WhaleLM: Finding Structure and Information in Sperm Whale Vocalizations and Behavior with Machine Learning

Pratyusha Sharma[1,2] **Shane Gero**[2,3]
Daniela Rus[1,2†] **Antonio Torralba**[1,2†] **Jacob Andreas**[1,2†]
[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology.
[2]Project CETI. [3]Department of Biology, Carleton University.
[†]These authors contributed equally to this work.

Abstract

Sperm whales (*Physeter macrocephalus*) communicate using patterned click sequences called codas. Whether there are any systematic patterns governing the structure of coda sequences, or how coda production influences group behavior, remain open questions. To answer these questions, we train neural sequence models ("sperm whale language models") on vocalization and behavior data from a population of sperm whales in the eastern Caribbean. By systematically manipulating models' training data and measuring changes in predictive power, we find that vocalizations exhibit order dependence and long-range dependencies on up to eight previous codas in an exchange. We additionally find that this structure encodes information about behavior: whales' current behavioral context and future actions are predictable with high accuracy from coda sequences. The methods developed for relating vocalization to behavior are general, and offer a flexible framework for using language models to investigate the structure and information content of unknown communication systems.

1 Introduction

Communication is a key characteristic of intelligence**???**. In humans, language allows us to share knowledge, coordinate actions, and establish social structures. Recently, modern language models (LMs)—neural sequence predictors trained to model the probability distribution of natural language text—have advanced our understanding of how efficiency and learnability constraints shape human languages **???** as well as scientific understanding of a number of other biological systems **??**. Humans are not the only animals that communicate to coordinate behavior; non-human organisms produce and perceive communicative signals in very different ways from humans, and many animal communication systems remain incompletely understood. Can neural sequence models aid and guide the scientific characterization of animal communication as well?

We use neural sequence models to characterize both the *structure* and *information content* of an animal communication system—specifically, to model communication and behavior in sperm whales (*Physeter macrocephalus*). Sperm whales exhibit a multi-level social structure **???**, coordinated group foraging and child-rearing behavior **???**, and a complex, socially learned communication system **???**. Sperm whale vocalizations consist of sequences of stereotyped, rhythmic click patterns called codas. Several recent studies have characterized codas' internal structure **?**, including with machine learning models **???**. But the patterns in which codas are combined into sequences, and their role in coordinating group behavior, are still not understood.

To obtain first answers to these questions, we train a collection of neural sequence models ("sperm whale language models") on several years of recordings from a population of sperm whales in the

eastern Caribbean, the EC1 clan. Models receive as input a "conversation history" (a sequence of vocalizations by one or more whales) and predict either the whales' future vocalizations, present behavior, or future behavior. By systematically manipulating the data these models are trained on (e.g. by restricting the length of the conversation history they have access to, or masking specific acoustic features of individual codas), and measuring the impact of these manipulations on predictive power, we can identify specific features of vocalizations that are predictive of future vocalizations or behavior.

We first show that vocal exchanges between sperm whales in our sample have complex internal structure: coda production exhibits long-range statistical dependencies, and is sensitive to the identity and ordering of the preceding 8 codas (up to 30 seconds in the past)—including not only codas produced by the vocalizing whale, but also those produced by conspecifics. Next, we show that these exchanges contain information about behavior: sequence models can predict both whales' present behavioral context and future actions from their vocalizations alone. By inspecting models' predictions, we identify a specific, multi-coda motif that is predictive of future diving when made by all whales present in an exchange. While past work has found specific individual codas that are predictably associated with (current) behavior **?**, these results provide the first evidence that some sperm whale vocalizations exhibit long-range structure above the single-coda level, and the first evidence that this structure encodes information about *future* behavior. As recently highlighted by Rutz et al. **?**, machine learning models hold great promise for advancing scientific understanding of communication systems across the tree of life, and the approach to sequence-model-guided discovery we present here can serve as a precursor to interactive playback experiments by enabling offline identification of informative features and predictive relationships—offering a flexible framework for using the tools of artificial intelligence to study complex biological systems.

2 Method

Using the Dominica Sperm Whale66 Project dataset (see supplementary for details), we train a neural sequence model to predict codas and behaviors from preceding coda sequences. We then examine the behavior of this model to determine what codas and behaviors are predictable and what features of vocalizations support these predictions.

