CIB) population (Goetz et al. 2012; NMFS 2008) (Figure 1). This region lies along the western shore of Lower Cook Inlet and is shaped by strong glacial and tectonic processes, resulting in steep coastal relief, soft sediment shorelines, and a dynamic estuarine environment (Molnia and Williams 2008). The estuaries in this region are defined by brackish waters, high turbidity due to glacial silt and sediment transport, seasonal ice cover during winter months, and relatively shallow depths, typically less than 60 m (Burell and Matthews 1974). Most estuaries receive freshwater input from multiple river systems, with the primary inflows originating from glacial rivers that drain into the heads of the bays.

The surrounding terrestrial landscape includes Lake Clark National Park and Preserve and the Chigmit Mountains, a rugged subrange of the Alaska Range characterized by volcanic peaks, alpine tundra, and extensive glacial coverage (Lanik et al. 2021). The region supports a diverse array of wildlife, including brown bears (*Ursus actos*), moose (*Alces alces*), and migratory bird species, and it serves as an important seasonal foraging and calving area for marine mammals (Hobbs et al. 2005; Hobbs et al. 2000). The coastal and estuarine ecosystems in this area are ecologically productive and dynamic, influenced by tidal mixing, sedimentation, and seasonal freshwater discharge. These characteristics make the southwestern Cook Inlet a biologically rich and environmentally complex region critical to the survival and recovery of the Cook Inlet beluga population (NMFS 2016).

3.2 | Instrumentation

Low-profile acoustic moorings evolved over the span of the program (Castellote et al. 2016; Castellote, Small, et al. 2020; Castellote et al. 2024; Lammers et al. 2013), but basically include an echolocation logger (C-POD until 2020 or F-POD starting in 2021, Chelonia Ltd., Cornwall, UK), and an acoustic recorder (Ecological Acoustic Recorder (EAR), Oceanwide Science Institute, Makawoa, HI, US for 2008–2016; DSG-ST, Loggerhead Instruments, Sarasota, FL, US for 2017–2019; and ST-500, Ocean Instruments, Auckland, NZ for 2019–2023) to sample background noise and record marine mammal social signals. The echolocation logger and acoustic recorder were housed within a hydrodynamically shaped syntactic foam mooring package equipped with an acoustical release (Edgetech PORT-LF, or Vemco AR-2, Innovasea Systems Inc., Boston, MA, US) for instrument recovery. These packages were deployed at no less than 20 m depth to avoid interaction with ice in winter, using a 160 lb (in-air weight) expendable steel anchor connected to the mooring package with 3 m of Vectran line (3/8 inch with polyester anti-strumming sleeve) to a swivel and to an acoustic release link. Recording configurations for the CIBA program increased in capacity over time: EAR units operated at 16 kHz with a 25% duty cycle, or 25 kHz with a 10% duty cycle; DSG-ST units at 24 kHz with a 50% duty cycle; ST-500 units at 48 kHz with a 50% duty cycle.

3.2.1 | Development Dataset Annotation and Labeling

Annotation sources included vocalizations from belugas, killer whales, and humpbacks (see Table 1). Annotations came from five CIBA datasets from 2017 to 2018 for belugas and four datasets from 2019 for killer whales and humpback whales (Table S1). Beluga annotations were a subset of the training data used by (Zhong et al. 2020) for their binary model. The subset was based on a proportional stratified sampling strategy to create a reduced dataset of 17,617 annotations from seven datasets totaling 429,293 annotations. This method ensured that the geographical distribution of the original data was accurately reflected in the smaller subsample.

Sound files were segmented into 2 s windows following (Zhong et al. 2020) as the duration of Cook Inlet beluga calls and whistles rarely exceeds this duration. For our subset annotations,

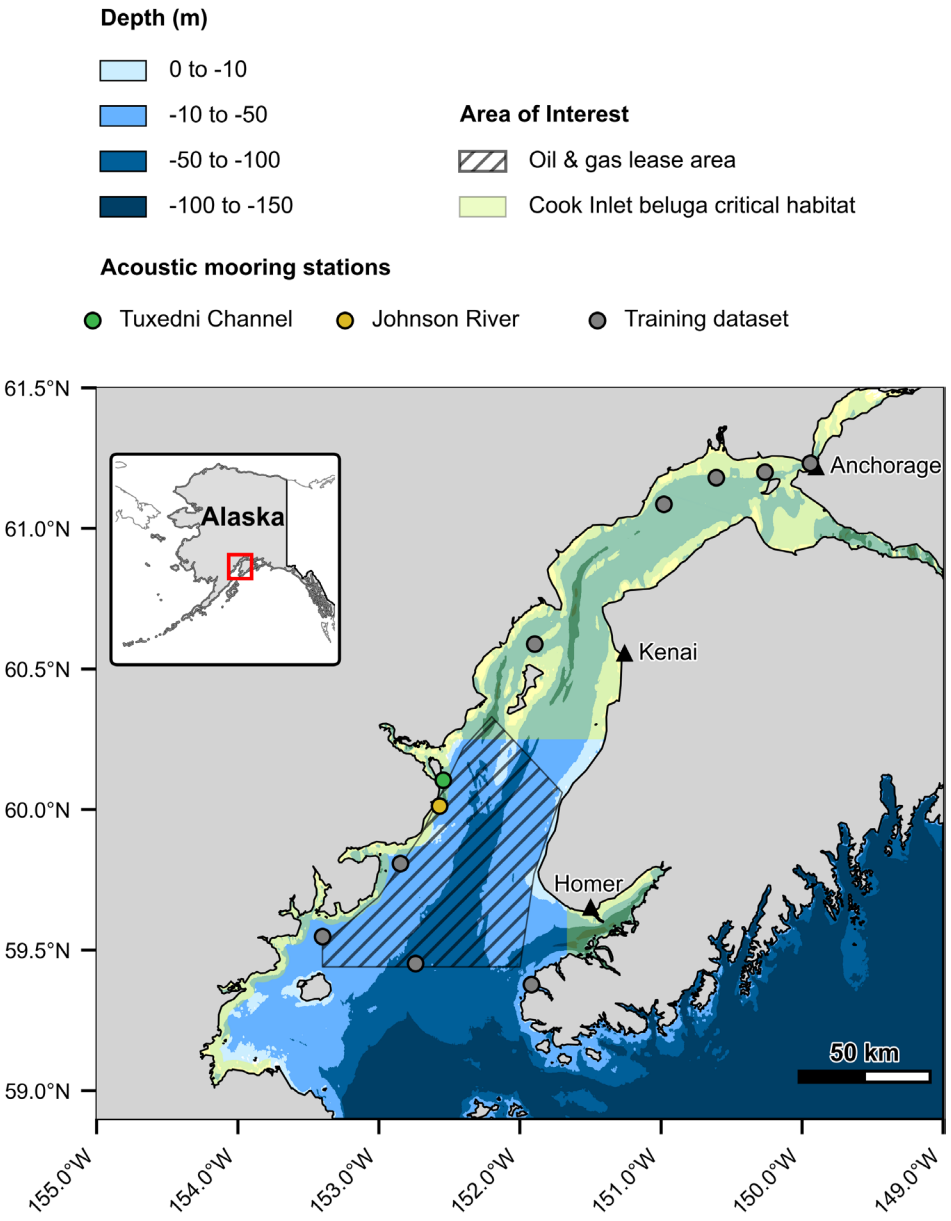

FIGURE 1 | Map of Cook Inlet, AK, acoustic mooring stations included in this study, federal oil and gas lease area, and beluga critical habitat.

TABLE 1 | Annotation sample size for the development dataset, including contrastive language-audio pretraining results, and proportions for the three species of cetaceans included in the classification.

Species	Annotations (percentage of total)	Total hours (percentage of total)	Spectrograms (percentage of total)	Positive spectrograms (percentage of all positives)	Negative spectrograms (percentage of all negatives)
Humpback	14,225 (31.56%)	53.61 (38.98%)	96,504 (38.98%)	24,126 (37.40%)	72,378 (39.54%)
Killer whale	13,233 (29.36%)	46.43 (33.76%)	83,576 (33.76%)	20,894 (32.39%)	62,682 (34.24%)
Beluga	17,617 (39.08%)	37.49 (27.26%)	67,484 (27.26%)	19,495 (30.22%)	47,989 (26.22%)
Total	45,075 (100%)	137.54 (100%)	247,564 (100%)	64,515 (100%)	183,049 (100%)

only four exceeded 2 s (0.03%) with a maximum duration of 3 s. During training, the 2 s windows were set to overlap their neighbors by 400 ms. This overlapping sliding-window strategy inherently provided a form of data augmentation by generating smoothly time-shifted views of acoustic events across adjacent segments. Unlike conventional augmentation

techniques that introduce synthetic perturbations (e.g., noise injection or zero-padding), this approach increases temporal variability while preserving the physical integrity of the original signal.

Each 2 s window was assigned a label using a temporal-overlap rule: if at least 150 ms of an annotated signal appeared within the spectrogram window, the spectrogram was labeled as positive. Species vocalization overlap was not expected in these datasets and therefore not considered in the processing workflow.

