Associative Syntax and Maximal Repetitions reveal context-dependent complexity in fruit bat communication

Luigi Assom
Alumnus, Department of Computer and Systems Sciences, Stockholm University
Stockholm, Sweden
luigi.assom@gmail.com

Abstract

This study presents an unsupervised method to infer discreteness, syntax and temporal structures of fruit-bats vocalizations, as a case study of graded vocal systems, and evaluates the complexity of communication patterns in relation with behavioral context. The method improved the baseline for unsupervised labeling of vocal units (i.e. syllables) through manifold learning, by investigating how dimensionality reduction on mel-spectrograms affects labeling, and comparing it with unsupervised labels based on acoustic similarity. We then encoded vocalizations as syllabic sequences to analyze the type of syntax, and extracted the Maximal Repetitions (MRs) to evaluate syntactical structures. We found evidence for: i) associative syntax, rather than combinatorial (context classification is unaffected by permutation of sequences, $F1 > 0.9$); ii) context-dependent use of syllables (Wilcoxon rank-sum tests, *p-value < 0.05*); iii) heavy-tail distribution of MRs (truncated power-law, exponent $\alpha < 2$), indicative of mechanism encoding combinatorial complexity. Analysis of MRs and syllabic transition networks revealed that mother-pupil interactions were characterized by repetitions, while communication in conflict-contexts exhibited higher complexity (longer MRs and more interconnected vocal sequences) than non-agonistic contexts. We propose that communicative complexity is higher in scenarios of disagreement, reflecting lower compressibility of information.

1 Introduction

Quantifying communication complexity in species with graded vocal systems remains a key challenge. We improved and extended an unsupervised pipeline to infer repertoire and syntax from vocalizations, applying it to fruit bats as a case study. We propose Maximal Repeats (MRs) as a novel metric to capture combinatorial complexity, extending variables of communication complexity rooted in information-theory to avoid non-independency between communication and sociality, which is a circularity pitfall in the social complexity hypothesis for communication complexity (SCHCC) [8].

Current methods face limitations. Sainburg et al. [13] [10] use manifold learning [11] to cluster vocal units, but this approach assumes discrete systems with clear unit boundaries and struggles with the continuous, graded vocalizations of species like fruit bats. Zhang et al. [15] analyze syntax of horseshoe bats using behavioral classifiers discriminating between aggressive and distressing calls, but require ground-truth syllable labels from experts, limiting scalability to other behavioral contexts and to other species.

Our work addresses two research questions:

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

- RQ1: How does dimensionality reduction affect unsupervised clustering on manifold learning for quantifying size and diversity of the repertoire ?
- RQ2: How do syntax and temporal structure encode contextual information?

To answer these, we first refined the method of Sainburg et al. [13] by inspecting how dimensionality reduction of mel-spectrograms affects clustering in manifold learning for the unsupervised labeling task; then, we used the labels to encode vocalizations as sequences and engineer features of a behavioral classifier, based on Zhang et al. [15], to test if order of syllables affect classification (i.e. compositional or associative type of syntax [14]). We extended the work with sequence analysis and introduce MRs - to our knowledge novel to animal communication - as a variable to analyze combinatorial complexity, motivated by their application in computational linguistics [4] and inspired by the analogical problem of how limited repertoires encode complex information in genetics (e.g. nucleotides in DNA sequences for protein expressions).

We used the fruit bat dataset [9] as a case study because it can be compared with the clustering baseline in [13] and because their authors provide domain reference of repertoire size and syntax-type for evaluating our unsupervised results [1].

Our contributions are:

1. A refined unsupervised pipeline for repertoire quantification in graded vocal systems, improving upon [13] and yielding results consistent with expert knowledge [1].
2. An analysis of context-dependent syntax, adapting the method of [15] to use syllables automatically labeled in multiple context-dependent repertoires.
3. The novel application of Maximal Repeats to animal communication, providing evidence for heavy-tailed distributions and proposing MR length as a metric of complexity.
4. Findings suggesting higher communicative complexity (longer MRs) in conflict behaviors versus cooperative ones.

We anticipate the limitation that the terms "conflictual" or "cooperative" are our interpretations of the behavioral annotation in the original dataset [9].

2 Background and Motivation

Discreteness in graded vocal systems A key challenge in deciphering animal communication is identifying the linguistic units relevant to the species, addressing discreteness as basis for syntax [2]. A method for unsupervised labeling clusters spectrogram representations of acoustic segments through manifold learning [13] [10], assuming that: i) vocal units are characterized by independent time-frequency features (e.g. clear unit boundaries as in discrete vocal systems); ii) acoustic similarity between units is relevant to the species [13]. However, in graded vocal systems the time-frequency features overlaps between syllables, and thus clustering performance degrade [8]. Poor efficacy on fruit bat vocalizations and other graded systems (e.g. mice, human phonemes) [13], prompted our study on how dimensionality reduction of input spectrograms may mitigate the challenges posed by vocal gradation.