In this paper, we are specifically concerned with sequence models parameterized by deep neural networks, which encode and then predict by first embedding input data in a high-dimensional vector space, then applying alternating linear and non-linear transformations to these token representations mapping them to a distribution over possible outputs. The parameters of the neural network are learned from data as described below. We train two families of neural sequence models, one of which predicts future coda production, and the other of which predicts a vocalizing whale's present or future behavior given coda sequences. We begin by formally defining these networks and their training objective, then describe how they can be used to analyze the structure and information content of sperm whale vocalizations.

Model training and evaluation: Our dataset (denoted $\mathcal{D}$) comprises a sequence of coda exchanges (each denoted e_i), each of which in turn comprises a sequence of codas (c^i_j), where c^i_j is the jth coda in the ith exchange. We 'tokenize' call sequences by assigning every coda a discrete identifier that captures the four defining coda features (rhythm, tempo, rubato, and ornamentation) previously described by Sharma et al. **?**, as well as the time elapsed since the preceding coda in the exchange, and the identity of the vocalizing whale. Each exchange e_i also takes place in a specific behavioral context (e.g. the beginning of a foraging dive or a period of socialization near the surface of the water; see Fig. **??**). We denote by b_i the behavioral context for the exchange e_i. Refer to Section 1.1 in the supplementary material for additional details on tokenization.

For each prediction task, we construct an encoder–decoder LSTM **?**, a type of recurrent neural network, that maps from a sequence of input codas to a distribution over next codas or behavior labels. To produce an accurate predictor, we train the network to imitate real coda sequences.

To do so, we first divide the dataset $\mathcal{D}$ into a training set $\mathcal{D}^{\text{train}}$ and a test set $\mathcal{D}^{\text{test}}$. When training models for coda prediction tasks, we choose parameters to maximize the log-likelihood $\sum_{i=1}^{|\mathcal{D}^{\text{train}}|} \sum_{j=1}^{|e_i|} \log p(c^i_j \mid \ldots, c^i_{j-2}, c^i_{j-1}; \theta)$, where $p(y \mid x; \theta)$ denotes the probability that the LSTM with parameters θ assigns to the output y given the input x. Intuitively, this choice of θ encourages

the model to assign a high probability to sequences that appeared in the training data and a low probability to all other sequences. When training models for behavior prediction, we optimize $\sum_{i=1}^{|\mathcal{D}^{\text{train}}|} \sum_{j=1}^{|e_i|} \log p(b_i \text{ or } b_{i+1} \mid \ldots, c_{j-2}^i, c_{j-1}^i)$, which encourages the model to assign high probability to the true behavioral context of training vocalizations. As described below, our experiments vary both the size of the context window and the features used to distinguish input codas.

As is standard when studying neural sequence models, we evaluate coda-prediction models according to their perplexity $\exp\{-\frac{1}{N}(\sum_{i=1}^{|\mathcal{D}^{\text{test}}|} \sum_{j=1}^{|e_i|} \log p(c_j^i \mid \ldots, c_{j-2}^i, c_{j-1}^i)\}$, where N denotes the total number of codas in $\mathcal{D}^{\text{test}}$. Perplexity is simply the exponentiated average log-likelihood per token. Example predictions are shown in Fig. **??**E. We evaluate behavior-prediction models according to their accuracy (whether the behavior assigned the highest probability matches the ground-truth behavior in the dataset). Averages for both evaluation metrics are computed over the full DSWP dataset using k-fold cross-validation ($k = 10$). Each cross-validation split holds out recordings from a distinct day for evaluation, and trains a sequence model on the remaining days, ensuring that models are evaluated on their ability to extrapolate to novel interactions. Refer to Section 1.2 in the supplementary materials for additional training details.

Experimental method: To understand what features of coda sequences contain information about future vocalizations or behavior, we repeat the training procedure described above while systematically varying the information available to the model. For example, to determine whether the next coda choice is solely influenced by the single preceding coda, we train two models, one of which conditions on a sequence of $n > 1$ input codas $c_{j-n}, \ldots, c_{j-2}, c_{j-1}$, and the other of which conditions on only the most recent coda c_{j-1}. If these two models exhibit similar perplexity on a held-out set, we conclude that longer contexts contain no additional information that is usable for prediction; if the long-context model performs better, we conclude that there is usable information in codas beyond the most recent. Formally, this procedure may be interpreted as measuring the transfer entropy **?** or $\mathcal{V}$-information **?** from the context to the next coda.