While the inference window duration of 2 s was targeted for Cook Inlet beluga vocalizations, this length was also adequate for killer whale and humpback whale signals. Annotations longer than 2 s in our training dataset typically occurred when multiple calls or whistles overlapped. The main exception was humpback whale feeding calls, which commonly last several seconds (D'Vincent et al. 1985). We could not find any published description of killer whale or humpback whale call repertoires specific to Cook Inlet, but the populations typically encountered in the lower inlet and adjacent regions are well characterized acoustically. The Cook Inlet beluga vocal repertoire comprises whistles, pulsed calls, and combined calls whose primary communicative energy is concentrated below 12 kHz (Brewer et al. 2023, 2026), the upper frequency limit of our model's analysis. Ultrasonic high-frequency burst pulse calls falling outside our analyzed frequency range have been documented in other beluga populations but not yet in Cook Inlet (Vergara et al. 2025). As these signals co-occur with non-ultrasonic call types, any such activity would still be accompanied by detectable signals within our analyzed frequency range. Humpback whales present in lower Cook Inlet during fall and winter are expected to produce primarily non-song vocalizations associated with feeding and social interactions, with energy concentrated well within the 0–12 kHz range analyzed by our model (Cerchio and Dahlheim 2001; Fournet et al. 2018). Killer whales in lower Cook Inlet and the adjacent northern Gulf of Alaska belong to resident and transient ecotypes acoustically characterized in Prince William Sound and Kenai Fjords (Myers et al. 2025; Saulitis et al. 2005). Their pulsed calls and whistles have peak energy between 1 and 12 kHz, with only the upper harmonics of whistles extending beyond this range, confirming that the bulk of communicative acoustic energy for all three species falls well within the frequency range captured by our model.

3.2.2 | Contrastive Language-Audio Pretraining-Based Expansion of Humpback and Killer Whale Annotations

We used contrastive language-audio pretraining (CLAP) to expand annotations for humpback and killer whales, which were underrepresented relative to belugas in our training data, and obtained a more balanced proportion of annotations across the three species for model training purposes. Additional vocalizations for both species were drawn from 3-month long continuous recording datasets collected in 2019 in areas adjacent to Tuxedni Bay (Chinitna and Iniskin bays, Castellote, Stocker, and Brewer 2020).

CLAP is a multimodal framework that aligns audio and text through dual encoders trained with contrastive learning (Wu et al. 2024). To identify effective prompts, we tested a range of positive and negative language queries on 10 preselected 10-min files from Tuxedni Bay and Chinitna Bay. Five files contained substantial environmental noise (e.g., waves, rain, sea-ice noise, birds, vessel noise), and five contained humpback and killer whale signals spanning varied signal-to-noise ratios. Positive prompts included both direct semantic descriptions of whale sounds (e.g., "whale vocalizations") and indirect semantic cues referring to acoustic attributes or analogues (e.g., "distant tones"). Negative prompts were used to suppress common non-target signals and included terms such as "noise," "wind," "waves," and more specific descriptors of ship or environmental noise (additional details in Supporting Information). While CLAP was used here to illustrate the potential of audio-language models as annotation tools, more domain-specific alternatives may offer improved retrieval performance for bioacoustic applications, as we expect this to be a rapidly evolving landscape.

For each candidate prompt, we quantified the number of correctly and incorrectly selected segments, and the prompt achieving the highest correct-to-incorrect ratio was selected for large-scale inference. For all results, the negative prompt was maintained fixed using "noise, wind, waves" as this combination kept the number of incorrect segment selections at a minimum when compared to each of these terms independently.

3.2.3 | Train/Validation/Test Partitions

Windows were partitioned into training (70%), validation (15%), and test (15%) sets following common machine learning practice for balanced datasets. The splitting procedure was performed only once and at the recording level (i.e., grouped by audio file), ensuring that windows originating from the same sound file were not distributed across different partitions. This guaranteed that evaluation was conducted on recordings entirely unseen during training, thereby preventing data leakage (Table 2). Stratification by class label was applied during splitting to preserve class proportions across all partitions.

For the three-class model, the identical data partitions were reused. After splitting, negative (noise) examples were excluded from each subset, as they do not constitute a class in this model. Thus, the original partitioning structure was preserved, while the class distribution was recalculated considering only the remaining positive categories.

3.2.4 | Spectrogram Generation

Each 2-s window was converted into a mel-spectrogram using a 2048-point fast Fourier transform (FFT), a 256-sample hop, and 224 mel filters. Spectrograms were created in the power–mel domain, transformed to decibels over an 80 dB range. Output tensors were normalized per sample during training.

TABLE 2 | Development dataset.

Species name	Train count	Train %	Val count	Val %	Test count	Test %
Noise	108,212	72.1	27,421	76.5	47,416	77.0
Humpback	15,722	10.5	3402	9.5	5002	8.1
Killer whale	13,049	8.7	1487	4.2	6358	10.3
Beluga	13,205	8.8	3512	9.8	2778	4.5
Total	150,188	100.0	35,822	100.0	61,554	100.0

Note: Annotated data partition per class label for training, testing and validating (val) the binary and multi-class models.

Because of the longevity of the CIBA program, sampling rates have been increasing over time (16, 24, 25, and 48 kHz). Therefore, all audio data were resampled to a uniform 24 kHz rate, providing a spectrogram bandwidth of 0–12 kHz, an optimal compromise focused on retaining the full frequency range of beluga vocalizations. This choice avoided too much compression of humpback whale vocalizations at lower frequencies while minimizing truncation of upper harmonics in killer whale calls and whistles. Recordings sampled above 24 kHz were downsampled, resulting in the loss of higher frequencies, whereas recordings at 16 kHz were upsampled. For upsampled files, mel-frequency bands above the original Nyquist limit (8–12 kHz) were populated using mean mel-dB values from sub-Nyquist bins within the same 2-s spectrogram segment. This approach approximated the background energy distribution of the original signal and ensured continuity, preventing abrupt gaps or low-amplitude artifacts in unsupported frequency regions.

Several denoising and normalization methods (median filtering, Wiener filtering, Per-Channel Energy Normalization PCEN; Lostanlen et al. 2019) were evaluated but did not improve generalization performance and were excluded from the final workflow. Explicit spectrogram augmentation techniques were also tested during base training; however, they did not improve performance on unseen datasets (see Supporting Information for details). Consequently, the final pipeline relies on the sliding-window strategy for augmenting data variability.

3.3 | Base Model Architecture and Training

Since the passive acoustic monitoring program is designed to support the conservation of this endangered population, the study was structured with beluga detection as the primary application objective. However, the model architecture and training procedure were developed to perform robust multi-class classification across all signal categories included in the dataset.

We adopted a sequential two-stage modeling framework in which a first model performs broad cetacean signal detection, and a second model resolves species identity. The binary detector and species classifier were trained as independent networks receiving identical Mel-spectrogram inputs. During inference, the classifier output was evaluated only when the detector probability exceeded the predefined threshold; otherwise the segment was labeled as no whale. Similar approaches have demonstrated strong performance in bioacoustic pipelines (e.g., Bergler et al. 2022; Clarfeld et al. 2025), particularly in scenarios characterized by rare biological events embedded within long-duration environmental recordings.

To evaluate this design, we tested three configurations. First, a single-stage four-class model with labels defined as no whale (0), humpback (1), killer whale (2), beluga (3); second, a two-stage approach combining a binary classifier (no whale and whale) followed by a three-class classifier (humpback, killer whale, and beluga); and finally, a two-stage model in which the first stage was again a binary classifier, but this time followed by a four-class classifier (no whale, humpback, killer whale, and beluga). The binary+four-class formulation was introduced to assess whether the second stage could correct false positives from the detector while still avoiding the need to learn global noise imbalance in a single-stage setting.

We used a ResNet-18 backbone for the binary and a ResNet-34 for the multiclass models (He et al. 2016), all pretrained on ImageNet (Russakovsky et al. 2015) as strong, computationally efficient, and widely adopted baselines for spectrogram classification in bioacoustics (Hagiwara et al. 2023), including prior marine mammal detection studies (Bergler et al. 2019; Zhong et al. 2020; Allen et al. 2021; Schall et al. 2024). The binary model was trained with binary cross-entropy loss label smoothing followed by temperature scaling (temperature = 3), and the final fully connected layer was replaced with a single-output linear layer producing a raw logit. Predictions were converted to probabilities using a sigmoid function at inference, and a decision threshold of 0.5 was applied to determine which windows were passed to the multiclass classifier. For the multiclass model, a three-logit output (or four-logit for the version with absence class) from a softmax function was used with categorical cross-entropy, followed by temperature scaling (temperature = 3) to reduce overconfidence. Additional details in Supporting Information.

All models were trained using the AdamW optimizer (Loshchilov and Hutter 2017), an initial learning rate of 0.01, and a weight decay coefficient of 0.0001. A batch size of 512 was used in all experiments. Optimization was performed using a cosine annealing learning-rate schedule, which gradually decreases the learning rate over training. Models were trained for 10 epochs using weights pretrained on ImageNet, which facilitated rapid convergence as shown in Figure S2. All experiments were conducted on an NVIDIA Tesla V100-SXM2 GPU with 32GB of memory (CUDA 12.8) with mixed-precision training enabled.

3.3.1 | Negative Window Sampling

Because the sound recordings contained substantially more background noise than vocalizations, the full set of extracted windows was highly imbalanced toward the negative class. To manage this imbalance, negative windows were randomly sub-sampled so that they comprised approximately 75% of the dataset. The 1st stage binary and single stage 4-class models were trained with this imbalance. For the second stage, the 3-class and 4-class models were trained on balanced datasets, as they operate only on positive windows passed from the binary model. The 3-class model was trained exclusively on positive windows from the three species, whereas the 4-class model included negative windows randomly sub-sampled from the original negative pool so that the negative class contained a number of examples comparable to each positive class.

