Efficient Auditory Coding for Bat Vocalizations

Testing Auditory Kernel Efficiency on Rhinolophus Affinis Calls

Aleksandra Savova[1]

Supervisor(s): Jorge Martinez[1]**, Dimme de Groot**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Aleksandra Savova
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez, Dimme de Groot, David Tax

***Abstract*—Efficient neural coding is a theoretical model in sensory neuroscience, positing that biological systems maximize information transfer to the brain while minimizing neural resources. While this concept has been extensively studied in the context of human speech perception and the human brain, its applicability to non-human vocalizations remains relatively unexplored. This study applies sparse coding to bat echolocation calls and demonstrates that the resulting kernel representations exhibit properties consistent with efficient coding principles, namely high compactness, sparsity, and functional specialization. Distinct kernel activation profiles were found to encode different echolocation call shapes and identify anomalous, irregular calls, indicating that the model captures biologically relevant features and exhibits sensitivity to deviations from stereotyped call structure.**

These findings underscore the advantages of sparse coding over traditional signal representations for modeling bat vocalizations and align with evidence that efficient coding strategies are shared across mammals, tuned to species-specific vocal patterns and conspecific vocalizations. This work improves the interpretability of animal auditory processing and provides a computational basis for modeling mammalian vocalizations, thereby supporting further research in decoding animal signals and interspecies communication.

I. Introduction

Human auditory perception excels at extracting meaningful information from speech, despite variability in vocal tract anatomy, prosody[1], speaking rate, dialects, and environmental noise in the raw acoustic signal across speakers and listening conditions [1–4]. Traditional linguistic theory posits that discrete units, such as phonemes[2] and morphemes[3], construct meaning; yet these units do not correspond to separable regions of the continuous acoustic waveform [5–7] [8]. A single phoneme may span multiple time frames, while a single frame may contain acoustic cues from multiple overlapping phonemes due to coarticulation[4] [8, 9]. This many-to-many mapping between acoustic features and linguistic units suggests that the auditory system employs an alternative, more efficient strategy to encode language structure rather than relying solely on discrete symbolic segmentation at the perceptual level [8–10].

Thus, it has been hypothesized that biological systems have evolved a highly efficient representation of sensory input to maximize the information conveyed to the brain and minimize the metabolic resources necessary for processing it [11]. In this view, called the efficient coding hypothesis, all perceptual content, including natural sounds, is expressed as sequences of neural spikes, where the number of spikes is kept as small as possible to reduce neural activation costs [11]. In the auditory system, this implies that early structures such as the cochlea and auditory nerve

[1] *prosody* refers to the rhythm, stress, and intonation patterns of speech.

[2] *phonemes* are the smallest sound unit in a language that can distinguish meaning, such as /p/ vs. /b/ in "pat" and "bat".

[3] *morphemes* are the smallest grammatical unit in a language that carry meaning, such as the root *fly* or the suffix *-ing*.

[4] *coarticulation* is the phenomenon where phonemes overlap in time, causing the articulation of one phoneme to be influenced by surrounding phonemes. For example, in the words "tea" and "too," the /t/ sound is produced differently: in "tea" the tongue is closer to the front of the mouth because of the front vowel /i/, while in "too" it is articulated further back due to the rounded back vowel /u/.

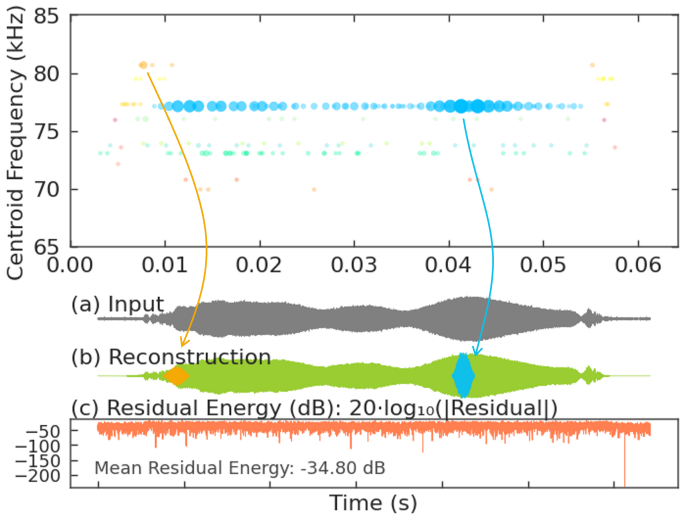

Fig. 1: **Sparse coding decomposition of a bat echolocation call.** Reconstruction used 200 spikes from 13 distinct kernels out of a 32-kernel dictionary. The top panel shows kernel activations as spikes (circles) over time and centroid frequency; point size reflects amplitude, and color denotes each of the 13 unique kernels. The original waveform (a) in gray, its sparse reconstruction (b) in green, and the residual (c) in red demonstrate how a small, specialized set of kernels efficiently captures complex acoustic structure, with mean squared error 1.918×10^{-4} and signal-to-noise ratio 20.62 dB.

are specifically optimized to produce sparse, information-rich representations of sound, filtering and compressing the acoustic signal before it reaches higher processing stages [12].

One computational approach aligned with this principle is *sparse coding*, where a signal is represented as a linear combination of only a few active elements chosen from a larger pool of possible components (Fig. 1). These elements, often called kernels or basis functions, are learned from data and tend to capture recurring patterns within it, allowing the input signal to be broken down into a *small* set of non-zero kernel activations [13]. Smith and Lewicki (2006) used sparse coding to test the efficient coding hypothesis in human speech perception [14].

This study extends Smith and Lewicki's approach for human speech to non-human vocalizations by focusing on the bat species *Rhinolophus affinis*. Bats provide a compelling test case for the efficient coding hypothesis due to their highly specialized audio-vocal systems [15, 16]. Most species, namely laryngeal echolocators, have evolved the ability to emit ultrasonic vocalizations, often exceeding 100 kHz, and analyze the returning echoes of their voice for spatial navigation and hunting prey – an adaptation called echolocation [15]. In addition to echolocation calls, bats are gregarious animals, living in complex social colonies, and as such produce various social calls to communicate distress, agony, isolation, territorial and foraging behaviors, etc [15].

Thus, to evaluate if sparse coding models trained on bat calls adhere to the principles of the efficient coding hypothesis, we pose the following questions:

RQ 1. *What spectral characteristics define the structure of the kernels learned through the sparse coding of bat*

vocalizations?

RQ 2. *To what extent do sparse representations achieve greater coding efficiency compared to traditional signal representation methods, such as Fourier and wavelet transforms?*

RQ 3. *To what degree do the learned kernels show functional specialization, with clusters of similar activation profiles encoding specific variations in bat calls?*

RQ 4. *To what extent do the learned representations exhibit sparsity, with a high prevalence of inactive (near-zero) coefficients across the kernel dictionary?*

To test whether efficient auditory coding applies to *R. affinis*, we followed Smith and Lewicki's methodology, using matching pursuit and gradient ascent to learn kernel dictionaries. Unlike Smith and Lewicki, however, we lacked access to species-specific biologically informed filters such as gammatone filters [17] or reverse correlation filters of real auditory responses, as such resources are unavailable for bats. Therefore, we tested the efficient coding hypothesis from a computational standpoint, analyzing the resulting kernels for reconstruction sparsity and kernel specialization by using unsupervised learning clustering techniques.

Our findings suggest that the learned kernel dictionaries do exhibit specialization, with some kernels consistently reconstructing certain call features, while others corresponding to background noise. Moreover, clustering analyses reveal distinct subtypes of echolocation calls and irregular echolocation structures that may correspond to behavioral or recording anomalies. These results support the notion that sparse, efficient representations can emerge from non-human vocalizations and support the idea that auditory coding strategies are shared across mammalian species [18–21]. The long-term implications include improved explainability in computational models of animal vocalizations and a potential foundation for interspecies communication.

II. Related Work

The human auditory system processes incoming sound by splitting it into its constituent frequencies using a small spiral-shaped organ in the inner ear, called the cochlea [22]. The cochlea is covered by a sensitive membrane with many hair cells vibrating at location-specific frequencies, effectively encoding sound on a continuous logarithmic scale [22, 23]. Johannesma (1972) modeled this process using cat revcor functions[5] to derive a mathematical approximation of cochlear frequency decomposition, called the gammatone filters [17].