Communication complexity & social complexity. The social complexity hypothesis for communicative complexity (SCHCC) recommends using quantitative information-theoretic metrics, such as the number of signaling units and the variety of their assembling patterns, to gauge communication complexity and to mitigate risks of non-independence between sociality and communication variables [8]. However, even if defining units through information theory helps to align definitions across fields (e.g. biology; linguistics), quantifying information does not necessarily imply meaning [6].

Syntax as systems conveying meaning. Syntax and temporal organization are considered systems that convey meaning [3]. The combination of units can form associative (order-independent) or combinatorial (order-dependent) syntax, which can yield to recursion or idiomatic types [14]. Heavy-tail distributions of signals (e.g. Zipf's law) are clues of compression mechanisms that minimize communication costs and imply rudimentary syntax [5]. Shannon entropy is a robust estimator of Zipf's power-law coefficient to assess repertoire variability (i.e. potential information content) [7], measuring uncertainty in repertoire draws [8]. However, it doesn't fully capture long-range dependencies and combinatorial capacity of a system.

Maximal Repeats: A metric for combinatorial complexity. These dependencies can be explored in terms of information-decay. Comparative studies show that in birdsong and human speech information decay is exponential for short sequences (e.g. notes) but follows power laws for longer ones (e.g. syllables or phrases); in humans, this mechanism predates the acquisition of language [12], implying potential common aspects across species phylogenetically distant. Maximal repetitions (MRs) offers a complementary tool. In computational linguistics, the scaling of MRs is mathematically linked to block entropy (entropy of n-grams); in written texts, the MR length was found to grow as a power of the logarithm of text length, consistent with Hilberg's conjecture, which posits that block entropy grows sub-linearly as a power-law and supports the view that natural language possesses strong long-range dependencies and is highly compressible [4]. We did not find studies employing MRs in animal communication and propose its use for estimating the combinatorial capacity of a system, thus extending variables of communication complexity in [8].

3 Methodology

We designed two experiments: unsupervised labeling to infer repertoire size and diversity (addressing RQ1); behavioral classification and statistical analysis of syllabic sequences to infer syntax type and temporal structures across behaviors (addressing RQ2).

Dataset. We used the annotated fruit bat vocalization dataset from [9], featuring 41 specimens with emitter, addressee, and behavioral context labels. Vocal units were automatically segmented by the authors. We analyzed vocalizations from the contexts: Mating Protests; Fighting; Threat-like; Biting; Feeding; Grooming; Kissing; Isolation (meaning mother-pup interactions). Contexts labeled as: Generic, Sleeping (utterances in sleeping area), or Unknown, were excluded due to ambiguity.

3.1 Size and Diversity of Repertoire

This experiment evaluated how the dimensionality of mel-spectrograms affects unsupervised clustering performance for quantifying the repertoire. The pipeline follows [13]: spectrograms (or their representations from autoencoders (AEs) [11]) are projected into a low-dimensional space with Uniform Manifold Approximation and Projection (UMAP), and then clustered with Hierarchical Density-Based Spatial Clustering (HDBSCAN).

We systematically varied the dimensionality of input representations to explore the clustering performance on graded vocalizations, through: i) Spectrogram settings (probing extreme time–frequency trade-offs to test if separability of clusters stems more from time or frequency); ii) Dimensionality Reduction (using PCA on AEs latent representations of spectrograms, like in [1], and testing different AEs architectures); iii) Segmentation (comparing the original procedure, which segmented audio where the amplitude envelope is above a fixed noise floor [9], with Dynamic Threshold Segmentation, which estimates the noise floor dynamically and is helpful to isolate shorter sub-units [12]) (see: settings for audio pre-processing in Table 2 and comments in Fig 4b). This analysis was conducted on the top-5 emitters; the best configuration was scaled to the full dataset.

Evaluation. We used a two-tiered strategy due to the lack of ground-truth labels:

1. *Internal Validation:* Silhouette Score to measure HDBSCAN cluster consistency.
2. *Agreement with Acoustic Similarity:* We generated a proxy for ground truth: for each emitter, we computed a pairwise distance matrix using Dynamic Time Warping (DTW) on Mel-Frequency Cepstral Coefficients (MFCCs) and performed Agglomerative Clustering with a quantile distance threshold ($q = 0.05$). This yielded 27 ± 2 syllable types per emitter, consistent with known bat repertoire sizes [1][15]. We measured agreement between these acoustic labels and HDBSCAN labels using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

3.2 Type of syntax and temporal structures conveying contextual information

This experiment investigated: 1) syntax type (associative/combinatorial), 2) context-dependent syllable usage, and 3) the distribution of Maximal Repeats (MRs). We tested three null hypotheses:

HP1$_0$: *Syllable order does not affect context classification.*

Method: We replicated the Random Forest (RF) in [15] to classify behavior based on

features from syllabic sequences (see predictors and their importance in Table 1, Fig 3). Unlike the original work, we used syllables from our unsupervised labels and extended the analysis to multiple behavioral classes.
Evaluation: Comparison of $F1 - scores$ between permuted and original sequences.