3.4 | Active Learning for Domain Adaptation

To enable reliable generalization across new recording environments, such as deployments with different background noise conditions, sound recorder characteristics, or geographic regions, typical of long-term programs, we incorporated an active learning component into the training workflow. We tested this workflow on a new domain dataset: annotated sections of CIBA data from Tuxedni Channel (1.2 h, 1474 annotations) and Johnson River (2.1 h, 78 annotations), not included in the base model training (Table S1).

Following an initial inference to pass by both stages (binary and multi-species class) on newly acquired, unannotated (or lightly annotated) wav data, an expert acoustic reviewer (M. Castellote) inspected a small proportion of the model's predictions. Because these data may differ substantially from the training distribution, model outputs in this stage often exhibit reduced precision and recall. The expert therefore evaluated a subsample of sound windows falling below a predefined confidence threshold and manually corrected them, in our case we used a confidence equal or smaller than 0.90 for the binary model and 0.98 for the multiclass model (Figure S1). A small subset of top confidence predictions was also evaluated to confirm absence of any major domain-shift noise artifacts on the novel data. Low confidence predictions for Tuxedni Channel and Johnson River were manually verified and corrected when necessary, whereas high confidence predictions proved to be reliable labels (Table S2). In the first active learning round, the curated data used to fine-tune both the binary and multiclass models was composed by 2190 windows for Tuxedni Channel and 3725 for Johnson River. After retraining, Johnson River performance was satisfactory after one round, whereas Tuxedni Channel still showed elevated false positives, motivating a second round that increased the reviewed Tuxedni set to 3215 windows and further improved performance.

This review–refine cycle can be repeated until the fine-tuned model achieves performance suitable for operational use. This process is a critical component of the deep learning workflow, as it ensures the accuracy, reliability, and generalizability of model outputs (Christin et al. 2019; Quinn et al. 2021). In the context of this study, verification was especially important for confirming the effectiveness of domain adaptation and assessing whether domain shifting had been successfully achieved without inducing bias.

3.5 | Comparison to Previous CIBA Pipeline Results

To evaluate performance and resistance to domain shift, the fine-tuned model was benchmarked against the results from the CIBA data analysis pipeline for the Johnson River (unpublished data; methods following Castellote et al. 2024) and Tuxedni Channel datasets (Castellote et al. 2024; Lammers et al. 2013). These datasets were selected for the comparison because the three species can be present in this region of the CIB critical habitat, an area facing potential increases in anthropogenic activity. Furthermore, because the data were collected over a decade apart using disparate recording systems, they present a significant challenge regarding domain shift. In previous studies, cetacean presence was calculated as detection positive hours (any hour with at least one species detection, DPH) for the three species by integrating echolocation detections from logger datasets with call and whistle detections from sound recordings from the same mooring package. To enable a direct comparison of DPH results between the previous study and the current work, we combine the echolocation logger data results from the past study with the new DNN pipeline results. This ensures that any observed differences in DPH are solely attributable to the performance of our new DNN data analysis pipeline.

Figure 2 summarizes the end-to-end 3-step data processing, model refinement, and testing workflow used in this study, including annotation expansion, multi-stage classification, and active learning.

4 | Results

We present results of improving annotations by adding new detections via the CLAP model, followed by the performance of the base model, the active learning improvement on data from Johnson River and Tuxedni Channel, and finally the comparison of obtained results between this workflow and results produced by the previous CIBA data analysis pipeline using heldout data from Johnson River (not previously seen by the model).

4.1 | CLAP-Based Annotations

Table 3 shows the highest performing prompts that elicited new annotations in the dataset. Prompt performance was highly sensitive to small linguistic variations, including the use of singular versus plural forms and the inclusion of adjectives. Negative prompts ("noise, wind, waves") were kept fixed after testing showed they produced fewer incorrect selections. Highest performing prompts applied to the 3 months-long continuous recording datasets from Chinitna and Iniskin bays (Castellote, Stocker, and Brewer 2020) yielded 1266 killer whale and 1322 humpback whale annotations.

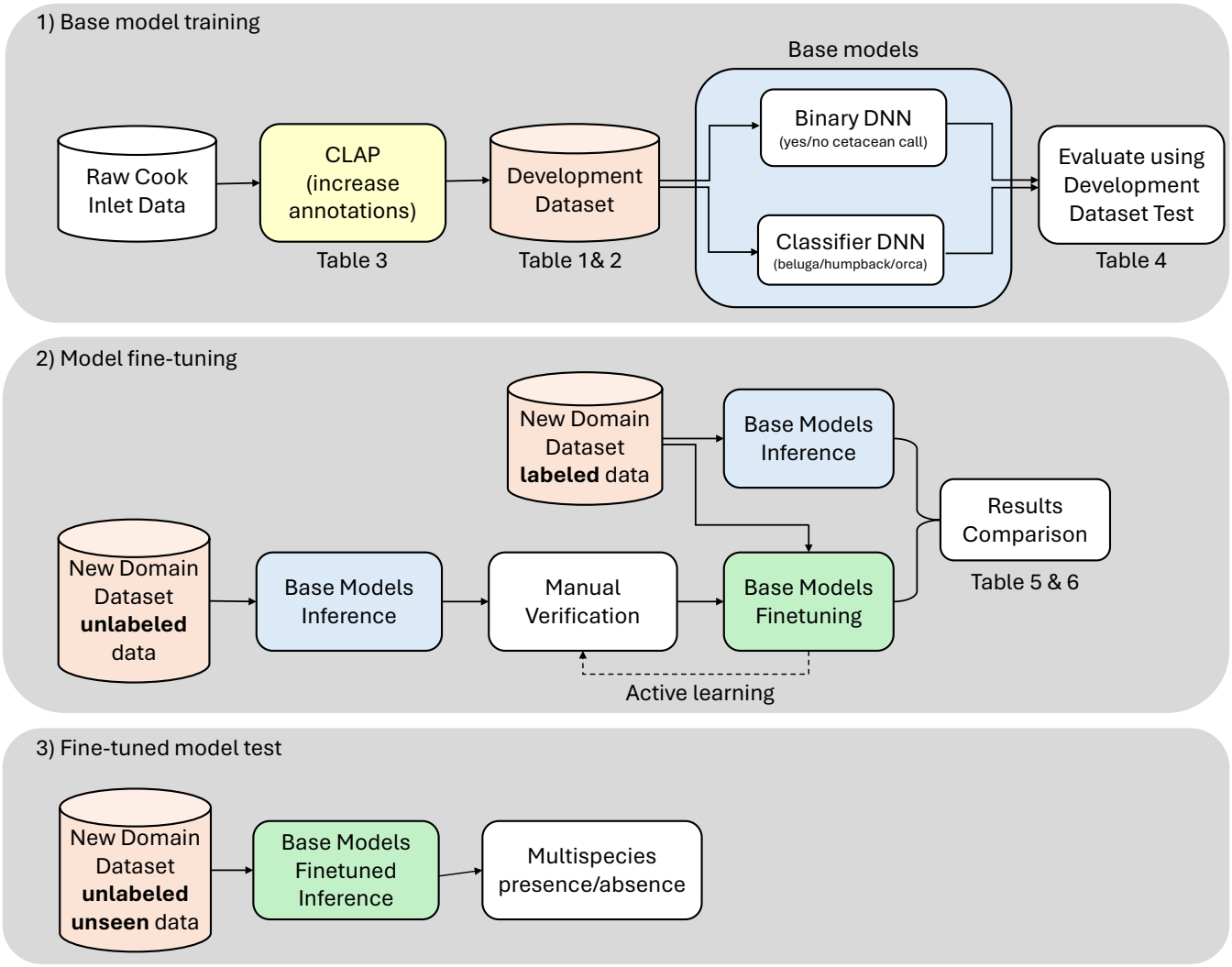

FIGURE 2 | End-to-end 3-step workflow used in this study for (1) base model training, (2) model fine-tuning through active learning, and (3) fine-tuned model testing. CLAP, contrastive language-audio pretraining; DNN, deep neural network. New Domain Dataset refers to Johnson River and Tuxedni Channel CIBA datasets.

4.2 | Base Model Performance

Table 4 summarizes the performance of the full models across the four classes (No Whale, Humpback, Killer whale, and Beluga). Although the single-stage and two-stage approaches demonstrated comparable overall performance, the binary+3-class configuration yielded slightly higher beluga recall while maintaining similar precision across classes, which is our primary interest.

4.3 | Active Learning Results

Active learning improved base model performance across both datasets, though the effect differed by class and location (Tables 5 and 6). For the "No whale" class, precision, recall, and F1 score increased in Tuxedni Channel and remained consistently high in Johnson River, with only minor changes. Beluga performance improved in Tuxedni Channel with gains in all metrics, but detections in Johnson River were negligible due to scarcity. For "Killer whale", Johnson River showed substantial gains in recall and F1 after active learning, while performance in Tuxedni Channel remained poor with no clear improvement. Neither dataset contained humpback whale vocalizations, so we could not assess their performance.

4.4 | Comparison With Older Johnson River and Tuxedni Channel Results

To ensure accuracy and eliminate false positives, all predictions from the fine-tuned model were manually verified. Cetacean presence was quantified as DPH to facilitate a direct comparison of this study's results with results reported by Lammers et al. (2013) for Tuxedni Channel and Castellote et al. (2024) for Johnson River.