Sparse Coding in Auditory Processing. Later, Lewicki (2002) demonstrated that sparse coding kernels learned on natural sounds (animal vocalizations and non-biological environmental noise) show a striking similarity to the gammatone filters, and thus to the auditory nerve fibers [12]. Smith and Lewicki (2006) expanded this into a mathematical model of human auditory processing by optimizing a set of kernels on a large corpus of human speech recordings. This involved gradient ascent to refine kernel parameters, and matching pursuit, a greedy algorithm that iteratively selects the kernel that reduces the residual energy the most. The resulting kernels (short waveform fragments) again resembled cochlear filters, alongside auditory nerve fiber frequency bandwidths of cat revcor filters and yielded greater coding efficiency than conventional signal representations [14].

Collectively, these findings suggest that early auditory coding adheres to information-theoretic principles, approaching a mathematical optimum, and as such can be approximated computationally. Moreover, they further imply that the peripheral auditory system and speech production properties of spoken language have co-evolved to maintain efficient processing. Put simply, humans speak the language they can hear, and hear the language they can produce.

Auditory Selectivity in Echolocation Bats. However, this principle of auditory tuning to one's own vocalizations is not unique to humans. Echolocating bats, which rely heavily on precise auditory processing for perception and survival, exhibit striking neural specialization for conspecific[6] vocalizations [15]. Wohlgemuth and Moss (2016) showed that midbrain neurons of the bat *Eptesicus fuscus* respond selectively to natural echolocation calls but show weak or nonspecific responses to synthetic sounds with matching acoustic characteristics [24]. Similarly, Park et al. (2021) managed to produce receptive fields (STRFs) similar to those of real bat neurons using an artificial neural network trained on bat social calls, further supporting that bat auditory processing is highly optimized for conspecific vocalizations [25].

Together, these studies reveal a consistent pattern: mammalian auditory systems prioritize efficient coding of same-species vocalizations, where both humans [4] and bats exhibit neural tuning to the natural statistics of their own calls and can discriminate real vocalizations from pseudonatural sounds. The present study extends prior work by focusing on how such selectivity emerges in the early auditory structures (the cochlea and auditory nerve), independent of higher-level neural processing. It examines what patterns in the call structure alone may account for this selectivity. In contrast to previous bat studies relying on black-box models, we apply sparse coding to test whether bat vocalizations are intrinsically optimized for sparse representation, as shown for human speech. This approach provides greater interpretability and a computational basis for animal auditory processing.

III. Methodology

The following section details the method used to address the sparse coding of bat vocalizations, first by introducing the dataset and the preprocessing steps, then delving into the theory behind the sparse representation model and the metrics used to evaluate the structural properties (**RQ 1**), compression efficiency (**RQ 2**), functional specialization (**RQ 3**), and activation sparsity (**RQ 4**) of the learned kernels.

A. The ChiroVox Dataset

To derive the auditory kernels of bat vocalizations, we utilized the ChiroVox database, the largest open-access bat call library to date, including 5793 recordings from 255 species [26]. Among these, *R. affinis* was selected for this study for

[5] *reverse correlation (revcor) functions* are mathematical representations of how a neuron transforms a stimulus (input sound, such as white noise) into its output; corresponds to the auditory nerve fiber responses.

[6] other member of the same species

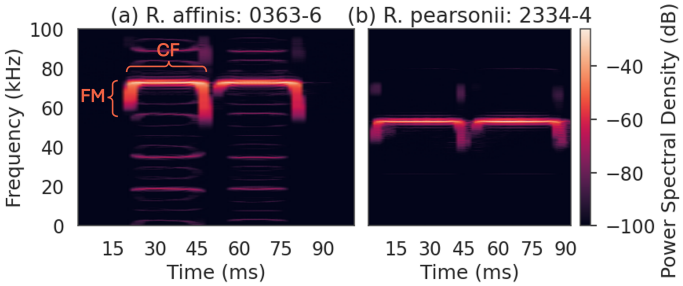

Fig. 2: **Echolocation call spectrograms of two Rhinolophidae species.** Both species display characteristic call structures with constant-frequency (CF) components surrounded by brief frequency-modulated (FM) sweeps in the onset and offset, typical of rhinolophid echolocation. (a) Two CF-FM *R. affinis* echolocation calls with dominant harmonic estimated at 72.9 kHz, with visible secondary harmonics appearing as faint horizontal lines. (b) Two CF-FM *R. pearsonii* echolocation calls with estimated dominant harmonic at 53.2 kHz.

having the highest number of publicly available recordings (262). This species belongs to the group of CF-FM bats, known for producing long-duration echolocation calls with a broadband.[7] frequency-modulated (FM) onset and offset surrounding a long narrowband[8] constant-frequency (CF) middle segment (see Fig. 2). These vocalizations are highly stereotyped across individuals, which reduces intra-species acoustic variability. This consistency was particularly advantageous due to the limited available training data, as it allowed the training algorithm to converge more rapidly and similarly across runs, thereby supporting reproducibility despite the constrained data conditions.

To better contextualize the efficiency and sparsity of the learned representations, we included a closely related species, *Rhinolophus pearsonii*, from the same family. *R. pearsonii* often coexists with *R. affinis* in overlapping habitats, including shared cave systems [27]. Its echolocation calls exhibit a similar CF-FM structure but differ in dominant frequency content: most of the call energy in *R. pearsonii* is concentrated between 57.6–70.0 kHz, compared to 70.0–88.5 kHz in *R. affinis* [28]. This acoustic resemblance (Fig. 2), combined with their ecological proximity, makes *R. pearsonii* a suitable comparative case for assessing the generalizability of the *R. affinis*-trained model to sister-species vocal patterns.

B. Bat Call Analysis

To derive a shared dictionary from *R. affinis* call recordings, all signals must be standardized to a common sample rate. However, selecting this rate involves a trade-off: higher sample rates preserve high-frequency detail, including biologically relevant features, while lower sample rates reduce computational cost and accelerate training.

One such biologically important feature is the harmonic structure of CF-FM vocalizations. Harmonics are frequency components at integer multiples of a fundamental frequency. In such calls, the dominant (most energetic) harmonic corresponds to the CF component, which falls within a narrow, species-specific range and plays a central role in echolocation and prey detection. Higher-order, fainter harmonics, as shown in Fig. 2 (a), improve spectral resolution, helping bats distinguish their target in cluttered environments [29]. Consequently, preserving these harmonics in the training data is essential for accurately modeling auditory processing.

Thus, spectral content was analyzed to identify the highest frequency of interest with the following metrics:

1) Peak Frequency f_{peak}

The frequency with the highest spectral energy across the signal duration, identifying the dominant harmonic:

$$f_{\text{peak}} = \arg\max_f \left(\sum_t S_{xx}(f,t) \right)$$

where $S_{xx}(f,t)$ is the signal power spectrogram.

2) Maximum Frequency Above Noise Floor f_{max}

The highest frequency at which the maximum dB-scaled power across time exceeds a noise threshold η, corresponding to secondary call harmonics:

$$f_{\max} = \max \left\{ f \,\middle|\, \max_t S_{xx}^{\text{dB}}(f,t) > \eta \right\}$$

where:

- $S_{xx}^{\text{dB}}(f,t) = 10\log_{10}(S_{xx}(f,t) + \varepsilon)$ is the dB-scaled spectrogram,
- $\varepsilon = 10^{-12}$ is a small constant to avoid logarithm of zero,
- $\eta = -30\,\text{dB}$ is the noise floor threshold.

Recordings were downsampled following the Nyquist criterion, ensuring the sampling rate was at least twice the first overtone of the dominant harmonic. Lastly, bat call segments were categorized into two groups based on their peak frequency (f_{peak}): social calls with $f_{\text{peak}} < 25\,\text{kHz}$ and echolocation calls with $f_{\text{peak}} \geq 25\,\text{kHz}$.

C. Denoising with Butterworth Filters

Environmental and anthropogenic noise were present in some recordings. To ensure that the resulting kernels captured only the structure of bat calls, recordings were denoised using a 5th-order Butterworth bandpass filter with a high-pass cutoff frequency at $0.7 \times f_{\text{peak}}$ to isolate the CF-FM call from low-frequency noise.

This approach was well suited for *R. affinis* echolocation calls, as noise interference is minimal in the ultrasonic range, with scarce natural sources emitting at such high frequencies. Alternatives such as standard denoising libraries were considered but found to degrade recording quality by attenuating calls, due to a lack of domain-specific tuning, as shown in Fig. 3.

D. Bat Call Detection

Bat calls were detected using an energy-based method adapted from [30], as illustrated in Fig. 4. Recordings were first segmented into 50 ms chunks and high-pass filtered with a cutoff frequency of $0.7 \times f_{\text{peak}}$ to suppress low-frequency noise. Within each segment, short-term energy was computed over 6.25 ms windows (1250 samples at 200 kHz), with

[7] *broadband* refers to a wide range of frequencies present simultaneously in a sound, resulting in a complex, spectrally rich signal.