HP2$_0$: *Syllable usage is identical across behaviors.*
Evaluation: Wilcoxon rank-sum test on the syllable frequency distributions between pairs of behaviors.

HP3$_0$: *The distribution of maximal repetitions follows an exponential distribution.*
Method: We extracted MRs - the longest repeating subsequences - using a prefix-suffix tree algorithm. An exponential distribution would signify simple memory-less information decay (lower probability to observe longer sequences). A heavy-tailed distribution (e.g. power-law) would signify long-range dependencies [4].
Evaluation: Likelihood ratio test (exponential vs. power-law).

Finally, we compared the mean MR length across behaviors and qualitatively inspected the syllabic transition networks.

4 Results

Cluster quality addressed the first question (RQ1). Coarse-graining the temporal dimension of spectrograms from vocal units segmented with dynamic segmentation yielded the best results ($Silhouette > 0.5$, 95% assignment accuracy; see: Fig 1b, Appendix), identifying seven types of vocal units and improving the previous baseline that discriminated only two (i.e. utterances of mother and pups in Isolation and utterances between adults in all the other contexts). The local dimensionality of the UMAP embedding, inspected with diagnostic tools, is visible in Fig 2 (Appendix) along with the spectrogram settings used. The acoustic similarity proxy (Agglomerative Clustering on DTW distance) yielded an average of 27 ± 2 syllable types per emitter, consistent with known fruit bat repertoire sizes [1] [15]. The agreement between this proxy and our best HDBSCAN clustering, using mel-spectrograms retaining higher dimensionality as in the original experiment, was moderate (Mean ARI = 0.12 ± 0.01, Mean NMI = 0.30 ± 0.01), and suggested a repertoire of 14 syllables.

Results on syntax and temporal structure (RQ2) are as follows:

Syntax Type (HP1). The permutation test revealed that syllable order did not affect classification performance ($F1 - score > 0.9$ for both original and permuted sequences). Failing to reject $HP1_0$ supports an *associative* rather than combinatorial type of syntax, consistent with findings in [1].

Syllabic distribution (HP2). Syllable distribution was significantly different between Isolation and other contexts ($p < 0.05$, Wilcoxon rank-sum test), aligning with observations in [1]. Although specific outcomes were dependent on the clustering methods defining the repertoire, we found no significant evidence to reject $HP2_0$ for the cooperative contexts of Feeding, Grooming, and Kissing in the majority of pairwise comparisons, suggesting more uniform syllable usage across these behaviors. Heatmaps of syllabic distribution also suggested that Emitters grew in the same colony may not have a different use of syllables.

Maximal Repeats Distribution (HP3). The likelihood ratio test rejected $HP3_0$ ($p < 0.05$). The distribution of MR lengths was best described by a truncated power-law ($\alpha = 1.79$), indicating a heavy-tailed distribution inconsistent with a memory-less process and instead indicative of long-range temporal structures, reflecting combinatorial capacity of syntactical patterns.

Behavioral Complexity through MRs and Networks. The average length of MRs was greater in conflict-related contexts (Mating Protest, Fighting, Threat-like) than in cooperative ones (see: Fig 5, Appendix). To further explore this complexity, we represented syllabic transitions as networks for each behavior. Quantitative analysis of these networks revealed a spectrum of structural properties: Conflict-related contexts exhibited network metrics indicative of a small-world architecture ($\omega \approx 0$), characterized by high local clustering (Avg $C > 0.4$) alongside efficient global connectivity; in contrast, cooperative contexts displayed metrics suggesting a more random, less structured network ($\omega > 0.5$) (see: Table 3, Fig 6a, Appendix). Qualitatively, graphs from the Isolation context showed simple repetitions of a specific syllable (see: Fig 4, Appendix), while graphs from conflict contexts revealed more interconnected, complex structures.

5 Conclusions & Discussion

We contributed with an unsupervised pipeline to quantify repertoire and syntax in a graded vocal system, using fruit bats as a case study. Our key finding is that communicative complexity, measured through Maximal Repeats (MRs) and network analysis, is higher in conflict contexts than in cooperative ones.

The finding that temporal compression aids cluster separation aligns with the nature of graded systems, where information is encoded in continuous acoustic modulation. We speculate that basic frequency-based utterances combine and are modulated in time to form more complex syllables, which are then assembled into sequences governed by combinatorial patterns (revealed by MRs) to convey behavioral meaning. We interpret our results through the lens of social complexity. Contexts like Mating Protest and Fighting likely may represent scenarios of social disagreement, requiring more complex signals to negotiate interactions. This is reflected in longer MRs with non-permuted counterparts, and small-world network structures within the syllabic transition graphs.

We propose the interpretation that higher-complexity observed in conflict-related communication may reflect lower compressibility of information conveying disagreement. We propose to test the use of MRs in other species as a proxy of combinatorial capacity.

Acknowledgments and Disclosure of Funding

This paper is adapted from work conducted during the author's Master's thesis. The author is grateful to Tim Sainburg for providing valuable feedback after the completion of that work. All interpretations and any errors remain the author's own.