For Johnson River, Figure 3 shows this comparison along a timeline from September 2022 to July 2023. For belugas, the current model identified three new detections in addition to the two previously reported; all five detections were brief, consisting of a single positive hour between September and November. Killer whale

TABLE 3 | Contrastive language-audio pretraining (CLAP) positive language prompt testing and fine-tuning used to obtain the highest number of correct segments with humpback and killer whale sounds and the lowest number of incorrect segments.

Language prompts targeting humpback whale sounds			
	Correct	**Incorrect**	**Correct/incorrect**
Direct semantic prompts			
Humpback whale calls	9	3	3.0
Humpback whale vocalizations	12	3	4.0
Whale call/s (same score singular or plural)	11	4	2.8
Whale vocalizations	12	3	4.0
Indirect semantic prompts			
Tonal signal/s	8	4	2.0
Tones	15	6	2.5
Distant tones	21	5	4.2
Distant tones no higher than 3 kHz	23	4	5.8
Distant tones no higher than 3 kHz in frequency	18	9	2.0
Distant tones no higher than 2 kHz in frequency	18	9	2.0
Distant tones with very low pitch	23	5	4.6
Distant tones in sequences	18	3	6.0
Low hums	12	26	0.5
Downsweeps	19	21	0.9
Distant trumpets	2	16	0.1
Very faint bird chirps	42	66	0.6
Bird chirps	40	12	3.3
Bird chirps shorter than 0.5 s	43	69	0.6
Very short bird chirps	40	68	0.6
Bird chirps only between 3 and 4 kHz	45	69	0.7
Bird chirps only between 3 and 4 kHz in frequency	49	75	0.7
Trill	12	4	3.0
Faint trill	17	4	4.3
Distant bird trills	42	60	0.7
Repetitive chirp	6	3	2.0
Repetitive bird	27	9	3.0
Repetitive bird whistle	14	16	0.9
Stereotypic bird	16	9	1.8
Stereotypic singing bird	19	9	2.1
Voice	4	1	4.0
Distant shouts	26	6	4.3
Unrecognizable shouts	30	2	15.0
Unrecognizable faint shouts	43	68	0.6
Brief female cries	45	6	7.5

(Continues)

TABLE 3 | (Continued)

Language prompts targeting humpback whale sounds	Correct	Incorrect	Correct/incorrect
Baby cries	45	3	15.0
Brief baby cries	46	3	15.3
Brief baby cries and distant bird trills	47	7	6.7
Brief baby cries and chirps	49	9	5.4
Brief baby cries and loud chirps	**49**	**2**	**24.5**
Language prompts targeting killer whale sounds			
Direct semantic prompts			
Orca calls	10	21	0.5
Orca whistles	13	11	1.2
Killer whale calls	8	29	0.3
Killer whale whistles	15	20	0.8
Dolphin whistles	10	13	0.8
Dolphin sounds	7	16	0.4
Orca whistles and other dolphins	18	12	1.5
Orca whistles and other sounds from dolphins	19	14	1.4
Whistles and sounds from orca and other dolphins	**22**	**9**	**2.4**
Indirect semantic prompts			
Brief baby cries and loud chirps	16	23	0.7
Short baby cries, loud chirps	19	24	0.8
Faint short baby cries, loud chirps	15	22	0.7
Faint baby cries, loud chirps	14	21	0.7
Faint baby cries, distant whistles	17	22	0.8
Chirps, distant whistles	9	41	0.2
Distant whistles	18	36	0.5
Short tones	3	14	0.2
Faint sweeps	9	39	0.2
Bird calls	2	10	0.2

Note: Prompts included direct semantic descriptions of whale sounds and indirect semantic descriptions of acoustic analogues (e.g., "repetitive chirp"). Top-performing prompts highlighted in bold.

presence was more extensive; the model identified 18 new encounters and successfully recovered all but one of the previous detections. While most killer whale encounters consisted of a single positive hour, longer durations of 2, 3, and 7 h were also recorded. Killer whales were detected in every month except December, with the highest prevalence occurring from September through November.

For Tuxedni Channel, Figure 4 shows the comparison along a timeline from December 2011 to April 2012. The current model identified all except 3 DPH from the previously reported, and 17 new DPH in January 2012, 13 DPH in February 2012, and 32 DPH in March.

5 | Discussion

This study aimed to improve upon an established, long-running operational pipeline by NOAA Fisheries tailored to a specific endangered population and acoustic environment: the endangered Cook Inlet beluga population. The challenges addressed here, rare beluga vocalizations, interference from other species, complex background noise in extreme tidal environments, and the use of different recording platforms over a decade of monitoring, are not represented in archival benchmark datasets, and the appropriate performance reference is the prior iteration of the CIBA analytical pipeline. Relative to that baseline, this new pipeline identified multiple beluga and

TABLE 4 | Comparative performance of single-stage (4-class) and base model's sequential two-stage architectures (binary+3-class; binary +4-class) across cetacean vocalization classes.

	Precision			Recall			F1		
Class	4-class	Binary+3-class	Binary+4-class	4-class	Binary+3-class	Binary+4-class	4-class	Binary+3-class	Binary+4-class
No whale	0.952	0.964	0.961	0.982	0.969	0.975	0.967	0.967	0.968
Humpback	0.939	0.863	0.898	0.653	0.678	0.724	0.770	0.759	0.802
Killer whale	0.868	0.811	0.872	0.871	0.901	0.899	0.870	0.853	0.885
Beluga	0.932	0.927	0.932	0.944	0.963	0.959	0.938	0.945	0.945

TABLE 5 | Tuxedni channel results from base model before and after active learning.

Model	Class	FN	FP	TP	TN	Precision	Recall	F1
Base (before fine-tuning)	No whale	29	19	255	87	0.93	0.90	0.91
	Humpback	0	2	0	388	—	—	—
	Killer whale	45	3	3	339	—	—	—
	Beluga	4	54	54	278	0.50	0.93	0.65
Fine-tuned	No whale	20	16	264	90	**0.93**	**0.95**	**0.94**
	Humpback	0	0	0	390	—	—	—
	Killer whale	48	1	0	341	—	—	—
	Beluga	3	54	55	278	**0.56**	**0.95**	**0.70**

Note: The best results are bold.

TABLE 6 | Johnson River results from base model before and after active learning.

Model	Class	FN	FP	TP	TN	Precision	Recall	F1
Base (before fine-tuning)	No whale	9	36	1054	93	0.97	**0.99**	**0.98**
	Humpback	0	0	0	1192	—	—	—
	Killer whale	125	0	1	1066	**1.00**	0.01	0.02
	Beluga	1	99	2	1090	0.02	0.67	0.04
Fine-tuned	No whale	48	16	1015	113	**0.98**	0.95	0.97
	Humpback	0	0	0	1192	—	—	—
	Killer whale	24	46	102	1020	0.69	**0.81**	**0.74**
	Beluga	3	13	0	1176	—	—	—

Note: The best results are bold.

killer whale vocalizations that were not detected using the previous approach. Importantly, this pipeline is designed so that it can be adapted to other endangered species, locations, and monitoring equipment. The dual-stage architecture, combined with text-guided audio representations (CLAP), improved the model's ability to capture rare acoustic signals present in the data, while the incorporation of an active learning framework enables adaptation of the pipeline to different acoustic environments. An important distinction is that the present contribution centers on workflow adaptation rather than backbone innovation. Recent studies have benchmarked general-purpose audio foundation models and specialized bioacoustic encoders for animal sound classification, often showing advantages over image-pretrained CNNs in some settings (Ghani et al. 2023; Schwinger et al. 2026; Miron et al. 2025). Our results should therefore be interpreted as demonstrating that the active learning framework is effective when paired with a standard ResNet baseline. A natural and immediate extension is to replace the current backbone with domain-specific pretrained encoders and evaluate whether their stronger initial representations further reduce annotation effort or improve transfer to new recording domains.