[8] *narrowband* refers to a sound with energy concentrated in a relatively small frequency range, typically producing a tonal, pure-tone-like signal.

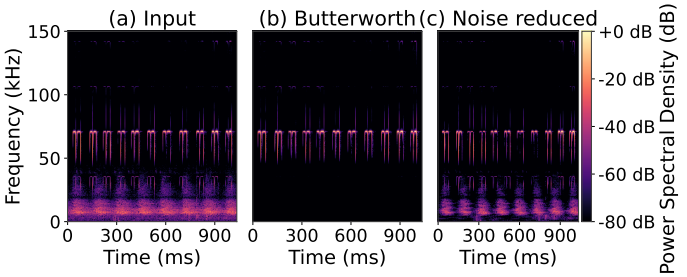

Fig. 3: **Denoising performance on *R. affinis* echolocation call.** Spectrograms showing (a) the original audio recording, (b) Butterworth highpass-filtered version, 5th order, cutoff at 70% of the dominant harmonic frequency, and (c) Python-denoised version using the `noisereduce` library. The clear separation between environmental noise (<50 kHz) and the bat's call (>50 kHz) makes the Butterworth filter particularly effective for denoising. Conversely, the Python denoising library lacks biological context about relevant call features, preserving low-frequency noise while attenuating bat call components.

50% overlap, yielding 15 energy values per segment. These values were normalized such that the median energy within each segment was set to 0 dB. A call was detected if the maximum energy within the segment exceeded a threshold of 3 dB.

Since the original method was designed for FM calls, an adjustment was made to accommodate the CF-FM structure of *R. affinis* calls, where most energy is concentrated in the central CF component. To capture the less energetic preceding and trailing FM sweeps, a 5 ms buffer was added before and after each detected call. Overlapping detections were merged into a single segment.

These segmented calls served as the input for the sparse coding analysis described next.

E. The Sparse Coding Problem

Sparse coding seeks to represent a time-domain signal $x(t) \in \mathbb{R}^T$ as a linear combination of a small number of atoms selected from an overcomplete dictionary $\mathcal{D} = \{\phi_\gamma(t)\}_{\gamma\in\Gamma}$:

$$x(t) \approx \sum_{k=1}^{K} a_k\, \phi_{\gamma_k}(t), \quad \text{with } K \ll |\Gamma|,$$

where:

- $a_k \in \mathbb{R}$ are scalar activation coefficients (or amplitude magnitudes),
- $\phi_{\gamma_k}(t)$ is the k-th selected atom (kernel) from the dictionary $\mathcal{D} = \{\phi_\gamma(t)\}_{\gamma\in\Gamma}$,
- $\gamma_k = (\tau_k, \omega_k, s_k) \in \Gamma$ is a parameter tuple specifying the time shift τ_k, center frequency ω_k, and scale s_k,
- Γ is the index set defining all possible combinations of parameters for atoms in the dictionary,
- K is the number of active atoms used in the sparse approximation, typically much smaller than $|\Gamma|$.

The sparse coding objective is to find the best set of K atoms and coefficients that minimize the reconstruction error:

$$\min_{\{a_k,\gamma_k\}_{k=1}^K} \left\| x(t) - \sum_{k=1}^{K} a_k\, \phi_{\gamma_k}(t) \right\|_2^2 \quad \text{s.t. } K \ll |\Gamma|$$

Here, $\|\cdot\|_2$ denotes the standard Euclidean (L2) norm. Since the problem is combinatorial and NP-hard, approximate methods are used to find sparse solutions.

F. Matching Pursuit

One such solution is Matching Pursuit, a greedy algorithm for approximating a signal by iteratively selecting dictionary atoms that best match the current residual [31]. Despite the sparse approximation alternatives [32], Matching Pursuit was chosen to maintain consistency with Smith and Lewicki's methodology.

Let the initial residual be $R^{(0)}(t) = x(t)$. At iteration k, the algorithm proceeds as follows:

1) Atom selection

$$\phi_k = \arg\max_{\phi\in\mathcal{D}} \left| \langle R^{(k-1)}, \phi \rangle \right|,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product on $\mathbb{R}^T$.

2) Coefficient computation

$$a_k = \langle R^{(k-1)}, \phi_k \rangle.$$

3) Residual update

$$R^{(k)}(t) = R^{(k-1)}(t) - a_k\, \phi_k(t).$$

The process is repeated until either a fixed number of atoms K has been selected or the residual energy $\|R^{(k)}\|_2^2$ falls below a specified threshold.

G. Compression Efficiency

Compression efficiency was evaluated by measuring reconstruction fidelity using signal-to-noise ratio (SNR) and bitrate, assuming 64-bit coefficients without time quantization (frame-level precision). This analysis compared matching pursuit reconstructions with standard signal representation methods, such as Fourier and Daubechies DB4 wavelet transforms, across both *R. affinis* and *R. pearsonii* recordings.

H. Kernel Analysis Metrics

To analyze the spectral properties of the learned kernels for **RQ 1**, three metrics were employed:

1) Bandwidth

Defined as the frequency range containing 90% of a kernel's spectral energy, this metric was used to assess kernel specialization (**RQ 3**) in encoding narrowband CF components or broadband FM sweeps, in line with the CF-FM structure of *R. affinis* calls.

2) Centroid Frequency

Calculated as the weighted mean frequency of a kernel's power spectrum, this metric indicated if kernels were concentrated in the ultrasonic range characteristic of Rhinolophidae echolocation.

3) Spectral Skewness

Lewicki's findings suggest that dictionaries trained on both natural noise and vocalizations tend to produce asymmetric sinusoidal shapes with rapid onsets and slow decays, whereas training on vocalizations alone yields more symmetric shapes. Skewness was used to quantify whether kernels optimized on bat vocalizations exhibited similar symmetry patterns.

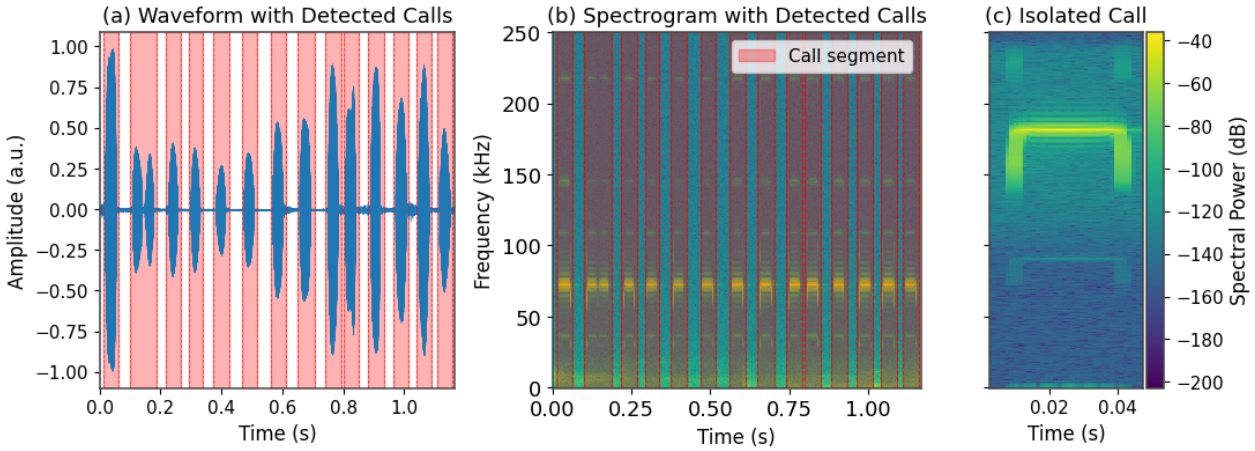

Fig. 4: **Energy-based call detection for *R. affinis*.** Audio recording is divided into segments based on its duration. Energy is calculated over 50% overlapping windows within each segment and normalized by subtracting the global median energy across all segments. Calls are identified where energy exceeds a 3 dB threshold and padded with 5 ms to make sure to capture FM sweeps. Panels display (a) the raw waveform with detected calls highlighted in red, (b) full recording spectrogram with detected calls, and (c) spectrogram of an isolated call after preprocessing.

I. Bat Call Variation Clustering

To assess kernel specialization (**RQ 3**), K-means clustering was applied to bat call kernel activation profiles. Each bat call segment i was represented as a vector $\mathbf{x}_i \in \mathbb{R}^{32}$, where $\mathbf{x}_i[k]$ denotes the L_1-normalized count of kernel k activations during reconstruction.