The research presented in this paper received no dedicated external funding. The author acknowledges partial financial aid from the Earth Species Project and from the Department of Computer and Systems Sciences at Stockholm University to support the presentation of this work.

Code Availability

The repository associated with the original Master's thesis, on which this paper is based, is available at: `https://github.com/gg4u/decodingNonHumanCommunication`

The repository contains the unrefactored thesis implementation. A cleaner and updated version is under development.

References

[1] Yoni Amit and Yossi Yovel. Bat vocal sequences enhance contextual information independently of syllable order. *Iscience*, 26(4), 2023.

[2] Jacob Andreas, Gašper Beguš, Michael M Bronstein, Roee Diamant, Denley Delaney, Shane Gero, Shafi Goldwasser, David F Gruber, Sarah de Haas, Peter Malkin, et al. Toward understanding the communication in sperm whales. *Iscience*, 25(6), 2022.

[3] Mélissa Berthet, Camille Coye, Guillaume Dezecache, and Jeremy Kuhn. Animal linguistics: a primer. *Biological reviews*, 98(1):81–98, 2023.

[4] Łukasz Dębowski. Maximal repetitions in written texts: Finite energy hypothesis vs. strong hilberg conjecture. *Entropy*, 17(8):5903–5919, 2015.

[5] Ramon Ferrer i Cancho, Oliver Riordan, and Béla Bollobás. The consequences of zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society B: Biological Sciences*, 272(1562):561, 2005.

[6] Arik Kershenbaum, Daniel T Blumstein, Marie A Roch, Çağlar Akçay, Gregory Backus, Mark A Bee, Kirsten Bohn, Yan Cao, Gerald Carter, Cristiane Cäsar, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 91(1):13–52, 2016.

[7] Arik Kershenbaum, Vlad Demartsev, David E Gammon, Eli Geffen, Morgan L Gustison, Amiyaal Ilany, and Adriano R Lameira. Shannon entropy as a robust estimator of zipf's law in animal vocal communication repertoires. *Methods in Ecology and Evolution*, 12(3):553–564, 2021.

[8] Louise Peckre, Peter M Kappeler, and Claudia Fichtel. Clarifying and expanding the social complexity hypothesis for communicative complexity. *Behavioral Ecology and Sociobiology*, 73(1):11, 2019.

[9] Yosef Prat, Mor Taub, Ester Pratt, and Yossi Yovel. An annotated dataset of egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny. *Scientific data*, 4(1):1–7, 2017.

[10] Tim Sainburg and Timothy Q Gentner. Toward a computational neuroethology of vocal communication: from bioacoustics to neurophysiology, emerging tools and future directions. *Frontiers in Behavioral Neuroscience*, 15:811737, 2021.

[11] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 2021.

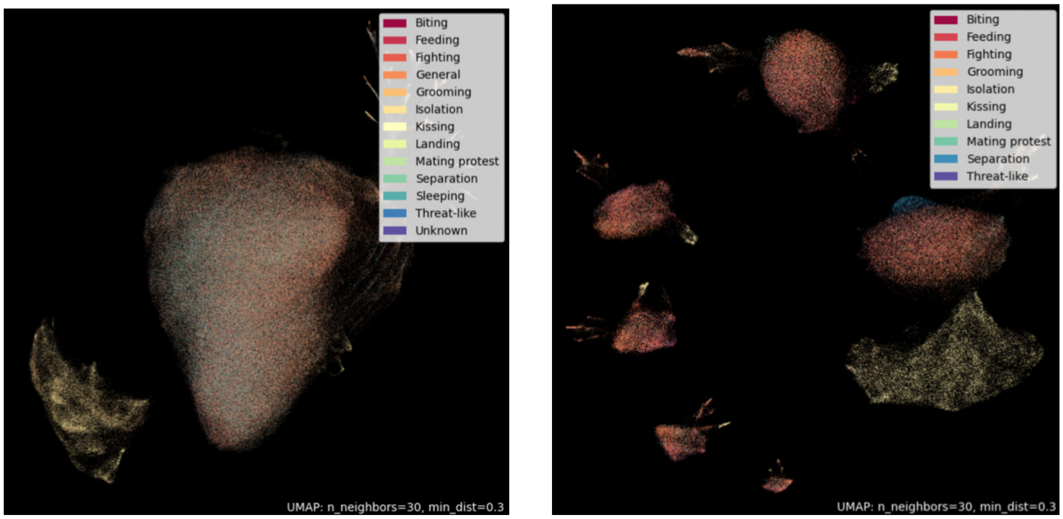
[12] Tim Sainburg, Brad Theilman, Marvin Thielk, and Timothy Q Gentner. Parallels in the sequential organization of birdsong and human speech. *Nature communications*, 10(1):3636, 2019.

[13] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228, 2020.

[14] Toshitaka N Suzuki, David Wheatcroft, and Michael Griesser. The syntax–semantics interface in animal vocal communication. *Philosophical Transactions of the Royal Society B*, 375(1789):20180405, 2020.