This same methodology can be used to evaluate the informativeness of individual *features* of codas; for example, the "rhythm" feature described by Sharma et al. **?**. To do so, we train one baseline model on full coda sequences as above, and one in which each *input* coda's identity is determined only by its tempo, rubato, and ornamentation features. For both models, we continue to identify *output* codas as before with all four features (to ensure predictions between the two models are directly comparable). If the second, "ablated" model produces less accurate predictions, we may conclude that the rhythm feature contains information useful for prediction.

Importantly, this method for quantifying the informativeness of features is self-supervised: it requires only communication (or communication and behavior) data, without additional labels or interventions from researchers. Below, we use it to identify aspects of sperm whale vocalizations that carry information about future vocalizations, as well as current and future behavior.

Results and Discussion

We first use neural sequence models to study the internal structure of coda sequences. To do so, we train next-coda prediction models while removing various sources of information from the input and measuring the effect on predictivity. Results are shown in Fig. **??**.

Vocalizations exhibit long-range dependencies and order-sensitivity

First, we investigate the effect of communicative context by studying how coda *sequencing* influences call production. As motivation, human languages exhibit complex structure in which words and morphemes must be combined and ordered in specific patterns to convey precise meanings: the sentence *The dog in the park was playing* is meaningful, while the sentence *Dog playing park in the the was* is not. Moreover, natural languages exhibit non-local statistical dependencies: if *dog* were replaced by *dogs*, then *was* would need to be replaced by *were* for the sentence to remain grammatical, even though these words are not adjacent to each other in the surface order of the sentence. Sequence-level structure is by no means unique to humans: past studies have shown that songbirds **??**, humpback whales **?**, and primates **???** also produce vocal sequences that exhibit statistical regularities over long distances.

In Fig. **??**B(i), we evaluate the informativeness of *call order*. We hold the context window fixed at the past two codas as well as a longer context of eight codas and train LMs on versions of the data in which these input codas arrive in (1) their natural order, or (2) are replaced with a uniformly random permutation of the input. In both cases, models are trained to predict future calls in their natural order. Removing order information from inputs increases perplexity (i.e. decreases predictivity) by up to 22.7%, indicating that ordering information is crucial for predicting future calls (Wilcoxon Sign-Ranked Test, sum of ranks $= 55$, p $= 0.001$).

In Fig. **??**B(ii), we evaluate the informativeness of *context length* by varying the number of preceding codas available to the LM during training and prediction. Short contexts (containing 6 or fewer preceding codas) substantially reduce the predictability of future codas (by up to 20.6%) relative to long contexts (Wilcoxon Sign-Ranked Test, sum of ranks $= 54$, p $= 0.002$). Together with the ordering information, these results indicate that the patterns governing call production depend on the ordering of a large number of preceding calls. The sperm whale communication system is sensitive to call order and exhibits statistical dependencies across calls separated by as much as 30 seconds (a typical duration for an 8-coda sequence).

Predicting vocalizations requires complex models and fine-grained coda representations

Having shown that sperm whale call production is sensitive to call history and call order, we next investigate which features within each individual call influence call production, and how expressive sequence models must be to capture this influence.

Past work **?** previously proposed to analyze codas as a combination of four features termed *rhythm, tempo, rubato, and ornamentation*. In our first experiment, we evaluate which of these features are needed to predict future vocalizations. To do so, we systematically ablate information about these features (one at a time) from the input, while leaving the model's output space unchanged. For example, to evaluate the role of the rhythm feature, we assign the same input token identifier to all codas that differ only in their rhythm type: for example, 4R/5 (where 4R denotes the rhythm category and 5 denotes the tempo category), 5R/5 and 1+1+3/5 codas are all mapped to the same input token ID. However, output coda IDs are kept unchanged to ensure all models make predictions over the same set of possible output tokens. We then compare the perplexity of this model to models with access to all features. If removing rhythm information from the input increases perplexity, we may conclude that this feature carries information about future vocalizations. Results of this experiment for rhythm and tempo, are shown in Fig. **??**C. It can be seen that, when considering a communicative context of only two codas, both features are predictive. Interestingly, with longer contexts, ablating the rhythm feature no longer meaningfully alters predictivity, indicating that it may be somewhat redundant with information conveyed by changes in the other three features over multiple time steps. Corresponding experiments for ornamentation and rubato are provided in Section 1.4 in the supplementary material.