For the scope of the present analysis, clusters were assumed to be convex. To mitigate issues related to distance concentration [33], only kernel activation counts were used, excluding amplitude and temporal shift dimensions, to keep the feature space dimension minimal. To visually inspect the presence of meaningful low-dimensional structure, Uniform Manifold Approximation and Projection (UMAP) was applied to assess whether the data formed distinct clusters or appeared uniformly distributed, which would suggest a lack of inherent cluster structure. However, due to its known limitations [34], UMAP results were used solely as a qualitative aid and not as a basis for clustering decisions.

Since no ground truth labels were available, clustering quality was instead evaluated across different reconstruction depths and values of K with the Elbow Method, Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI).

J. Sparsity Metrics

To evaluate sparsity, call segments were represented by their activation count vectors (see Sec. H). Sparsity was assessed using three measures: the Gini index, the Hoyer index, and the (p, q)-norm with $p = 0.5, q = 2$, following the recommendations from Hurley and Rickard [35].

Let $\mathbf{x} \in \mathbb{R}^n$ denote the non-negative, L1-normalized activation vector for a segment.

1) Gini Index

$$G(\mathbf{x}) = \frac{1}{n-1}\left(n + 1 - 2 \cdot \frac{\sum_{i=1}^{n}(n+1-i)\cdot x_{(i)}}{\sum_{i=1}^{n} x_i}\right),$$

where $x_{(i)}$ are the entries of $\mathbf{x}$ sorted in ascending order. Higher values indicate greater sparsity.

2) Hoyer Index

$$H(\mathbf{x}) = \frac{\sqrt{n} - \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2}}{\sqrt{n} - 1},$$

which ranges from 0 (dense) to 1 (maximally sparse).

3) (p, q)-Norm Ratio (with $p = 0.5, q = 2$)

$$S_{(0.5,2)}(\mathbf{x}) = \frac{\|\mathbf{x}\|_{0.5}}{\|\mathbf{x}\|_2},$$

where $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$. This ratio favors sparsity with near-zero values when few large coefficients dominate.

Sparsity metrics were computed per segment and summarized across the dataset to assess overall sparsity. In addition, the raw distributions were visualized as qualitative confirmation.

IV. Experimental Setup

Sparse Dictionary Learning. Segmented bat call recordings, resampled to 200 kHz were used for six separate trainings to derive sparse dictionaries of 32 auditory kernels each, varying in kernel initialization size (400 samples and 100 samples) and training subset (all calls, echolocation-only, or social-only). The 100-sample initialization follows Smith and Lewicki's approach, while the 400-sample size accounts for differences in sample rate and call duration between human speech syllables [36] and bat vocalizations [37]. Kernel number was chosen for consistency with Smith and Lewicki's original work [14], since dictionary cardinality alone is not a reliable measure of diversity [38]. Instead, applying the kernels convolutionally already guarantees that the effective dictionary becomes highly overcomplete [39], known to produce more efficient representations with higher sparsity and lower reconstruction error [40, 41].

Matching Pursuit Reconstruction. The echolocation training set was encoded using Matching Pursuit and the echolocation-trained dictionary (400 samples) with a rate of 18,000 kernels per second for high reconstruction precision. The reconstruction leveraged code developed by Dimme

de Groot (supervisor), which was used both for dictionary learning and signal reconstruction [42].

Data Preprocessing Pipeline and Analysis. Data preprocessing and result analysis were performed using the following open-source libraries, with the full Python script available on the 4TU repository [43]:

- Librosa [44, 45] and Soundfile were used for audio file handling and signal processing.
- NumPy [46], SciPy [47, 48], scikit-learn [49, 50], and Pandas [51, 52] supported statistical analysis and data manipulation.
- Matplotlib [53, 54] was used for visualizing experimental results.

V. Results

Below, we present the key findings from the analysis of the learned sparse representations.

A. Spectral Characteristics of the Learned Kernels

Addressing **RQ 1**, the spectral properties of the learned kernels were analyzed in terms of their spectral centroid and bandwidth, as summarized in Table I. The spectral skewness across all kernels exhibited a mean of –0.0020 with a low standard deviation of 0.0472, indicating highly symmetric spectral shapes.

The spectral centroids of the kernels spanned a broad frequency range, from as low as approximately 10 kHz to above 85 kHz. Bandwidths also varied widely, reflecting differences in spectral selectivity among kernels.

Notably, two distinct spectral groups emerged:

1) **Narrowband high-frequency kernels** with centroids above 70 kHz and very narrow bandwidths, often below 1 kHz (e.g., kernels 1, 6, 8, 9, 15, 17).
2) **Broadband low-frequency kernels** with relatively low centroids and very broad bandwidths, some exceeding 30 kHz (e.g., kernels 28, 29, 30, 32)

Other kernels displayed intermediate bandwidths and centroids across the mid-frequency range.

Figure 5 shows the kernel activation counts for the first 70 most energetic spikes in the matching pursuit reconstruction, illustrating the relative usage of individual kernels. It is important to note that increase in representation precision shifts the preference from narrowband to broadband kernels, as shown by the activation counts for the full 18,000 kernels/s reconstruction, provided in Appendix B.

B. Reconstruction Fidelity and Compression

Figure 6 presents fidelity–rate curves comparing Matching Pursuit (MP) spike coding using *R. affinis*-trained kernels, Fourier transforms, and Daubechies wavelets for echolocation calls of *R. affinis* and *R. pearsonii*. The testing dataset comprised 1000 calls per species.

As shown in the figure, MP coding for *R. affinis* achieves significantly higher reconstruction fidelity than both Fourier and wavelet methods, reaching an SNR of 31 dB at higher bitrates. For *R. pearsonii*, Fourier outperforms MP with the *R. affinis*-trained dictionary. The narrow 95% confidence intervals (under 1 dB) confirm the reconstruction fidelity results for **RQ 2** are consistent.

TABLE I: **Spectral centroid and bandwidth (BW) of the learned kernels.** Kernels show diverse frequency selectivity, ranging from narrowband high-frequency to broadband low-frequency responses.

Kernel (ID)	Centroid (kHz)	BW (kHz)	Centroid (kHz)	BW (kHz)
1 / 2	72.14	0.78	72.76	14.06
3 / 4	73.77	12.50	69.96	14.06
5 / 6	73.11	13.28	73.58	1.56
7 / 8	77.18	11.72	72.92	1.56
9 / 10	73.11	0.78	85.57	7.81
11 / 12	76.08	11.72	73.92	14.06
13 / 14	79.59	3.13	77.45	4.69
15 / 16	67.83	0.81	64.73	2.79
17 / 18	80.72	0.78	76.05	25.78
19 / 20	84.38	7.81	75.17	17.19
21 / 22	61.07	8.89	87.36	10.16
23 / 24	9.97	3.13	89.21	17.97
25 / 26	65.56	21.09	85.45	14.06
27 / 28	85.58	17.97	13.26	43.75
29 / 30	60.41	40.63	74.99	30.47
31 / 32	83.79	29.69	16.88	35.16

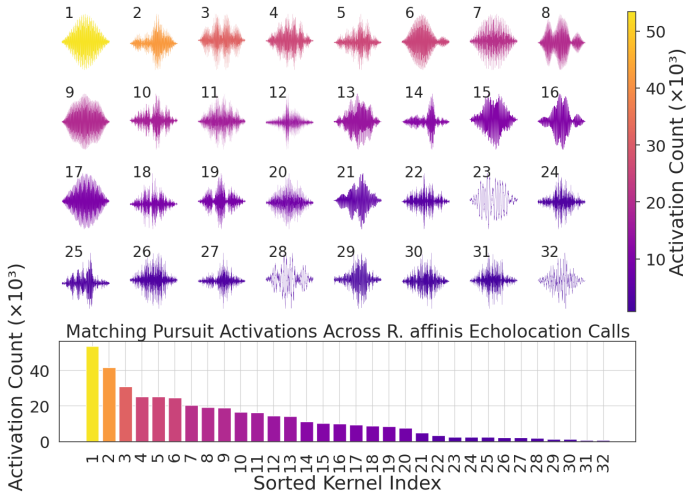

Fig. 5: **Kernel activation across *R. affinis* echolocation calls.** Activation counts for a dictionary of 32 kernels (top), initialized at 400 samples each, trained on 6026 recordings using matching pursuit with 10,000 gradient ascent iterations and expansion/trimming every 50 iterations. The histogram shows total kernel activations across reconstructions of 70 spikes. The most frequently used kernel is a tonal, narrowband high-frequency kernel, likely corresponding to the CF component of the call, which dominates energy and is captured early in matching pursuit iterations. This suggests kernel functional specialization and high information content, as the CF component plays a central role in echolocation navigation.