[15] Kangkang Zhang, Tong Liu, Muxun Liu, Aoqiang Li, Yanhong Xiao, Walter Metzner, and Ying Liu. Comparing context-dependent call sequences employing machine learning methods: an indication of syntactic structure of greater horseshoe bats. *Journal of Experimental Biology*, 222(24):jeb214072, 2019.

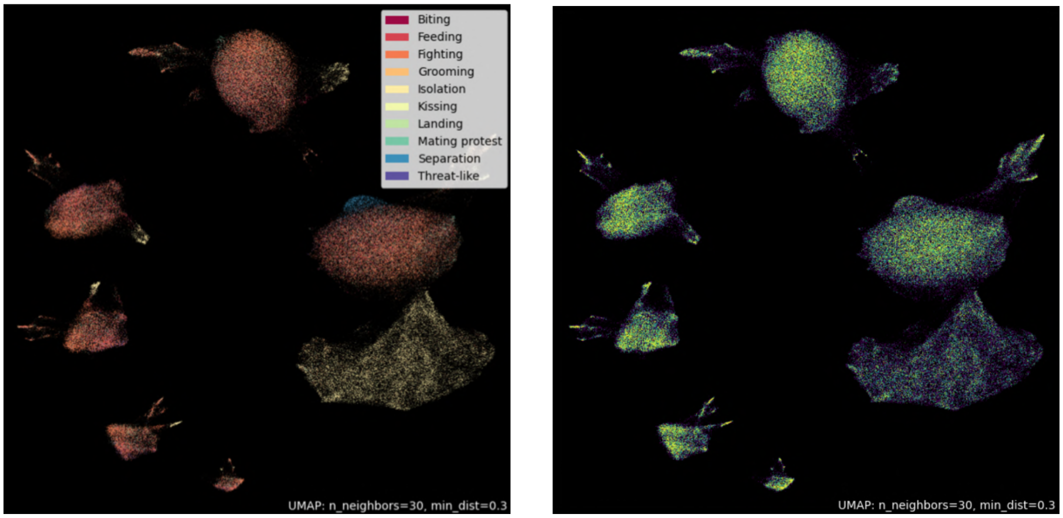
A Technical Appendices and Supplementary Material



(a) Benchmark results (replicated from [13]).

(b) Our improved results with refined pipeline.

Figure 1: Improved clustering quality of continuous-type vocalizations. The left panel shows the original benchmark results, which primarily separate isolation calls from adult vocalizations. The right panel demonstrates our improved clustering, which identifies seven distinct syllable types through optimized dimensionality reduction and segmentation techniques applied to the graded vocal system.



(a)

(b)

Figure 2: Diagnostic of the local dimensionality of manifold learning. Clustering obtained from Non-parametric UMAP applied on Mel-Spectrograms (6x32) preprocessed by Mel-filterbank (hop size equal to FFT length) and dynamic segmentation. Inputs: 152,578 data-points from all bats (41 individuals). Bluish colors represent the lowest local dimensionality, which corresponds to underdeveloped vocalizations from the Isolation context (i.e., simpler, more uniform spectrograms). Warmer colors (yellow/red) indicate regions of higher local dimensionality and greater acoustic complexity.

Table 1: Predictors used in ML classifiers – Adapted from: [15]

ID	Description	Formula
a	Syllable richness	Total number of syllable types in a sequence
b	Sequence length	Total number of any syllable-type in a sequence
c	Transition count	Total number of bi-gram transitions in a sequence
d	Linearity index	a/c
e	Contextual variety	Total number of transition types under the behavioral context
f	Sequence entropy	$H = -\sum_i p_i \log p_i$
g	Pattern commonness	$\prod_i p(s_i)$
h	Contextual transition strength	$\prod_i p(t_i)$
i	Versatility ratio	Ration between features: a/b
j	Transition uncertainty	Entropy of transition probabilities
k	Graph transition strength	Product of transitional probabilities (occurring in any context)
l	Local predictability	$p_{\text{cond}} \log(p_{\text{cond}})$
m	Global frequency weight	$p_{\text{trans}} \log(p_{\text{trans}})$
n	Conditional syllable chain	$\prod_i p_{\text{cond}}(s_i)$
o	Conditional bi-gram chain	$\prod_i p_{\text{cond}}(B_i)$
p	Transitional bi-gram chain	$\prod_i p_{\text{trans}}(B_i)$
q	2-step Markov predictability	$\prod_i p(B_i \mid B_{i-1})$
r	Sequence perplexity	$\left(\prod_{i=1}^{N} \frac{1}{p_{\text{cond},i}}\right)^{1/N}$