The preceding experiments have all used a specific recurrent neural sequence model for prediction. Our final coda prediction experiments evaluate the role that the choice of sequence model plays in these findings. To do so, we compare the predictive accuracy of the model in Fig. **??**D with four other neural and non-neural sequence models: (a) a **linear** model in which the input sequence is represented by concatenating indicator features for each input coda in order, then mapped directly to a distribution over next codas, (b) a **multi-layer perceptron** which uses the same input representation as the linear model, but passes these inputs through a neural network with an additional hidden layer, (c) an **n-gram** model which predicts next items by counting empirical frequencies of different input coda sequences, and (d) a **LSTM model with attention**, which augments the sequence-to-sequence LSTM model with a single attention head, as in a pointer-generator network **?**. See Section 1.3 in the supplementary materials for implementation details of all models. Results are shown in Fig. **??**D: expressive models with explicit sequential structure predict the next call in a sequence more accurately. Surprisingly, n-gram models perform nearly as well as recurrent models, while models based on a (fixed, non-recurrent) input feature representation obtain significantly worse perplexities. The addition of an attention mechanism does not substantially alter predictivity. These results show that vocalizations can be predicted accurately by a range of learned sequence models, but that recurrent neural models enjoy a slight advantage over their classical counterparts.

Current and future diving behavior are predictable from vocalizations alone

We next apply neural sequence models to predict not vocalizations, but behavior. When not floating on the surface of the water, sperm whales in the Eastern Caribbean community alternate between three high-level behavioral states during their active period: conducting deep foraging dives (at depths of over 600 meters), shallow dives (at depths of less than 200 meters), and sleep (during which whales are perpendicular to the surface of the water at depths of less than 100 meters). These behavioral states can be distinguished using accelerometry data that is captured and is aligned to acoustic data captured by the tags. While past work has found that some individual codas are predictably associated with whales' *current* behavioral state **?**, the question of whether vocalizalizations also carry information about *future* behaviors has remained open for decades.

Using accelerometry data, we automatically annotated the DSWP dataset with the behaviors that accompany vocalizations. These annotations are shown in Fig. **??**A: we split foraging dives into their descent and ascent phases, and additionally mark periods of sleep and shallow dives. See Tab. 1 in the supplementary materials for definitions of the behavioral phases and Section 2.1 in the supplementary materials for details of the annotation procedure. Using these annotations, we then train sequence models with the structure described in Fig. **??**B(i) and Fig. **??**C(i) to perform two prediction tasks. We first predict the *current* behavioral state (i.e. the state of the whale at the moment a particular coda was produced). Because no vocalization occurs during sleep, this involves discriminating between three states: the descent and ascent phases of foraging dives, along with shallow dives. We additionally also predict the *future* behavioral state of the vocalizing whale. We train a model to predict the tagged whale's next action after the call sequence is produced: whether it will be a deep foraging dive or some other behavior (e.g. another shallow dive or sleep).

Results for both prediction tasks are shown in Fig. **??**B(ii, iii) and C(ii, iii). We evaluate on balanced test test set containing equal portions of the three current-behavior and two future-behavior labels. It can be seen that, for both the present and future prediction tasks, a neural sequence model trained to predict behavior from vocalizations can do so non-trivially, obtaining 72.8% accuracy on the current behavior prediction task (Wilcoxon Sign-Ranked Test, sum of ranks $= 54$, p $= 0.002$) and 86.4% accuracy on the future behavior prediction task (Wilcoxon Sign-Ranked Test, sum of ranks $= 55$, p $= 0.001$), compared to a random baseline at 33.3% and 50% for the tasks respectively.

Behavior prediction is possible from coda sequences, but not isolated codas

As in Fig. **??**B(iv) and Fig. **??**C(iv), we conclude by evaluating what aspects of whales' vocalizations are necessary for accurate behavior prediction. We first evaluate coda-level features by training predictors on rhythm features or tempo features alone. As can be seen in Fig. **??**C(ii), tempo features alone suffice to match the accuracy of the full model at both behavior prediction tasks, while rhythm features alone provide reduced (but still non-trivial) predictive accuracy. This demonstrates that specific combinations of rhythm and tempo are uniquely produced before and during foraging dives.