C. Variations in Bat Call Structure and Kernel Specialization

To address **RQ 3**, clustering was performed on kernel activations extracted from bat call recordings. Clustering quality was evaluated across varying reconstruction depths (top n spikes per recording, with $n \in \{100, 150, 200, 300, 400, 500, 1000\}$) and numbers of clusters $k \in [2, 40]$, using the Silhouette score, Davies–Bouldin Index, and Calinski–Harabasz Index. The optimal clustering parameters were selected as $n = 200$ spikes and $k = 27$ clusters (see Fig. 7).

Two alternative clustering strategies were considered:

1) **Encoding rate-based encoding:** Clustering using variable spike counts defined by a target encoding

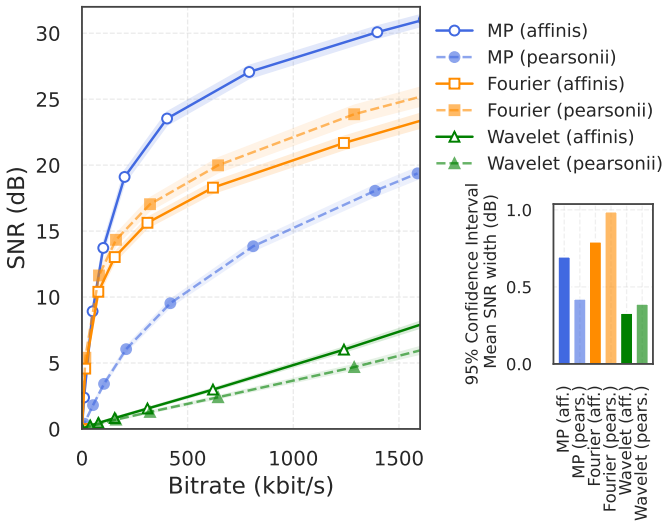

Fig. 6: **Fidelity–rate curves for spike, Fourier, and wavelet coding of *R. affinis* and *R. pearsonii* calls.** Trade-offs between coding cost and signal fidelity are shown for 1000 test echolocation calls per species, comparing Matching Pursuit (MP) spike coding with *R. affinis*-trained kernels (blue), Fourier transforms (orange), and Daubechies wavelets (green). Solid lines indicate *R. affinis*; dashed lines, *R. pearsonii*. Shaded regions denote tight (<1 dB) 95% confidence intervals (CI), with mean CI widths histogram in the bottom right. MP coding consistently outperforms Fourier and wavelet, achieving up to 31 dB SNR for *R. affinis* at high bitrates. Lower SNR for *R. pearsonii* with the same kernels reveals that species-specific acoustic differences hinder cross-species reconstruction. This demonstrates the advantage of species-tailored spike coding in representing bat calls.

rate, preserving longer call sequences (1–2 sec) better and capturing trends such as gradual decreases in call frequency and variations in call repetition rates, such as shorter inter-call intervals.

2) **Amplitude-weighted aggregation:** Clustering based on kernel activations weighted by their amplitudes.

Among these, the fixed spike count approach was preferred, as visualization of individual calls was more interpretable for analysis purposes, despite encoding rate-based clusterings showing comparable performance on quantitative metrics. The amplitude-weighted aggregation method yielded significantly poorer cluster quality.

Figure 8 illustrates the clustering results from the fixed spike count method, with grouped calls exhibiting similar dominant harmonics and spectral shapes. Some clusters mainly contain sequences of calls, but most group individual echolocation calls. Notable cluster features include calls with prominent secondary harmonics, calls with smeared patterns and vertical line artifacts, likely of acoustic origin, and calls characterized by very narrow constant-frequency (CF) components. Lastly, clusters comprising empty recordings and anomalous, irregularly shaped calls were also identified and are shown alongside all 27 clusters in Appendix C.

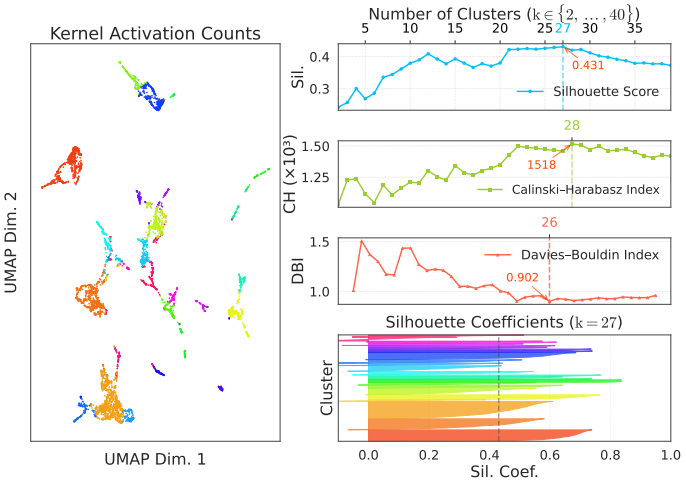

Fig. 7: **Clustering metrics and UMAP projection of kernel activations.** Clustering was performed on the top n kernel activations per recording, with $n \in \{100, 150, 200, 300, 400, 500, 1000\}$. The selected reconstruction depth n and number of clusters k were chosen based on the values where the Silhouette score (Sil.), Calinski–Harabasz Index (CHI), and Davies–Bouldin Index (DBI) agreed and indicated strongest cluster quality (metrics shown on the right).The UMAP embedding on the left offers a qualitative visualization of cluster assignments. The silhouette coefficient plot at the bottom right provides more detailed information on cluster cohesion for the chosen clustering.

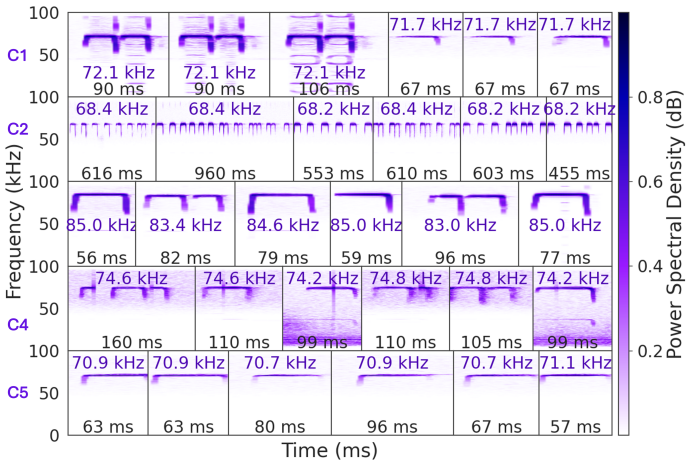

Fig. 8: **Auditory kernel-based clustering captures call diversity.** Five clusters (C1 to C5) are shown, each row displaying the six most central calls with scaled durations. Clustered calls exhibit similar dominant harmonics and shapes. C2 groups call sequences, while others group individual calls. C1 has prominent secondary harmonics, C4 shows smeared calls with vertical line patterns (likely acoustic artifacts), and C5 contains very narrow CF components. These patterns suggest auditory kernel coactivation profiles may encode call structure variations, potentially conveying information about the calls' broader behavioral context.

D. Representation Sparsity

For **RQ 4**, kernel activations per cluster were visualized in Fig. 9, where large white regions indicate near-zero kernel activations. On average, clusters exhibited a Gini coeffi-

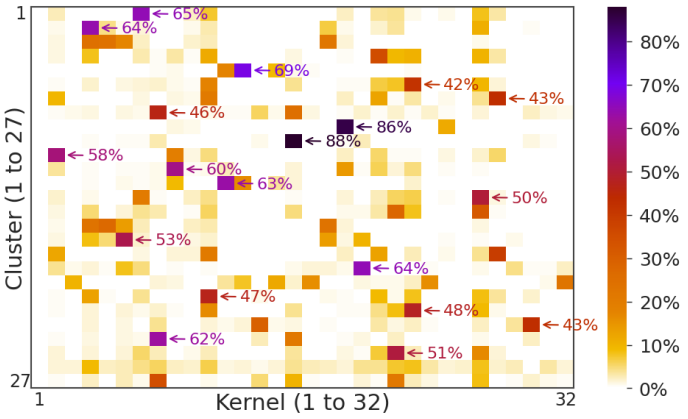

Fig. 9: **Cluster sparsity.** Qualitative visualization of kernel activation distribution for 20 dB (SNR) reconstructions across clusters, with annotated contributions above 40%. Clusters exhibit a mean Gini coeff. of 0.8745, indicating that few kernels dominate within the same cluster, implying that many calls share similar underlying acoustic structures. This supports the idea that bat vocalizations are naturally structured for compact representations, following the biological principle of efficient sensory coding. The observed sparsity, therefore, provides evidence for the biological relevance of the representation method.

cient of 0.8745, suggesting that a small subset of kernels dominates the activation pattern within each cluster. This implies that many calls share similar underlying acoustic structures.