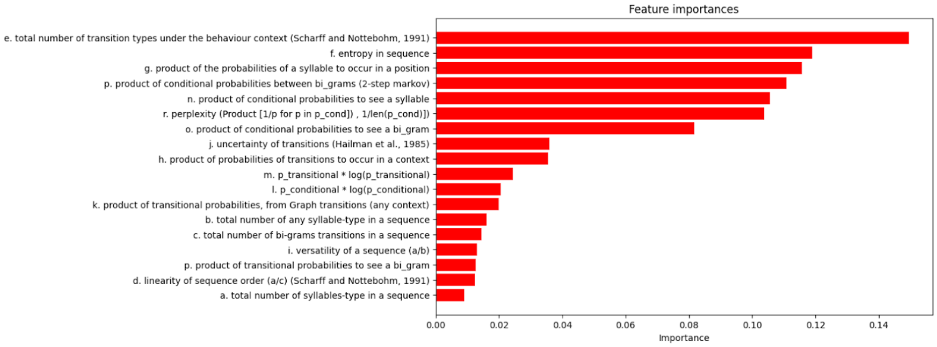
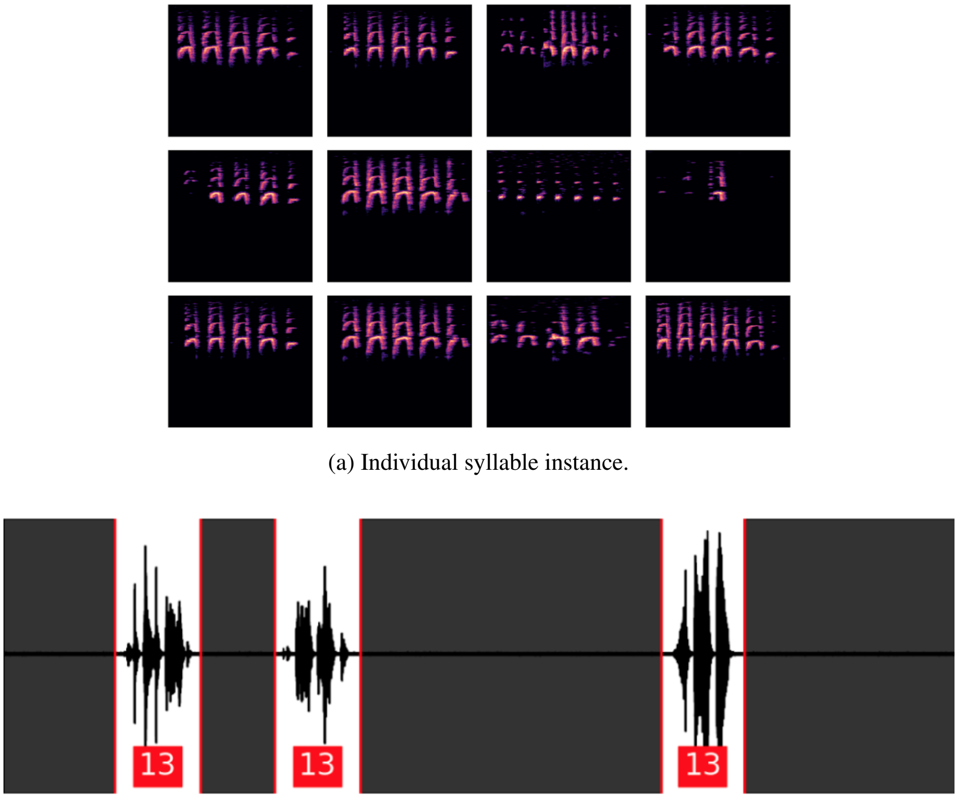


Figure 3: Importance of features used for the Random Forest classifier. Features representing richness of contextual syntax, unpredictability of sequences, commonness of patterns and strength of short transitions (respectively, features: e, f, g, p) account for about 50% of the total feature importance, suggesting a predominant temporal organization of short transitions and repetitive patterns.



(a) Individual syllable instance.

(b) Sequence of repetitive occurrences.

Figure 4: Syllable-type unique to the Isolation context (mother-pupil interactions), isolated through agglomerative clustering. (a) Randomized sampling of the syllable-type, displaying its uniform spectral structure. (b) A sequence of this syllable, demonstrating the characteristic repetition patterns consistent with underdeveloped vocalizations. *Note* – Unsupervised labeling in these figures used the acoustic segments from the original dataset [9], whose boundaries were computed by thresholding the amplitude envelope above a fixed noise floor. When using algorithms that estimate the noise floor dynamically (as in [12]), syllables could be further subdivided into smaller segments; in this example, the three bursts visible in the waveforms were separated into distinct sub-units.

Table 2: Parameters used for audio preprocessing.