In the future behavior prediction task, these results are partially explained by a single coda type that individually predicts future diving behavior. In Fig. **??**B(iv), it can be seen that most pre-dive calls, i.e. calls produced within 15 minutes before the onset of a foraging dive, contain 1+1+3/5 codas (where 1+1+3 denotes the rhythm category and 5 denotes the tempo category), while these are comparatively infrequent in contexts that are not followed by a dive. In fact, considering only the subset of exchanges from Fig. **??**B(iv) in which *all* calls (from both the tagged whale and conspecifics) are of the long 1+1+3 type, we find that 67.4% of these exchanges are followed by a dive, while only 19.6% of other exchanges are followed by a dive (Fisher's exact test (two-sided), odds ratio: 8.94, $p = 8.2e^{-7}$).

Conversely, for the task of predicting the whales' current rather than future behavior, we find that single coda types are not strongly predictive of behavioral context, i.e., ascent, descent and shallow dives: when limiting the number of preceding codas available to the model, as in Fig. **??**D (*left*), accuracy on this task is substantially degraded relative to performance with long input sequences (of seven or more codas). This result indicates distinctive *sequences* of codas discriminate different behavioral contexts from each other. This pattern may again be seen visually: in Fig. **??**D (*left*), we embed all codas from our dataset in two dimensions using t-SNE, then draw lines connecting codas produced sequentially in different behavioral contexts. Each context exhibits a distinctive sequence of coda transitions, even though some individual coda types are produced in multiple contexts. In some cases, these coda sequences are reproduced identically on different days and by different individuals

when the same behavior occurs. If codas are taken to be the atomic units of the communication system, this finding may be interpreted as revealing a kind of "behavior-dependent syntax" governing coda production (analogous to that observed in house finches **?**).

Concluding Remarks

Machine learning offers promising directions for advancing our understanding the complex communication systems of sperm whales. We have shown that the neural sequence models can identify novel structure within vocalizations produced by sperm whales in the EC1 clan, predict likely future vocalizations, and in some cases link vocalization to behavior. A major challenge in studying an animal communication system is simply identifying, from within an enormous hypothesis space, which features of the system are likely to be information-carrying, and how these features relate to behavior. Our results show that neural sequence models for animal communication, analogous to language models for human languages, can play a key role in meeting these challenges.

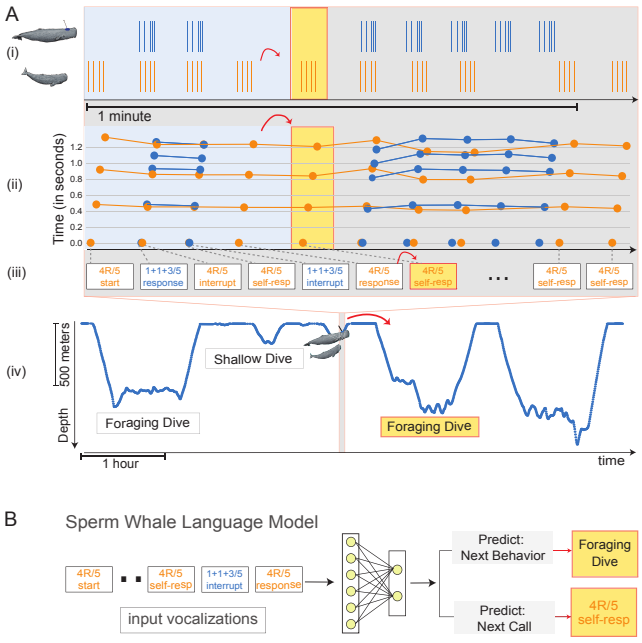

Figure 1: **Sperm Whale Language Model.** Sperm whales produce sequences of clicks grouped into distinct clusters called codas. **A** shows an (approximately) minute-long interaction between two whales, with the tagged whale's calls in blue and a conspecific's calls in orange. (ii) highlights the corresponding exchange plot **?** of the interaction. (iii) depicts this exchange represented as a series of discrete tokens (with codas organized into discrete types following Sharma et al. **?**. (iv) shows the depth profile of the tagged whale over a longer time window. **B** depicts a language model trained on the task of the next coda and behavior prediction. Like language models trained on human-generated text data, this model is trained autoregressively on sequences of vocalizations like the one depicted in (iii).