This qualitative observation was supported by the quantitative sparsity metrics, Gini, Hoyer, and PQ, computed per recording. Aggregated results for reconstructions with 200 and 2400 spikes are presented in Table II. All metrics confirm high sparsity levels across both species, with consistently higher sparsity observed for *R. pearsonii*. It is worth noting, however, that for high-bitrate reconstructions at 18,000 kernels/s, *R. pearsonii* activation profiles show only two consistently active kernels. The rest remained rarely used, which explains the higher sparsity scores for *R. pearsonii*, despite lower reconstruction fidelity.

TABLE II: **Sparsity metrics at varied reconstruction depths.** The table shows mean ± standard deviation of Gini, Hoyer, and PQ metrics calculated on 1000 recordings per *R. affinis* and *R. pearsonii*, indicating very high reconstruction sparsity, with *R. pearsonii* being consistently sparser.

Metric	200 kernels	2400 kernels
Gini (aff.)	0.985 ± 0.008	0.997 ± 0.001
Gini (pea.)	0.994 ± 0.004	0.998 ± 0.001
Hoyer (aff.)	0.960 ± 0.028	0.981 ± 0.011
Hoyer (pea.)	0.996 ± 0.013	0.993 ± 0.006
PQ (aff., p=0.5, q=2)	$(3.50 \pm 0.31) \times 10^{-3}$	$(2.3 \pm 0.1) \times 10^{-4}$
PQ (pea., p=0.5, q=2)	$(6.9 \pm 1.5) \times 10^{-4}$	$(1.0 \pm 0.1) \times 10^{-4}$

VI. Responsible Research

This section highlights the core principles of responsible research that have guided this study, focusing on reproducibility, replicability, ethical conduct, and transparent acknowledgment of limitations throughout the research process.

A. Reproducibility and Open Science

This study adheres to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [55] to ensure reproducibility and transparency. In line with these principles, all project code and data have been made publicly available under an open-source license: the code developed by the author is hosted on the 4TU repository [43], while the supervisor's code by Dimme de Groot is openly accessible on GitHub [42]. The dataset used, ChiroVox, is also open-source and fully documented, with all recordings unique identifiers included in the code to facilitate reproducibility [43].

This work employs widely-used open-source libraries, including librosa, pandas, numpy, soundfile, scipy, scikit-learn, matplotlib, and others, all of which are freely available and clearly cited to encourage reuse. Additionally, to ensure consistent results across runs, all randomization processes within the code are seeded with the fixed values 42 and 84.

B. Replicability

In this study, we selected a bat species characterized by highly stereotyped vocalizations to enhance the reliability and consistency of training convergence across multiple runs. This choice was deliberate given the limitations in the available training data, as stereotyped calls reduce variability and facilitate more stable model training despite the relatively small dataset, thereby supporting future reproducibility and replicability.

However, the dataset has inherent limitations. Most recordings in the ChiroVox library, nearly 75%, were collected in controlled environments such as closed spaces or from hand-held bats [26], where specimens were often handled for species identification. As a result, these recordings may not fully reflect typical bat behavior in the wild. In particular, some calls may represent agonistic or distress vocalizations rather than natural communication, potentially biasing the data. Because recordings often involve direct interaction with the bats or their habitat, ethical considerations naturally arise. Given that these recordings were conducted by researchers within the scientific community, we place trust in the established ethical standards and practices upheld by those responsible for the data.

Lastly, the replicability of this research for *R. affinis* is challenging due to the scarcity of comprehensive bat vocalization datasets. Nonetheless, the analytical pipeline developed here is generalizable and can be applied to closely related sister species to assess whether similar patterns and results hold, offering a way to validate and extend our findings beyond the current dataset.

C. Limitations

An important limitation of this study is that the model was trained and evaluated on the same dataset, with the test data being a subset of the training data. Therefore, assessing overfitting or generalization beyond the specific *R. affinis* recordings is not possible. This design choice was driven by practical constraints. Specifically, curating a separate test set that matched the spectral and metadata distribution of the training set would have required considerable additional time and manual effort. Moreover, retraining the model was not feasible within the scope of

the Bachelor Research Project, as a single training cycle on DelftBlue, including queue waiting time, typically required approximately one week. Given these constraints, a separate evaluation phase was considered out of scope. Consequently, reconstruction fidelity and compression efficiency metrics may be influenced by overfitting. Nonetheless, this setup still provides meaningful insight into the spectral structure and specialization of the learned kernels, which was the primary aim of this work.

D. Use of Large Language Models

In the preparation of this manuscript, large language models (LLMs) were employed to assist with tasks such as drafting, grammar correction, and improving the clarity and flow of the text. While AI tools supported the writing process, all the conceptual content, research ideas, analysis, and interpretations remain the original work of the author. The use of LLMs did not influence the scientific conclusions, and the author retains full responsibility for the accuracy and integrity of the research presented.

VII. Discussion and Conclusions

This study explored the spectral characteristics, compression efficiency, and potential biological relevance of auditory kernels trained on *R. affinis* vocalizations. This section places the present findings within a broader scientific context and outlines several directions for future development.

A. Symmetry and Spectral Structure: Insights from Kernel Shapes

The distribution of kernel shapes exhibited symmetry consistent with Lewicki's findings, which showed that auditory kernels trained on animal vocalizations alone tend to be symmetric. Based on this, we expect that training on a combination of bat vocalizations and natural environmental sounds present in bat habitats might yield representations that closely reflect bat auditory processing. However, due to the absence of revcor data or other biologically-informed filters, we are currently unable to evaluate this hypothesis.

Analysis of kernel usage across matching pursuit iterations revealed a pattern consistent with the energy structure of bat calls. Narrowband, high-frequency kernels were predominantly selected in the early iterations of matching pursuit, indicating that these kernels capture the most energetic, constant-frequency (CF) components of the call. In later iterations, the increased selection of kernels with broader bandwidth suggests a shift toward reconstructing frequency-modulated (FM) sweeps. Additionally, three kernels exhibited centroid frequencies below 20 kHz and very broad bandwidths, likely involved in reconstructing background noise.

B. Species-Specific Coding: Sparse, Tuned, and Non-Generalizable

Reconstructions of *R. pearsonii* calls showed lower fidelity and higher sparsity, with only two kernels consistently active across reconstructions. This indicates limited inter-species generalization of kernels trained on *R. affinis*, despite the two species' phylogenetic and acoustic proximity [27]. The finding aligns with established evidence that bat auditory systems are finely tuned to conspecific vocalizations, particularly the dominant harmonic of the CF component, which is species-specific [24, 25].

C. Beyond the Signal: Capturing Call Variation and Degradation

Expert evaluation is required to determine whether the observed clusters capture biologically meaningful variation in bat calls. However, several hypotheses can be proposed. Clustering based on consistent encoding reconstruction rate might be detecting Doppler shift compensation, an adaptation allowing some bat species to lower their call frequency during flight to counteract the frequency increase caused by their own motion, and thus help them lower the call echo to their optimal auditory range [56]. Since this phenomenon manifests as a downward trend in call sequences [56] and takes place only during flight [56], auditory kernel-based clustering might potentially distinguish resting calls from in-flight vocalizations. Additionally, clusters containing temporally smeared calls likely reflect echoes or increased distance from the microphone, suggesting that auditory kernels can capture degraded or noisy vocalizations. Given that *R. affinis* is widespread in Southeast Asia and includes numerous subspecies [37], it would be valuable to assess whether kernel-based clustering distinguishes among them. Unfortunately, for now, the ChiroVox dataset lacks subspecies annotations, preventing us from comparing clustering results against ground truth labels. Although the biological significance of the identified call structure variations remains inconclusive, they pose a compelling case for further investigation by specialists.

D. Future Directions

Several additional directions remain for directly extending this work. First, the method could be applied to a larger dataset of social calls, as currently their number was too limited to derive meaningful results, and the focus had to be shifted to echolocation calls alone. For future exploration, applying the method to other bat families or animal species might further confirm its broader generalizability. Another line of development is to examine the conditional distributions of kernel co-activations, or how the presence of certain kernels determines the presence (or lack thereof) of others, as this could potentially offer insights into the information gain of certain call features. Finally, expert evaluation and long-term integration of the present results with biologically-informed auditory models, if such are developed, could significantly improve results interpretability and give invaluable insight into the biological relevance of the current model.