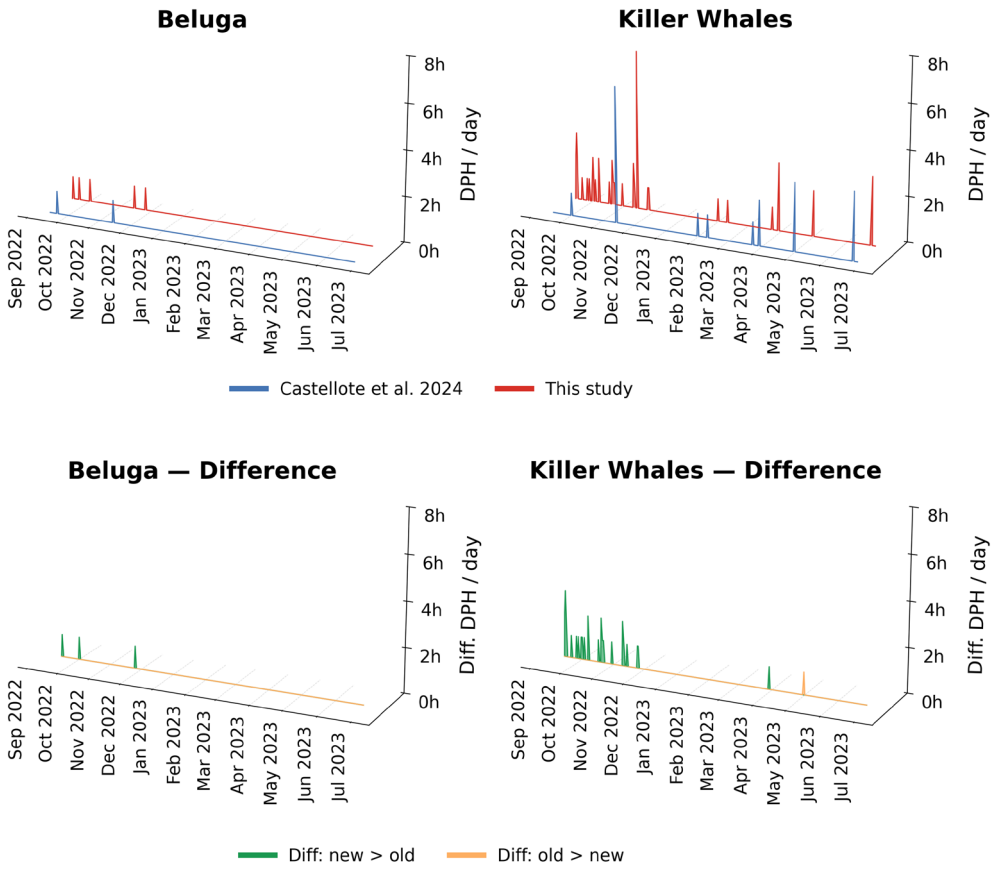

FIGURE 3 | Daily detection positive hours (DPH) for beluga and killer whales in the Johnson River dataset: a comparison between Johnson River dataset (unpublished; following methods from Castellote et al. 2024) and the current study. Lower panels denote the difference in detections between the two timeseries.

The field of PAM for underwater applications has quickly evolved in the last few years in its technical capacity. A larger selection of more power efficient recorders with increased memory capacity and faster sampling rate capabilities is now available (Pavan et al. 2022). New PAM platforms are now available beyond moorings or landers, such as drifters, gliders, and AUVs (Cauchy et al. 2023). This has led to an increase in the use of PAM in ecological studies, with the consequent drastic increase in the volume of data collected. However, for the last few years, the analytical capacity in this field has remained relatively quiescent except for the notable development of AI applications. The fast progression of computer processing power has catapulted the development of machine learning algorithms which now very efficiently automate a large portion of the passive acoustic data analysis process (Kershenbaum et al. 2025). Bioacousticians are now challenged by the pace of progress in computer science, and the constant evolution of signal processing pipelines. The spatial expansion of the CIBA program prompted the need to improve its analytical methods in such a way.

Given the conservation focus on the endangered Cook Inlet beluga, multi-species modeling is treated as a supporting mechanism to improve beluga detection reliability, with performance for other species interpreted in that context. The jump from one species to multi-species detection prompted the exploration of a dual stage model pipeline, which in turn allows tuning the system's performance at two different levels of the decision tree. The binary model threshold can now be set low to avoid type II error (false negative), which is the preferred strategy in conservation biology of endangered species, where sensitivity for rare encounters is valuable. This dual stage model is particularly beneficial when data is highly imbalanced, such as PAM data for marine mammals; most sound files contain just noise. Multi-stage classification is used to handle difficult discrimination tasks by breaking the problem into stages (Kershenbaum et al. 2025). In our design, we perform signal detection first, then species classification on the spectrograms we believe have a signal, making learning more robust under class imbalance. There might also be a benefit for generalization and domain shift, as the challenge is now broken into two distinct tasks with different goals: the initial one to distinguish between all noise and cetacean signals, where fundamental features of cetacean calls and whistles are likely sufficient to determine its presence during inference, compared to more specialized learned features needed for discriminating between species' signals.

The use of CLAP to expand the annotation sample size for humpback and killer whales proved beneficial, as prior CIBA data and analyses had been largely focused on beluga detection. Identifying effective text prompts and reviewing model outputs was much more efficient than manually searching for these signals across large datasets. However, our experience with prompt design indicates that variations in concept phrasing, such as the use of singular versus plural forms or the positioning

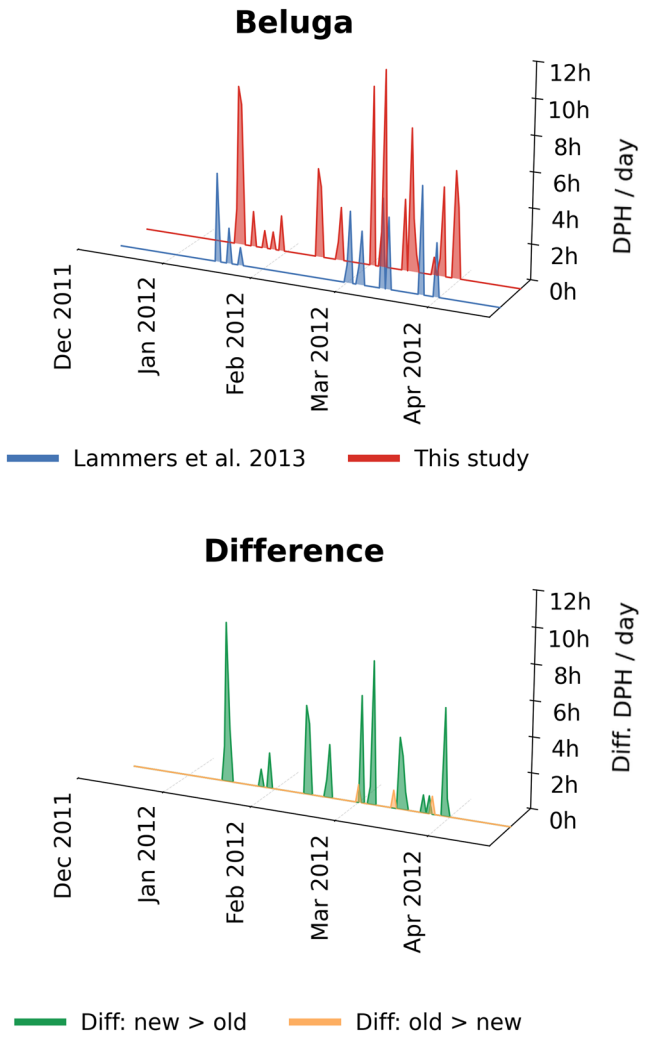

FIGURE 4 | Daily detection positive hours (DPH) for beluga and killer whales in the Tuxedni Channel dataset: A comparison between Lammers et al. (2013) and the current study. Lower panel denotes the difference in detections between the two timeseries.

of adjectives before or after the noun, affected retrieval performance in ways that were difficult to predict or interpret. This is not uncommon on other zero-shot contrastive audio-language models (Yuksekgonul et al. 2023) or even LLMs (Salinas and Morstatter 2024) and suggests improvements in training the encoders would be advantageous. Indirect, semantically broad prompts using common-language descriptors were most effective at identifying humpback and killer whale calls and whistles. This result highlights the need for training datasets that incorporate more technical acoustic descriptors. Such improvements would be particularly valuable for targeted searches of species-specific signals, where attributes such as frequency range or call duration can readily discriminate among species. Using textual descriptions that prioritize acoustic features of sound events to disambiguate between classes as suggested by Olvera et al. (2024) did not help in our case. This prompt sensitivity has direct implications for the reproducibility and transferability of CLAP-based annotation strategies, and practitioners applying this approach to new species or acoustic domains should adopt a structured prompt evaluation workflow. We recommend evaluating candidate prompts on a small, manually verified subset of the target dataset before committing to large-scale annotation, even a few hours of reviewed recordings can reveal which phrasings retrieve the target signal most reliably in the specific acoustic environment of interest. Where a single reliable prompt cannot be identified, prompt ensembling (averaging the text embeddings of multiple semantically equivalent phrasings of the same target sound) can provide a more stable and reproducible classification signal than any individual formulation (Olvera et al. 2024). Importantly, all prompts tested and the selection procedure used should be fully reported in the methods to ensure reproducibility. In the CIBA program, the prompt sensitivity we encountered reinforces the value of the active learning loop as a downstream correction mechanism: rather than depending on any single zero-shot prompt to produce clean annotations, we use CLAP as an efficient first-pass label generator whose errors are corrected iteratively through targeted manual review and model retraining. This positions CLAP as a scalable annotation accelerator rather than a standalone classifier, which we argue is the appropriate operational role for prompt-based audio-language models in long-term conservation monitoring programs at their current stage of development. Future work should systematically compare text- and audio-based retrieval strategies using contrastive audio-language models under real PAM deployment conditions, where recording context and soundscape complexity may differentially affect retrieval performance across query modalities.

Inference by the two-stage base model on unseen data showed mixed results. The model correctly classified beluga signals, but its first stage failed to detect a small number of them. Some true negatives were classed as beluga, and most detected killer whale signals were incorrectly classed as beluga. These results suggest that the base model did not generalize as well as expected; however, while initial models provide a starting point, they often perform suboptimally, whereas active learning typically yields substantially better performance for the same number of annotations (Kath et al. 2024).

The review of predictions as part of the active learning process improved recall and precision for the target species as well as a reduction in false positives and provided a good insight into the nuances the model might have encountered. For example, in the Tuxedni Channel data, we often encountered what we believe are calls/wing flaps from a seaduck species, likely large rafts of wintering long-tail ducks (*Clangula hyemalis*), harlequin ducks (*Histrionicus histrionicus*), or white-winged and surf scoters (*Melanitta deglandi* and *M. perspicillata*) at the surface very close to the mooring location, commonly classed as beluga prior to fine-tuning. Upon inspection of the training data annotations, we found that beluga calls were often associated with periods of intense marine bird calling at this location. The western coast of lower Cook Inlet is an important wintering area for sea ducks which gather in large flocks (ABR 2011; Angler 1995; Renner et al. 2017). We hypothesize that, during training, the model extracted features related to seaduck calls in inference windows labeled as beluga. As a result, it predicted beluga with high confidence in sound recordings containing seaduck calls/wing flaps with a high signal-to-noise ratio, even when no actual beluga calls or whistles were present (Figure 5).