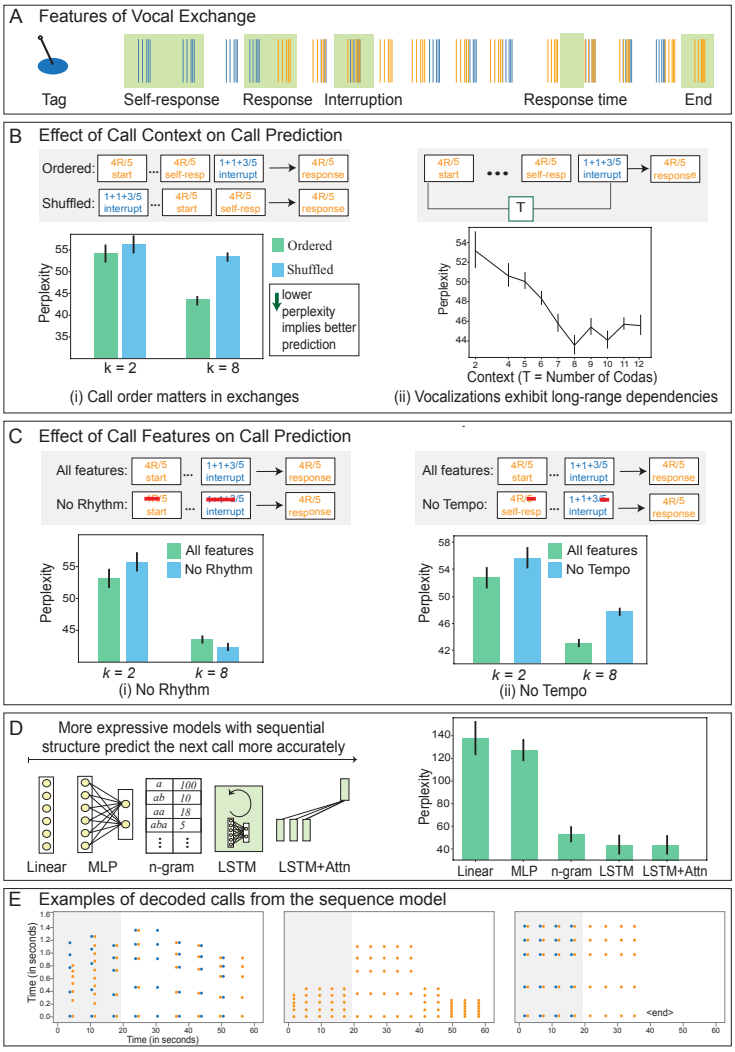

Figure 2: **Structure of Sperm Whale Exchanges.** We build a sequence model over sperm whale calls to identify what aspects of their call sequence encode information that is predictive of the next call. **A** depicts the sequence of calls exchanged between two whales and various turn-taking and response patterns. In **B**, we analyse the effect of communicative context on call production. We find that coda production is sensitive to the ordering of previous codas (left) and sensitive to a history of up to 8 codas (right). In **C** we evaluate the effect of specific features of the calls on the predictability of the next call in the sequence. We find a considerable decrease in models' predictive ability when the tempo feature is removed, indicating that the rhythm feature also carries information about future vocalizations. For models with longer input contexts, we observe that omitting the rhythm of the calls has little effect on model performance; however, doing so has a detrimental effect on models with shorter communicative contexts. **D** evaluates the change in model performance with the complexity of the model class. We find that more expressive neural models fit the data distribution better than linear models or (count-based) n-gram models. This shows the existence of long-range dependencies in the communication system that are difficult to model with surface statistics alone. **E** shows example vocalizations generated from the trained model when "prompted" with the sequence of vocalizations shaded in gray. For **B**, **C**, and **D**, Cousineau–Morey error bars are plotted; see manuscript for statistical tests.