E. Broader Implications and Final Remarks

Despite present limitations, results are highly encouraging, uncovering a fascinating and largely unexplored area of research. Demonstrating the generalizability of auditory kernels is a crucial step toward revealing the fundamental information units of the raw acoustic waveform. This progress can not only advance the development of artificial, explainable models capable of synthesizing and modifying mammalian vocalizations but also opens exciting possibilities for deeper insights into animal communication.

This work lays a foundation for computational modeling of animal auditory processing, thereby supporting future research in decoding animal vocalizations and interspecies communication.

References

[1] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, Apr. 2015.

[2] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, May 2008.

[3] K. I. Kirk, D. B. Pisoni, and R. T. Miyamoto, "Effects of stimulus variability on speech perception in listeners with hearing impairment," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 6, pp. 1395–1405, Dec. 1997.

[4] A. Vouloumanos and J. F. Werker, "Tuned to the signal: The privileged status of speech for young infants," *Developmental science*, vol. 7, no. 3, pp. 270–276, 2004.

[5] P. Ladefoged, K. Johnson, and P. Ladefoged, *A course in phonetics*. Thomson Wadsworth Boston, 2006, vol. 3.

[6] Department of Linguistics, Ohio State University, "Phonology," in *Language Files: Materials for an Introduction to Language and Linguistics*, 11th ed., Columbus, OH: Ohio State University Press, 2011, pp. 102–135.

[7] V. Fromkin, R. Rodman, and N. Hyams, "Morphology: The words of language," in *An Introduction to Language*, 10th ed. Cengage Learning, 2013, pp. 33–75.

[8] M. S. Lewicki, "A signal take on speech," *Nature*, vol. 466, no. 7308, pp. 821–822, Aug. 2010.

[9] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological Review*, vol. 74, no. 6, pp. 431–461, 1967.

[10] C. E. Stilp and K. R. Kluender, "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," *Proceedings of the National Academy of Sciences*, vol. 107, no. 27, pp. 12 387–12 392, Jun. 2010.

[11] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, W. A. Rosenblith, Ed., Cambridge, MA: MIT Press, 1961, pp. 217–234.

[12] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, Apr. 2002, Published online: March 18, 2002.

[13] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[14] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, 2006.

[15] A. Salles, K. M. Bohn, and C. F. Moss, "Auditory communication processing in bats: What we know and where to go," *Behavioral Neuroscience*, vol. 133, no. 3, pp. 305–319, Jun. 2019, Epub 2019 May 2.

[16] H. M. Ter Hofstede and J. M. Ratcliffe, "Evolutionary escalation: The bat–moth arms race," *Journal of Experimental Biology*, vol. 219, no. 11, pp. 1589–1602, 2016.

[17] P. I. M. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," in *IPO Symposium on Hearing Theory*, Eindhoven, the Netherlands, 1972, pp. 58–69.

[18] B. Grothe and M. Pecka, "The natural history of sound localization in mammals – a story of neuronal inhibition," *Frontiers in Neural Circuits*, vol. 8, Oct. 2014.

[19] S. M. Woolley and C. V. Portfors, "Conserved mechanisms of vocalization coding in mammalian and songbird auditory midbrain," *Hearing Research*, vol. 305, pp. 45–56, Nov. 2013.

[20] M. Steinschneider, K. V. Nourski, and Y. I. Fishman, "Representation of speech in human auditory cortex: Is it special?" *Hearing Research*, vol. 305, pp. 57–73, 2013.

[21] J. Hurford, "Origin and evolution of language," in *Encyclopedia of Language Linguistics*, K. Brown, Ed., Second Edition, Oxford: Elsevier, 2006, pp. 91–98, ISBN: 978-0-08-044854-1.

[22] A. J. Oxenham, "How we hear: The perception and neural coding of sound," *Annual Review of Psychology*, vol. 69, pp. 27–50, Jan. 4, 2018.

[23] G. Buchsbaum, "The possible role of the cochlear frequency-position map in auditory signal coding," *The Journal of the Acoustical Society of America*, vol. 77, no. 2, pp. 573–576, Feb. 1985.

[24] M. J. Wohlgemuth and C. F. Moss, "Midbrain auditory selectivity to natural sounds," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 9, pp. 2508–2513, Mar. 1, 2016.

[25] S. Park, A. Salles, K. Allen, C. F. Moss, and M. Elhilali, "Natural statistics as inference principles of auditory tuning in biological and artificial midbrain networks," *eNeuro*, vol. 8, no. 3, ENEURO.0525–20.2021, Jun. 16, 2021.

[26] T. Görföl *et al.*, "Chirovox: A public library of bat calls," *PeerJ*, vol. 10, e12445, 2022.

[27] T. Jiang, J. Feng, K. Sun, and J. Wang, "Coexistence of two sympatric and morphologically similar bat species rhinolophus affinis and rhinolophus pearsoni," *Progress in Natural Science*, vol. 18, no. 5, pp. 523–532, 2008, ISSN: 1002-0071.

[28] L. Zhang *et al.*, "Recent surveys of bats (mammalia: Chiroptera) from china. i. rhinolophidae and hipposideridae," *Acta Chiropterologica*, vol. 11, no. 1, pp. 71–88, Jun. 2009.

[29] M. E. Bates, J. A. Simmons, and T. V. Zorikov, "Bats use echo harmonic structure to distinguish their targets from background clutter," *Science*, vol. 333, no. 6042, pp. 627–630, Jul. 2011.

[30] M. D. Skowronski and M. B. Fenton, "Quantifying bat call detection performance of humans and machines," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 513–521, 2009.

[31] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[32] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.

[33] N. Tomašev and M. Radovanović, "Clustering evaluation in high-dimensional data," in *Unsupervised learning algorithms*, Springer, 2016, pp. 71–107.

[34] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Journal of Machine Learning Research*, vol. 22, no. 201, pp. 1–73, 2021.

[35] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[36] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in english," *The Journal of the Acoustical Society of America*, vol. 32, no. 6, pp. 693–703, 1960.

[37] C. Burgin, *Rhinolophidae*, Oct. 2019.

[38] P. Honeine, "Entropy of overcomplete kernel dictionaries," *Bulletin of Mathematical Sciences and Applications*, vol. 16, Nov. 2014.

[39] V. Papyan, Y. Romano, and M. Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 2887–2938, Jan. 2017, ISSN: 1532-4435.

[40] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.

[41] B. A. Olshausen, "Highly overcomplete sparse coding," in *Human Vision and Electronic Imaging XVIII*, B. E. Rogowitz, T. N. Pappas, and H. de Ridder, Eds., ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 8651, Mar. 2013.

[42] D. de Groot, *Sparse dictionary learning and matching pursuit code*, https://github.com/D1mme/rp_auditory_kernels/tree/main, Open-source code, accessed 2025-06-21, 2025.

[43] A. Savova, *Bat call data preprocessing and analysis notebooks*, https://data.4tu.nl/private_datasets/HVMmIz6HFsWkl7qZ2UXMiYN6M8CDJfnZAg3pBKmKdUY, Publicly available code, accessed 2025-06-21, 2025.

[44] B. McFee *et al.*, "Librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, Citeseer, 2015, pp. 18–25.

[45] B. McFee *et al.*, *Librosa/librosa: 0.11.0*, version 0.11.0, Mar. 2025.

[46] C. R. Harris *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[47] P. Virtanen *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[48] R. Gommers *et al.*, *Scipy/scipy: Scipy 1.16.0rc2*, version v1.16.0rc2, Jun. 2025.

[49] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[50] T. scikit-learn developers, *Scikit-learn*, version 1.6.0, Dec. 2024.

[51] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61.

[52] T. pandas development team, *Pandas-dev/pandas: Pandas*, version v2.2.3, Sep. 2024.

[53] L. Malfait, J. Berger, and M. Kastner, "P.563—the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006. DOI: 10.1109/TASL.2006.883177.

[54] T. M. D. Team, *Matplotlib: Visualization with python*, version v3.10.1, Feb. 2025.

[55] M. D. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160 018, 2016.

[56] W. Metzner, "An audio-vocal interface in echolocating horseshoe bats," *The Journal of Neuroscience*, vol. 13, no. 5, pp. 1899–1915, 1993.