Function	**Description and Best Practices**	**Settings**

Bandpass	Cutoff for low and high frequencies	low_freq = 256
		high_freq = 120000

Noise-Removal	Non-stationary noise removal	time_constant_s = 0.2
		time_mask_smooth_ms = 5
		stationary = False
		freq_mask_smooth_hz = 256

Pre-Emphasis	Emphasis high-frequencies	pre_emphasis = 0.97

Short-Time Fourier Transform (STFT)	STFT for power-to-decibel spectrograms. Use a n_fft of 8ms, with a window of 4ms and 1ms overlaps. Normalize with respect to median power values.	n_fft = 2048
		fmin = 256
		fmax = 120000
		hop_length = 256
		win_length = 1024
		sr = 250000

Mel-Frequency Cepstral Coefficients (MFCCs)	Use 64 mel-bins	as in STFT n_mels = 64

Dynamic Threshold Segmentation	Tokenize original audio segments into shorter sub-components	as in STFT
		db_delta = 5
		ref_level_db = 20
		pre_emphasis = 0.97
		min_silence_for_spec = 0.1
		max_vocal_for_spec = 1 (# second)
		min_level_db = -60 (# threshold of sound or noise)
		silence_threshold = 0.1
		verbose = True
		min_syllable_length_s = 0.01
		spectral_range = [2000, 60000]

MEL-Filterbank	Compute Log-MEL-Spectrograms. Use a MEL-filter bank to increase the frequency resolution of spectrograms to 4093 (fft_length // 2 + 1); map them into 32 mel_bins; increase the relative distance of decibels to 120 db	fft_size = 8192 (# samples per frame)
		hop_size = 8192 (# samples to step)
		fft_length = 8192 * 2 (# size of the FFT)
		n_mels = 32
		f_min = 500
		f_max = 120000

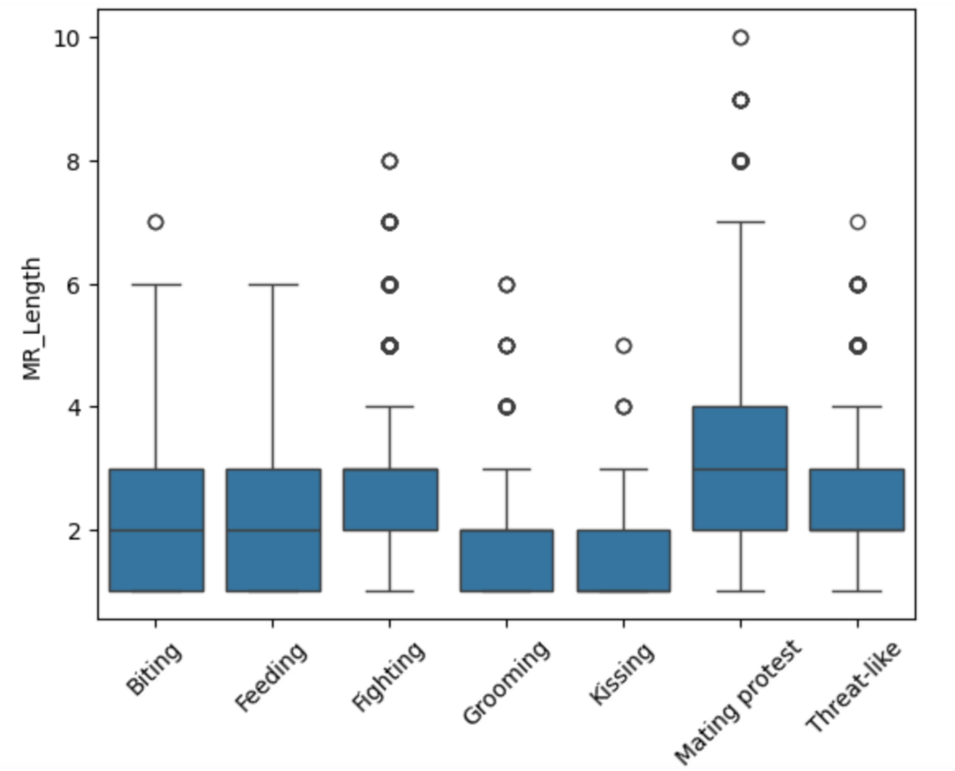
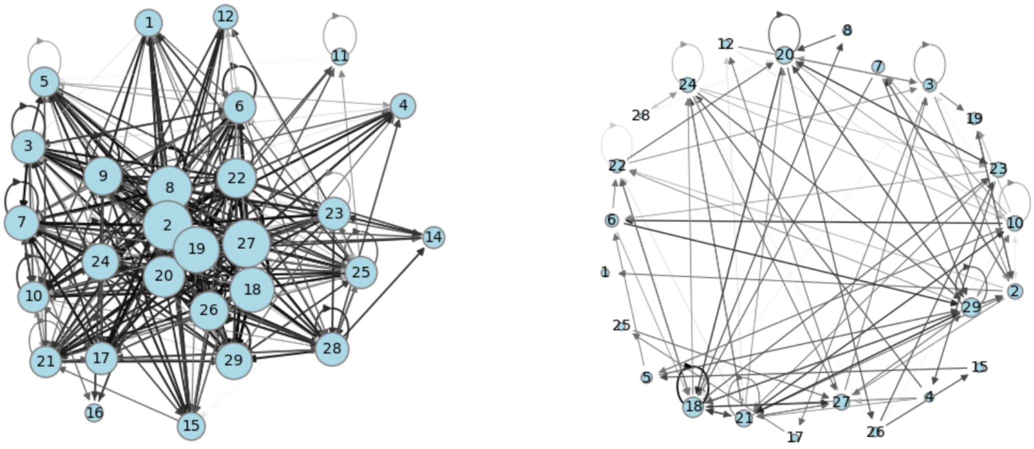


Figure 5: Distribution of Maximal Repetition (MR) lengths across behavioral contexts (sequences with at least 50 support). Conflict-related contexts (Mating Protest, Fighting, Threat-like) show heavier-tailed distributions with longer MRs, indicating more complex temporal structures and lower compressibility of information. Cooperative contexts (Feeding, Grooming, Kissing) exhibit shorter MR distributions, suggesting higher redundancy and more compressible communication patterns. The Isolation context shows a unique pattern dominated by short, repetitive sequences.