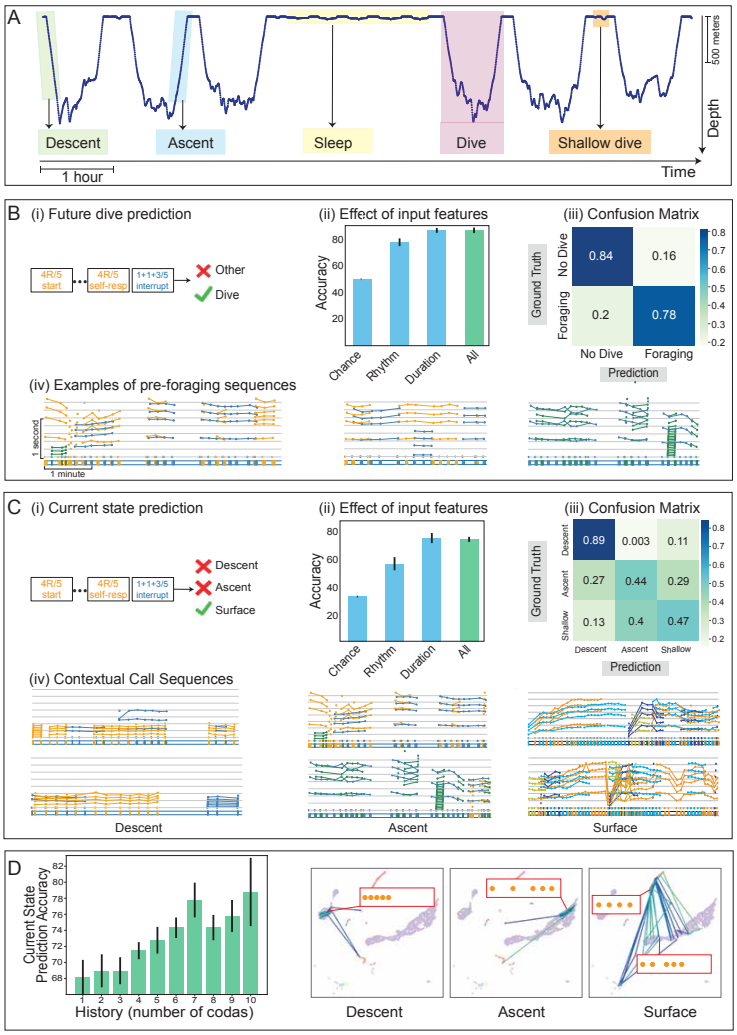

Figure 3: **Predicting behavior from Vocalizations.** **A** shows depth profile of a tagged whale and the corresponding behavioral states of the whale across the period depicted. **B** (i) depicts a neural sequence model trained to predict the future diving behavior of the whale based on its current sequence of calls. (ii) The model predicts the future behavioral state of the whale correctly 86.4% of the time, significantly better than random-chance baseline of 50%. The sequence of durations of the calls in the sequence is most informative of the next state. (iii) shows the confusion matrix evaluating the model's performance on the test set over different classes. (iv) Examples of pre-dive codas **C** (i) shows model trained to predict the current state of the whale given a sequence of calls. (ii) The model predicts the current state with an accuracy of 72.8% accuracy, again significantly greater than a chance baseline at 33%. Here too we see that duration information is independently informative about current behavior. (iii) Confusion matrix for the task of current state prediction. (iv) Sample calls for different behavioral states. **D** (Left) Models with a larger input context predict the current behavioral state of the whale better. (Right) By embedding codas in two dimensions using t-SNE, and connecting codas produced during the same exchange, we observe characteristic *sequences* of codas associated with different behavioral states, even when some of the constituent codas in these sequences recur across contexts. For **B**, **C**, and **D**, Cousineau–Morey error bars are plotted; see manuscript for statistical tests.

References

Philip Lieberman. The biology and evolution of language. 1984.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, 2002.

Ray Jackendoff. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, UK, 2002.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44, 04 2023.

Christo Kirov and Ryan Cotterell. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665, 2018. doi: 10.1162/tacl_a_00247. URL https://aclanthology.org/Q18-1045.

Shane Steinert-Threlkeld and Jakub Szymanik. Learnability and semantic universals. *S&P*, 12:4:1–39, November 2019.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL https://api.semanticscholar.org/CorpusID:235959867.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

H Whitehead. Sperm whales: social evolution in the ocean. *Choice*, 41(06):41–3452–41–3452, February 2004.

Maurício Cantor, Lauren G Shoemaker, Reniel B Cabral, César O Flores, Melinda Varga, and Hal Whitehead. Multilevel animal societies can emerge from cultural transmission. *Nat. Commun.*, 6:8091, September 2015.

Shane Gero, Anne Bøttcher, Hal Whitehead, and Peter Teglberg Madsen. Socially segregated, sympatric sperm whale clans in the atlantic ocean. *R Soc Open Sci*, 3(6):160061, June 2016.

M Marcoux, H Whitehead, and L Rendell. Sperm whale feeding variation by location, year, social group and clan: evidence from stable isotopes. *Mar. Ecol. Prog. Ser.*, 333:309–314, March 2007.

Maurício Cantor and Hal Whitehead. How does social behaviour differ among sperm whale clans? *Mar. Mamm. Sci.*, 31(4):1275–1290, October 2015.

Hal Whitehead and Luke Rendell. Movements, habitat use and feeding success of cultural clans of South Pacific sperm whales. *J. Anim. Ecol.*, 73(1):190–196, January 2004.