Appendix A
Learned Dictionaries

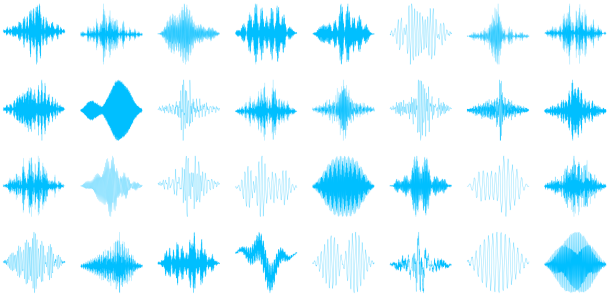

Fig. 10: **Learned Dictionary 1.** Kernels, initialized at 400 samples and trained on a combination of social and echolocation calls.

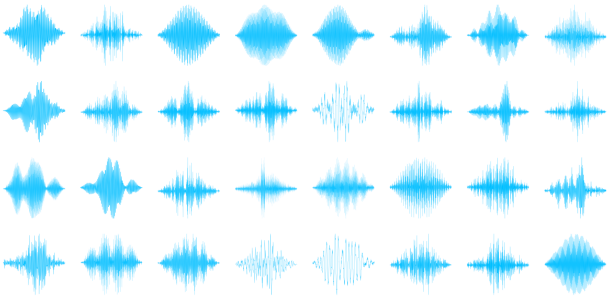

Fig. 11: **Learned Dictionary 2.** Kernels, initialized at 400 samples and trained on echolocation calls only.

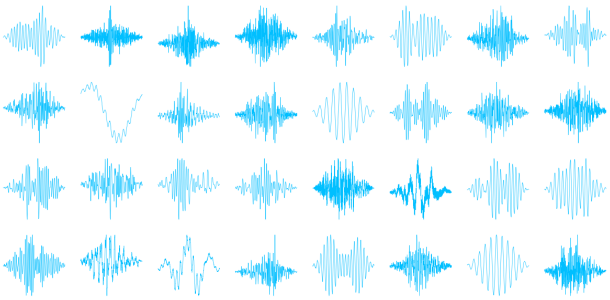

Fig. 12: **Learned Dictionary 3.** Kernels, initialized at 400 samples and trained on social calls only.

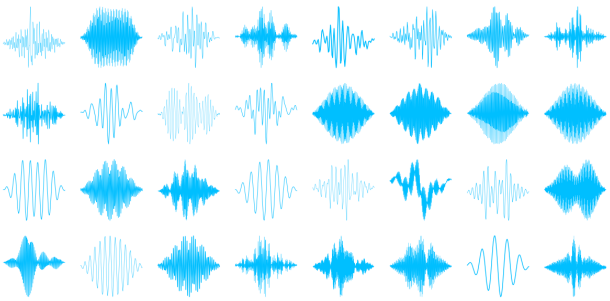

Fig. 13: **Learned Dictionary 4.** Kernels, initialized at 100 samples and trained on a combination of social and echolocation calls.

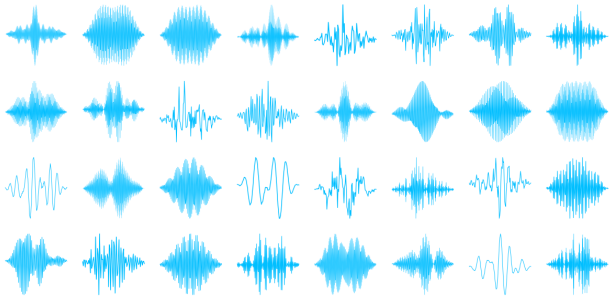

Fig. 14: **Learned Dictionary 5.** Kernels, initialized at 100 samples and trained on echolocation calls only.

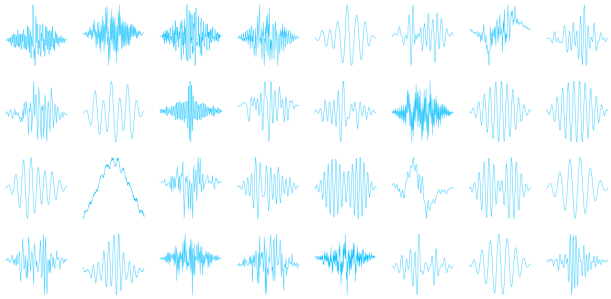

Fig. 15: **Learned Dictionary 6.** Kernels, initialized at 100 samples and trained on social calls only.

Appendix B
Kernel Activations

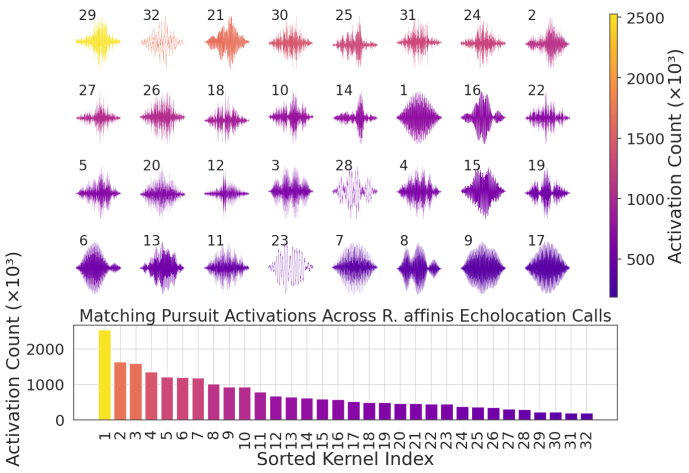

Fig. 16: **Kernel activation counts across high-fidelity reconstructions** Shows the relative kernel usage for reconstructions with an encoding rate of 18000 kernels/s. Notably, the most frequently used kernels exhibit lower centroid frequency and broad bandwidth, which hints that they reconstruct FM sweeps and background noise.

Appendix C
All Clusters

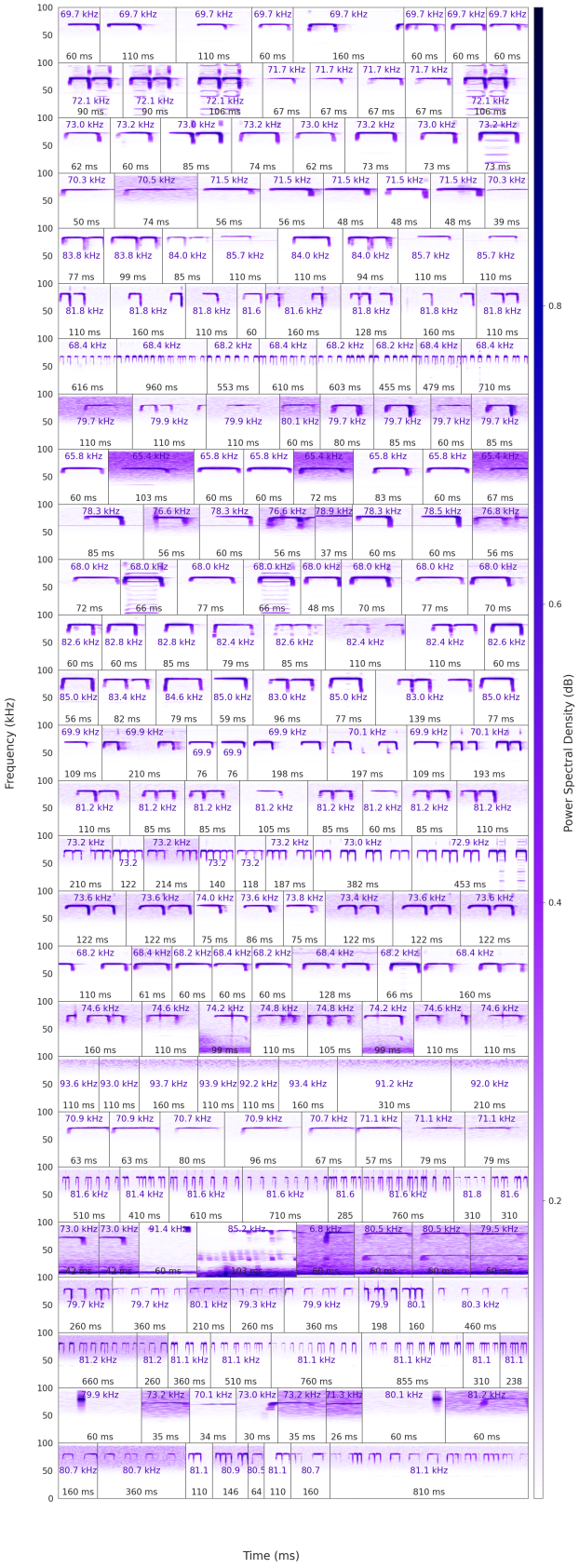

Fig. 17: **Clustering results.** The top eight most central recordings per cluster (row) are shown. Clusters 20, 23 and 26 seem to be grouping together empty recordings and irregularly-shaped calls, respectively.