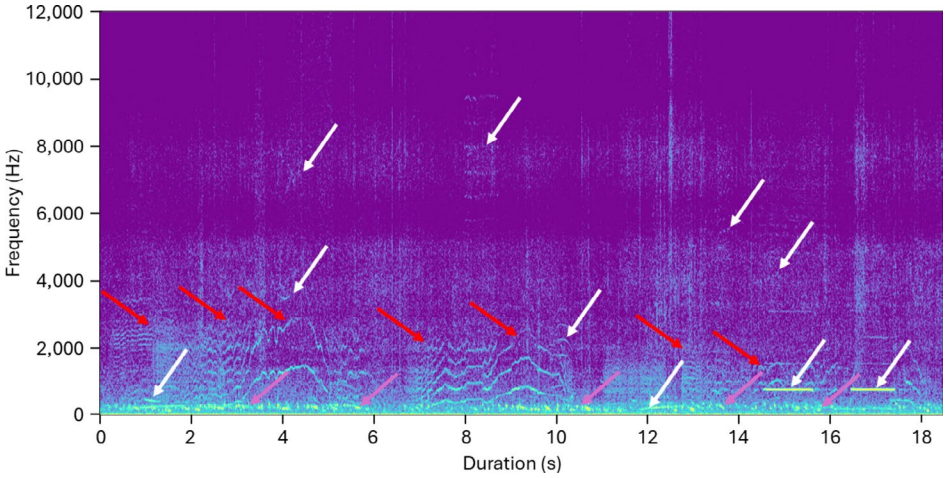

FIGURE 5 | Spectrogram of a 18-s recording (0–12 kHz) featuring concurrent sea ice frequency modulated tonal noise from pressure stress (red arrows), beluga vocalizations (white), and seaduck calls/wing flaps (pink). Processing resolution: 23.8 Hz and 10.7 ms.

This finding raises interesting ecological questions, as the co-occurrence could be driven by shared prey preferences. Seaducks and belugas both forage on marine invertebrates like crustaceans and annelids (Ouellet et al. 2013; Quakenbush et al. 2015). In the only study to look at the diet of Cook Inlet belugas, Quakenbush et al. (2015) found that of the 18 stomachs that contained prey, 50% contained up to eight species of invertebrates, predominantly shrimp, followed by polychaetes and amphipods. Beluga may be passively cued on these seabird sounds to locate areas of concentrated prey, a behavior supported by evidence of odontocetes listening for cues related to prey concentration (Lawrence et al. 2016; Thode et al. 2015). This association occurred primarily during winter, December to March, a time when little is known about the prey preferences of this endangered beluga population. However, the potential for increased anthropogenic noise in this acoustically pristine area may have the potential to compromise this acoustic association. If there is auditory masking, these subtle biological cues may disrupt the belugas' ability to locate concentrated prey. Further research should focus on understanding the specific prey driving these species to lower Cook Inlet to better assess the risks posed by industrial development. Another notable nuance observed during the active learning process was the occurrence of high-confidence false positives (2.4%) for beluga vocalizations in the presence of sea ice pressure noise. Despite the absence of beluga whistles or calls, the model was misled by high-amplitude, frequency-modulated tonal signals generated by ice stress (Figure 5). Such noise is characteristic of Arctic and subarctic winter soundscapes and has been documented for its structural similarities to marine mammal vocalizations (Kinda et al. 2015).

Model performance improved substantially after a single iteration of active learning across both datasets. In Tuxedni Channel data, confusion between beluga vocalizations and nonbiological signals was reduced, leading to a decrease in beluga false positives and improved performance for the 'No Whale' class. Beluga detection benefited from this refinement, with recall remaining high (0.95) and precision increasing from 0.50 to 0.56 (Table 5). In Johnson River data, active learning led to a marked improvement in killer whale detection: signals that were previously misclassified as beluga were correctly identified, increasing killer whale recall from 0.01 to 0.81 and F1 from 0.02 to 0.74 (Table 6). Despite these gains, the first-stage noise–whale classifier continued to miss a portion of true whale detections, limiting overall recall for less prevalent classes. Overall, active learning substantially improved detection performance for the dominant species at each site while highlighting remaining challenges in detecting less prevalent species.

Although the fine-tuned model demonstrated performance more consistent with the requirements of a multi-species framework, the advantages of the integrated classifier were attenuated by inter-species confusion between beluga and killer whale vocalizations. Furthermore, the model exhibited an unexpected sensitivity to marine bird signals, leading to an inflation of false positives for belugas. To mitigate these limitations, future iterations could incorporate a 'noise' class to the classifier for nontarget signals, such as marine bird calls/wing flaps or sea ice and employ these signals as hard negatives during the training of the base model. Despite these limitations, the comparative analysis of daily DPH for belugas and killer whales in the Johnson River and Tuxedni Channel remains highly significant. The fine-tuned model identified three previously undetected beluga encounters in Johnson River in addition to the two known detection periods, resulting in a 2.5-fold increase in recorded species presence, and five previously undetected beluga encounters in Tuxedni Channel, resulting in a 1.5-fold increase in recorded species presence. Improved detection rates and resilience to hardware-driven domain shift highlight the efficacy of the updated analysis pipeline. These findings will help inform management decisions and government permitting in this area, which is acoustically a relatively pristine winter foraging ground for Cook Inlet belugas (Castellote et al. 2024). Furthermore, the doubling of killer whale detections strengthens the evidence for potential predation pressure on this beluga population. Although ecotype could not be confirmed in many of the encounters due to the brevity and scarcity of calls, this characteristic of a low calling rate is typical of the mammal eating ecotype (Deecke et al. 2005). Heightened predation risk intensifies acoustically cryptic behavior in belugas as a defensive mechanism. This suggests that beluga presence and habitat utilization, particularly in

periods of elevated killer whale activity, may be underestimated by passive acoustic data unless sampling is very focalized, warranting caution in drawing conclusions about population distribution in larger scales. While the sample size is modest, this case study underscores the necessity of implementing the most current methodologies in long-term PAM programs and highlights the continuous need for upgrading analytical frameworks to ensure accurate ecological assessments.

6 | Conclusions

This study presents an adaptive, open-source deep learning framework that advances passive acoustic monitoring of the endangered Cook Inlet beluga whale by improving analytical scalability, domain generalization, and multi-species detection capacity. By transitioning from a binary to a two-stage, multi-class workflow, we demonstrate how sensitivity can be prioritized for rare species detection while maintaining control over classification uncertainty in highly imbalanced and acoustically variable datasets.

The integration of zero-shot contrastive audio-language models enabled efficient expansion of training data for additional cetacean species, highlighting both the potential and current limitations of prompt-based annotation strategies. Active learning proved essential for improving performance on unseen data, reducing false positives and revealing soundscape-specific challenges that would have remained obscured in fully automated pipelines. These findings emphasize the importance of iterative human–model interaction for both methodological robustness and ecological interpretability.

Applied to the Cook Inlet beluga critical habitat, this framework enhances the reliability and efficiency of PAM analyses in support of conservation decision-making. More broadly, the approach is highly transferable and provides a practical pathway for modernizing large-scale marine mammal acoustic monitoring as data volumes, species complexity, and management demands continue to grow.

Author Contributions

Pablo Arbelaez: conceptualization, methodology, writing – review and editing, formal analysis. **Lori Polasek:** writing – review and editing. **Rahul Dodhia:** conceptualization, data curation, methodology, project administration, writing – review and editing, formal analysis. **Zhongqi Miao:** writing – review and editing, conceptualization, methodology, formal analysis. **Manuel Castellote:** conceptualization, data curation, project administration, methodology, visualization, writing – original draft, formal analysis. **Juan M. Lavista Ferres:** writing – review and editing. **Daniela Ruiz:** conceptualization, data curation, methodology, writing – review and editing, formal analysis. **Verena Gill:** writing – review and editing.

Acknowledgments

Data collection for the CIBA program was funded by the Alaska Department of Fish and Game's Marine Mammal Program, National Oceanic and Atmospheric Administration (NOAA) Alaska Fisheries Science Center's Marine Mammal Laboratory, Alaska Region of the Bureau of Ocean Energy Management, U.S. Department of the Interior, and the Alaska Region of NOAA Fisheries. The findings and conclusions in this article are those of the authors and do not necessarily represent the views of NOAA Fisheries. Tuxedni Channel 2011-2012 data, analyzed by Marc Lammers (NOAA Fisheries), were provided as part of the CIBA program for our comparative analysis. Marian Blaze (AI for Good Lab) assisted in the formatting and preparation of this paper for its submission. We thank the Associate Editor, Dr. Christine Erbe, and three anonynous reviewers for their careful evaluation of the manuscript and their constructive comments, which helped improved the quality and presentation.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The model and pipeline scripts supporting this study are openly available in https://github.com/microsoft/cookinlet_belugas. Data and annotations are available upon request to the corresponding author.

References

ABR Inc. 2011. Pebble Project Environmental Baseline Document 2004 Through 2008.

Allen, A. N., M. Harvey, L. Harrell, et al. 2021. "A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset." *Frontiers in Marine Science* 8: 607321. https://doi.org/10.3389/fmars.2021.607321.