(a) Mating Protest context.

(b) Kissing context.

Figure 6: Examples of networks of syllabic transitions for Emitter ID 215 (syllables based on agglomerative clustering). These networks visually represent the transition probabilities between different syllable types within a specific behavioral context. The network structure for Mating Protest (a) is denser and more interconnected, indicative of higher complexity, contrasting with the sparser structure of the Kissing context (b).

Table 3: Graph metrics for syllabic transition networks across behavioral contexts for Bat#215. Metrics include sequence support (number of transitions), small-world coefficients (Sigma, Omega), maximal clique statistics, graph density, and average clustering coefficient (Avg C). A Sigma (σ) > 1 and Omega (ω) $\approx$ 0 indicates small-world structure.

Context	Support	σ	ω	# Big Clique	# All Clique	Density	Avg C
Biting	292	1.02	0.05	9	99	0.40	0.46
Feeding	96	0.84	0.53	4	42	0.15	0.13
Fighting	50	1.00	0.03	4	12	0.26	0.44
Grooming	48	1.15	0.65	4	35	0.11	0.09
Isolation	78	–	–	2	11	0.10	0.00
Kissing	48	0.86	0.63	4	39	0.13	0.12
Mating Protest	629	1.00	0.00	17	25	0.81	0.62
Threat-like	46	1.08	0.10	5	37	0.18	0.35

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The methodology describes extracting syntactical patterns through unsupervised labeling and testing hypotheses about their distribution and relationship to behavior. The claims match experimental results and corroborate theoretical conjectures in linguistics, but in animal communication. Aspiration goal is to provide metrics of communication complexity that can support experiments in ethology and AI models - such as synthetic evaluations of generative models that replicate the syntactical patterns observed here (not included in the paper).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of interpretability of labels is mentioned in the introduction, even if briefly. Limitations of results of clustering quality are described by reporting best results but also intermediate results depending on the unsupervised technique used, to inform the reader about limitations on internal validity.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We stated the assumptions of the unsupervised method we aimed to improve, reporting from the original work. Due to economy in the article, we did not include theoretical aspects linking block entropy, maximal repetitions and heavy tail distributions, which can be found in the paper that supported our motivation for using MR.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information for replicating the experiment is available on github repository and on the master thesis from which we extracted this work; settings and description mentioned in this paper should allow to replicate the core procedures of the experiments, although full details can only be found in the repository.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Access to data of the original dataset is available on Figshare, and we published a github repository with notebooks to run the experiments; all the steps (e.g. data preparation, settings) are reported in the supplemental material and repository.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [No]

Justification: Our work utilized experiments that required many steps, and thus many parameter settings, from data preprocessing, to testing parameters from dimensionality reduction of spectrograms, autoencoder architectures that we used. They are specified in the original thesis that motivated this paper. Full details ar provided as supplemental material (thesis and github are publicly accessible) but not in the appendix to make the submission compact.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: systematic errors, error bars in box plots, and statistical significance are reported.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: This comments are also included in the master thesis that motivated this work, publicly available for consultation, with details on computation time for each component of the experiments (e.g. each of the Autoencoders tested to reduce dimensionality of spectrograms).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: vocalizations of animals were recorded in lab settings in the original study from which we sourced the dataset, which is available to public for allowing reproducibility of results. The authors of the study were employed at another university at the time of conducting the experiments, but have no more affiliation (now working at another university). We were ansure which emails utilize to submit, we had to use emails from the current employer.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper does not address societal impact due to the economy of a short paper and due to the goal may be more a possible outcome than a direct impact. We envision the possibility to facilitate hypothesis testing communicatin complexity depending on behavior (SCHCC hypothesis), and extend applicability of the technique to inquire aspects as cognition through in many other species, or propose a pipeline for testing syntactical structures of synthetic vocalizations, but they are out of the scope of this paper. We do not expect negative impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We mentioned the reference to the dataset, available on open access. We mentioned the author of the two main baselines we utilized to run the experiment (i.e. the unsupervised clustering based on manifold learning, and random forest classifier).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The linked repository is released under MIT license, and contains the original implementation from the author's Master's thesis, which served as the basis for this work; however, since it is an archival work that needs to be refactored for clarity, and not yet fully compliant with the best practices indicated in `https://nips.cc/public/guides/CodeSubmissionPolicy`, the author considered "Not Applicable" as the appropriate answer.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.