Angler, B. A., 1995. Estimates of Marine Bird and Sea Otter Abundance in Lower Cook Inlet, Alaska During Summer 1993 and Winter 1994.

Bergler, C., H. Schröter, R. X. Cheng, et al. 2019. "ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning." *Scientific Reports* 9, no. 1: 10997. https://doi.org/10.1038/s41598-019-47335-w.

Bergler, C., S. Q. Smeele, S. A. Tyndel, et al. 2022. "ANIMAL-SPOT Enables Animal-Independent Signal Detection and Classification Using Deep Learning." *Scientific Reports* 12, no. 1: 21966. https://doi.org/10.1038/s41598-022-26429-y.

Brewer, A. M., M. Castellote, A. M. Van Cise, T. Gage, and A. M. Berdahl. 2023. "Communication in Cook Inlet Beluga Whales: Describing the Vocal Repertoire and Masking of Calls by Commercial Ship Noise." *Journal of the Acoustical Society of America* 154, no. 5: 3487–3505. https://doi.org/10.1121/10.0022516.

Brewer, A. M., A. M. Van Cise, C. Garner, et al. 2026. "Cook Inlet Beluga Whale Calling Varies by Group Characteristics, Behavior, and Tidal State." *Behavioral Ecology and Sociobiology* 80: 62. https://doi.org/10.1007/s00265-026-03740-6.

Burell, D., and J. Matthews. 1974. "Turbid Outwash Fiords." In *Coastal Ecological Systems of the United States*, edited by H. Odum, B. Copeland, and E. McMahan, vol. 3, 12–17. Conservation Foundation.

Castellote, M., V. A. Gill, C. D. Garner, et al. 2024. "Using Passive Acoustics to Identify a Quiet Winter Foraging Refuge for an Endangered Beluga Whale Population in Alaska." *Frontiers in Marine Science* 11: 1393380. https://doi.org/10.3389/FMARS.2024.1393380/TEXT.

Castellote, M., R. J. Small, M. O. Lammers, et al. 2020. "Seasonal Distribution and Foraging Occurrence of Cook Inlet Beluga Whales Based on Passive Acoustic Monitoring." *Endangered Species Research* 41: 225–243. https://doi.org/10.3354/esr01023.

Castellote, M., R. J. Small, M. O. Lammers, J. J. Jenniges, J. Mondragon, and S. Atkinson. 2016. "Dual Instrument Passive Acoustic Monitoring

of Belugas in Cook Inlet, Alaska." *Journal of the Acoustical Society of America* 139, no. 5: 2697–2707. https://doi.org/10.1121/1.4947427.

Castellote, M., M. Stocker, and A. Brewer. 2020. Passive Acoustic Monitoring of Cetaceans and Noise During Hilcorp 3D Seismic Survey in Lower Cook Inlet, AK. Accessed November 23, 2025. https://inletkeeper.org/wp-content/uploads/2024/04/final-report-10-11-20-preview.pdf.

Cauchy, P., K. J. Heywood, N. D. Merchant, D. Risch, B. Y. Queste, and P. Testor. 2023. "Gliders for Passive Acoustic Monitoring of the Oceanic Environment." *Frontiers in Remote Sensing* 4: 1–13. https://doi.org/10.3389/frsen.2023.1106533.

Cerchio, S., and M. Dahlheim. 2001. "Variation in Feeding Vocalizations of Humpback Whales *Megaptera novaeangliae* From Southeast Alaska." *Bioacoustics* 11, no. 4: 277–295. https://doi.org/10.1080/09524622.2001.9753468.

Christin, S., É. Hervet, and N. Lecomte. 2019. "Applications for Deep Learning in Ecology." *Methods in Ecology and Evolution* 10, no. 10: 1632–1644. https://doi.org/10.1111/2041-210X.13256.

Clarfeld, L., K. Gieder, R. Abrams, et al. 2025. *Two-stage models improve machine learning classifiers in wildlife research: A case study in identifying false positive detections of Ruffed Grouse: U.S. Geological Survey data release*.

Deecke, V. B., J. K. B. Ford, and P. J. B. Slater. 2005. "The Vocal Behaviour of Mammal-Eating Killer Whales: Communicating With Costly Calls." *Animal Behaviour* 69, no. 2: 395–405. https://doi.org/10.1016/j.anbehav.2004.04.014.

D'Vincent, C. G., R. M. Nilson, and R. E. Hanna. 1985. "Vocalization and Coordinated Feeding Behavior of the Humpback Whale in Southeastern Alaska." *Scientific Reports of the Whales Research Institute* 36: 41–47.

Fournet, M. E. H., L. P. Matthews, C. M. Gabriele, D. K. Mellinger, and H. Klinck. 2018. "Source Levels of Foraging Humpback Whale Calls." *Journal of the Acoustical Society of America* 143, no. 2: EL105–EL111. https://doi.org/10.1121/1.5023599.

Ghani, B., T. Denton, S. Kahl, and H. Klinck. 2023. "Global Birdsong Embeddings Enable Superior Transfer Learning for Bioacoustic Classification." *Scientific Reports* 13, no. 1: 22876.

Goetz, K., K. Shelden, C. Sims, J. Waite, and P. Wade. 2023. *Abundance and Trend of Belugas (Delphinapterus leucas) in Cook Inlet, Alaska, June 2021 and June 2022*. National Marine Fisheries Service.

Goetz, K. T., R. A. Montgomery, J. M. Ver Hoef, R. C. Hobbs, and D. S. Johnson. 2012. "Identifying Essential Summer Habitat of the Endangered Beluga Whale *Delphinapterus leucas* in Cook Inlet, Alaska." *Endangered Species Research* 16, no. 2: 135–147. https://doi.org/10.3354/esr00394.

Hagiwara, M., B. Hoffman, J. Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian. 2023. "BEANS: The Benchmark of Animal Sounds." In *In 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90.

Himes Boor, G. K., T. L. McGuire, A. J. Warlick, et al. 2022. "Estimating Reproductive and Juvenile Survival Rates When Offspring Ages Are Uncertain: A Novel Multievent Mark-Resight Model With Beluga Whale Case Study." *Methods in Ecology and Evolution* 14: 631–642. https://doi.org/10.1111/2041-210x.14032.

Hobbs, R., K. Laidre, B. Mahoney, and M. Eagleton. 2005. "Movements and Area Use of Belugas, *Delphinapterus leucas*, in a Subarctic Alaskan Estuary." *Arctic* 58, no. 4: 331–340. https://doi.org/10.14430/arctic447.

Hobbs, R. C., D. J. Rugh, and D. P. DeMaster. 2000. Abundance of Belugas, *Delphinapterus leucas*, in Cook Inlet, Alaska, 1994–2000. http://hdl.handle.net/1834/26383.

Kath, H., P. P. Serafini, I. B. Campos, T. S. Gouvêa, and D. Sonntag. 2024. "Leveraging Transfer Learning and Active Learning for Data Annotation in Passive Acoustic Monitoring of Wildlife." *Ecological Informatics* 82: 102710. https://doi.org/10.1016/j.ecoinf.2024.102710.

Kershenbaum, A., Ç. Akçay, L. Babu-Saheer, et al. 2025. "Automatic Detection for Bioacoustic Research: A Practical Guide From and for Biologists and Computer Scientists." *Biological Reviews* 100, no. 2: 620–646. https://doi.org/10.1111/brv.13155.

Kinda, G. B., Y. Simard, C. Gervaise, J. I. Mars, and L. Fortier. 2015. "Arctic Underwater Noise Transients From Sea Ice Deformation: Characteristics, Annual Time Series, and Forcing in Beaufort Sea." *Journal of the Acoustical Society of America* 138, no. 4: 2034–2045.

Lammers, M. O., M. Castellote, R. J. Small, et al. 2013. "Passive Acoustic Monitoring of Cook Inlet Beluga Whales (*Delphinapterus leucas*)." *Journal of the Acoustical Society of America* 134, no. 3: 2497–2504. https://doi.org/10.1121/1.4816575.

Lanik, A., J. Rogers, and K. RD Jr. 2021. Lake Clark National Park and Preserve Geologic Resources Inventory Report. https://irma.nps.gov/datastore/downloadfile/666832.

Lawrence, J. M., E. Armstrong, J. Gordon, S. M. Lusseau, and P. G. Fernandes. 2016. "Passive and Active, Predator and Prey: Using Acoustics to Study Interactions Between Cetaceans and Forage Fish." *ICES Journal of Marine Science* 73, no. 8: 2075–2084. https://doi.org/10.1093/icesjms/fsw013.

Loshchilov, I., and F. Hutter. 2017. "Decoupled Weight Decay Regularization." In *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019*. ICLR.

Lostanlen, V., J. Salamon, M. Cartwright, et al. 2019. "Per-Channel Energy Normalization: Why and How." *IEEE Signal Processing Letters* 26, no. 1: 39–43. https://doi.org/10.1109/LSP.2018.2878620.

McGuire, T. L., G. K. Himes Boor, J. R. McClung, et al. 2020. "Distribution and Habitat Use by Endangered Cook Inlet Beluga Whales: Patterns Observed During a Photo-Identification Study, 2005–2017." *Aquatic Conservation: Marine and Freshwater Ecosystems* 30, no. 12: 2402–2427. https://doi.org/10.1002/aqc.3378.

McGuire, T. L., K. E. W. Shelden, G. K. Himes Boor, et al. 2020. "Patterns of Mortality in Endangered Cook Inlet Beluga Whales: Insights From Pairing a Long-Term Photo-Identification Study With Stranding Records." *Marine Mammal Science* 37, no. 2: 492–511. https://doi.org/10.1111/mms.12766.

McGuire, T. L., A. D. Stephens, J. R. McClung, et al. 2020. "Reproductive Natural History of Endangered Cook Inlet Beluga Whales: Insights From a Long-Term Photo-Identification Study." *Polar Biology* 43, no. 11: 1851–1871. https://doi.org/10.1007/s00300-020-02750-y.

McGuire, T. L., A. D. Stephens, J. R. McClung, et al. 2021. "Anthropogenic Scarring in Long-Term Photo-Identification Records of Cook Inlet Beluga Whales, *Delphinapterus leucas*." *Marine Fisheries Review* 82, no. 3–4: 20–40. https://doi.org/10.7755/mfr.82.3-4.3.

McHuron, E. A., M. Castellote, G. K. Himes Boor, et al. 2023. "Modeling the Impacts of a Changing and Disturbed Environment on an Endangered Beluga Whale Population." *Ecological Modelling* 483: 110417. https://doi.org/10.1016/j.ecolmodel.2023.110417.

Miron, M., D. Robinson, M. Alizadeh, et al. 2025. What Matters for Bioacoustic Encoding. https://doi.org/10.48550/arXiv.2508.11845.

Molnia, B., and R. S. Williams. 2008. *Glaciers of North America Glaciers of Alaska*, edited by B. F. Molnia, R. S. Williams Jr., and J. G. Ferrigno. United States Government Printing Office.

Myers, H. J., D. W. Olsen, B. H. Konar, et al. 2025. "Killer Whale Call Detection Rates Vary Among Subspecies and Populations in the North Pacific." *Scientific Reports* 15, no. 1: 21072. https://doi.org/10.1038/s41598-025-06041-6.

NMFS. 2008. *Conservation Plan for the Cook Inlet Beluga Whale (Delphinapterus leucas)*. National Marine Fisheries Service.

NMFS. 2016. *Recovery Plan for the Cook Inlet Beluga Whale (Delphinapterus leucas)*, 284. National Marine Fisheries Service Alaska Region.

Olvera, M., P. Stamatiadis, and S. Essid. 2024. A Sound Description: Exploring Prompt Templates and Class Descriptions to Enhance Zero-Shot Audio Classification. Detection and Classification of Acoustic Scenes and Events. http://arxiv.org/abs/2409.13676.

Ouellet, J. F., C. Vanpé, and M. Guillemette. 2013. "The Body Size-Dependent Diet Composition of North American Sea Ducks in Winter." *PLoS One* 8, no. 6: e65667. https://doi.org/10.1371/JOURNAL.PONE.0065667.

Pavan, G., G. Budney, H. Klinck, H. Glotin, D. Clink, and J. Thomas. 2022. "History of Sound Recording and Analysis Equipment." In *Exploring Animal Behavior Through Sound: Volume 1: Methods*, edited by J. A. E. Christine, 1–36. Springer Nature. https://doi.org/10.1007/978-3-030-97540-1_1.

Quakenbush, L. T., R. S. Suydam, A. L. Bryan, L. F. Lowry, K. J. Frost, and B. A. Mahoney. 2015. "Diet of Beluga Whales, *Delphinapterus leucas*, in Alaska From Stomach Contents." *Marine Fisheries Review* 77, no. 1: 70–84. https://doi.org/10.7755/MFR.77.1.7.

Quinn, T. P., V. Le, and A. P. A. Cardilini. 2021. "Test Set Verification Is an Essential Step in Model Building." *Methods in Ecology and Evolution* 12, no. 1: 127–129. https://doi.org/10.1111/2041-210X.13495.

Renner, M., K. J. Kuletz, and E. A. Labunski. 2017. "Seasonality of Seabird Distribution in Lower Cook Inlet." OCS Study BOEM 2017-011. U.S. Department of the Interior, Bureau of Ocean Energy Management, Headquarters.

Rugh, D. J., K. E. W. Shelden, and R. C. Hobbs. 2010. "Range Contraction in a Beluga Whale Population." *Endangered Species Research* 12, no. 1: 69–75. https://doi.org/10.3354/esr00293.

Russakovsky, O., J. Deng, H. Su, et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115, no. 3: 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Salinas, A., and F. Morstatter. 2024. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. arXiv, (2401.03729v3). http://arxiv.org/abs/2401.03729.

Saulitis, E., L. A. Holmes, C. Matkin, K. Wynne, D. Ellifrit, and C. St-Amand. 2015. "Short Note: Biggs Killer Whale (*Orcinus orca*) Predation on Subadult Humpback Whales (*Megaptera novaeangliae*) in Lower Cook Inlet and Kodiak, Alaska." *Aquatic Mammals* 41, no. 3: 341–344. https://doi.org/10.1578/am.41.3.2015.341.

Saulitis, E. L., C. O. Matkin, and F. H. Fay. 2005. "Vocal Repertoire and Acoustic Behavior of the Isolated AT1 Killer Whale Subpopulation in Southern Alaska." 83, no. 8: 1015–1029. https://doi.org/10.1139/Z05-089.

Schall, E., I. I. Kaya, E. Debusschere, P. Devos, and C. Parcerisas. 2024. "Deep Learning in Marine Bioacoustics: A Benchmark for Baleen Whale Detection." *Remote Sensing in Ecology and Conservation* 10, no. 5: 642–654. https://doi.org/10.1002/rse2.392.

Schwinger, R., P. V. Zadeh, L. Rauch, et al. 2026. "Foundation Models for Bioacoustics–a Comparative Review." *Ecological Informatics* 96: 103765.

Shelden, K. E. W., K. T. Goetz, D. J. Rugh, D. G. Calkins, B. A. Mahoney, and R. Hobbs. 2016. "Spatio-Temporal Changes in Beluga Whale, *Delphinapterus leucas*, Distribution: Results From Aerial Surveys (1977-2014), Opportunistic Sightings (1975-2014), and Satellite Tagging (1999-2003) in Cook Inlet, Alaska." *Marine Fisheries Review* 77, no. 2: 1–31. https://doi.org/10.7755/mfr.77.2.1.

Shelden, K. E. W., D. J. Rugh, B. A. Mahoney, and M. E. Dahlheim. 2003. "Killer Whale Predation on Belugas in Cook Inlet, Alaska: Implications for a Depleted Population." *Marine Mammal Science* 19, no. 3: 529–544.

Small, R. J., B. Brost, M. Hooten, M. Castellote, and J. Mondragon. 2017. "Potential for Spatial Displacement of Cook Inlet Beluga Whales by Anthropogenic Noise in Critical Habitat." *Endangered Species Research* 32, no. 1: 43–57. https://doi.org/10.3354/esr00786.

Stowell, D. 2022. "Computational Bioacoustics With Deep Learning: A Review and Roadmap." *Peer J* 10: e13152. https://doi.org/10.7717/peerj.13152.

Thode, A., D. Mathias, J. Straley, et al. 2015. "Cues, Creaks, and Decoys: Using Passive Acoustic Monitoring as a Tool for Studying Sperm Whale Depredation." *ICES Journal of Marine Science* 72, no. 5: 1621–1636. https://doi.org/10.1093/icesjms/fsv024.

Vergara, V., M. A. Mikus, C. Chion, D. Lagrois, M. Marcoux, and R. Michaud. 2025. "Effects of Vessel Noise on Beluga (*Delphinapterus leucas*) Call Type Use: Ultrasonic Communication as an Adaptation to Noisy Environments?" *Biology Open* 14, no. 3: bio061783. https://doi.org/10.1242/BIO.061783/367434.

Warlick, A. J., G. K. Himes Boor, T. L. McGuire, et al. 2024. "Identifying Demographic and Environmental Drivers of Population Dynamics and Viability in an Endangered Top Predator Using an Integrated Model." *Animal Conservation* 27, no. 2: 240–252. https://doi.org/10.1111/acv.12905.

Wu, Y., K. Chen, T. Zhang, et al. 2024. Large-Scale Contrastive Language-Audio Pretraining With Feature Fusion and Keyword-to-Caption Augmentation. arXiv 2211.06687v4. http://arxiv.org/abs/2211.06687.

Yuksekgonul, M., F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. 2023. When and Why Vision-Language Models Behave Like Bags-of-Words, and What to Do About It? arXiv 2210.01936v3. http://arxiv.org/abs/2210.01936.

Zhong, M., M. Castellote, R. Dodhia, J. Lavista Ferres, M. Keogh, and A. Brewer. 2020. "Beluga Whale Acoustic Signal Classification Using Deep Learning Neural Network Models." *Journal of the Acoustical Society of America* 147, no. 3: 1834–1841. https://doi.org/10.1121/10.0000921.

Zimmer, W. M. X. 2011. Passive Acoustic Monitoring of Cetaceans. https://doi.org/10.1017/CBO9780511977107.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Datasets from the CIBA program used in this study for training and testing DNN models. Data from the Tuxedni Channel and Johnson River areas were used for steps 2 and 3 in Figure 2, with step 3 performed using unseen data only. The remaining data were used for the development dataset. **Figure S1:** Relationship between the confidence threshold and the percentage of results to manually validate. **Table S2:** Error rates among confident predictions accepted as labels. **Figure S2:** Training and validation loss curves over 10 epochs for (a) the binary classification model and (b) the multiclass classification model.