William A Watkins. Sperm whale codas. *J. Acoust. Soc. Am.*, 62(6):1485, 1977.

Linda Weilgart and Hal Whitehead. Coda communication by sperm whales (physeter macrocephalus) off the galápagos islands. *Canadian Journal of Zoology*, 71(4):744–752, 1993. doi: 10.1139/z93-098. URL https://doi.org/10.1139/z93-098.

L E Rendell and H Whitehead. Vocal clans in sperm whales (Physeter macrocephalus). *Proc. Biol. Sci.*, 270 (1512):225–231, February 2003.

Pratyusha Sharma, Shane Gero, Roger Payne, David F Gruber, Daniela Rus, Antonio Torralba, and Jacob Andreas. Contextual and combinatorial structure in sperm whale vocalisations. *Nat. Commun.*, 15, May 2024.

Antonio Leitao, Maxime Lucas, Simone Poetto, Taylor A. Hersh, Shane Gero, David F. Gruber, Michael Bronstein, and Giovanni Petri. Evidence of social learning across symbolic cultural barriers in sperm whales, 2024.

Gašper Beguš, Andrej Leban, and Shane Gero. Approaching an unknown communication system by latent space exploration and causal inference. *arXiv preprint arXiv:2303.10931*, 2023.

Shafi Goldwasser, David Gruber, Adam Tauman Kalai, and Orr Paradise. A theory of unsupervised translation motivated by understanding animal communication. In *NeurIPS 2023*, December 2023.

Alexandros Frantzis and Paraskevi Alexiadou. Male sperm whale (physeter macrocephalus) coda production and coda-type usage depend on the presence of conspecifics and the behavioural context. *Canadian Journal of Zoology*, 86(1):62–75, 2008.

Christian Rutz, Michael Bronstein, Aza Raskin, Sonja C Vernes, Katherine Zacarian, and Damián E Blasi. Using machine learning to decode animal communication. *Science*, 381(6654):152–155, 2023.

Pratyusha Sharma, Shane Gero, Roger Payne, David F. Gruber, Daniela Rus, Antonio Torralba, and Jacob Andreas. Contextual and combinatorial structure in sperm whale vocalisations. *bioRxiv*, 2023. doi: 10.1101/2023.12.06.570484. URL https://www.biorxiv.org/content/early/2023/12/08/2023.12.06.570484.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Joe O'Connor and Jacob Andreas. What context features can transformer language models use? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021.

William A Searcy, Jill Soha, Susan Peters, and Stephen Nowicki. Long-distance dependencies in birdsong syntax. *Proceedings of the Royal Society B*, 289(1967):20212473, 2022.

Takashi Morita, Hiroki Koda, Kazuo Okanoya, and Ryosuke O Tachibana. Birdsong sequence exhibits long context dependency comparable to human language syntax. *bioRxiv*, 2020.

Jenny A Allen, Ellen C Garland, Rebecca A Dunlop, and Michael J Noad. Network analysis reveals underlying syntactic features in a vocally learnt mammalian display, humpback whale song. *Proceedings of the Royal Society B*, 286(1917):20192014, 2019.

Yoichi Inoue, Waidi Sinun, Shigeto Yosida, and Kazuo Okanoya. Note orders suggest phrase-inserting structure in male mueller's gibbon songs: a case study. *acta ethologica*, 23:89–102, 2020.

Esther Clarke, Ulrich H Reichard, and Klaus Zuberbühler. The syntax and meaning of wild gibbon songs. *PloS one*, 1(1):e73, 2006.

Maël Leroux, Alexandra B Bosshard, Bosco Chandia, Andri Manser, Klaus Zuberbühler, and Simon W Townsend. Chimpanzees combine pant hoots with food calls into larger structures. *Animal Behaviour*, 179:41–50, 2021.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017. URL https://arxiv.org/abs/1704.04368.

Ivy Ciaburri and Heather Williams. Context-dependent variation of house finch song syntax. *Animal Behaviour*, 147:33–42, 2019.

Mark P Johnson and Peter L Tyack. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE journal of oceanic engineering*, 28(1):3–12, 2003.

Felicia Vachon, Luke Rendell, Shane Gero, and Hal Whitehead. Abundance estimate of eastern caribbean sperm whales using large scale regional surveys. *Marine Mammal Science*, 2024.

Andrea Ravignani, Daniel L Bowling, and W Tecumseh Fitch. Chorusing, synchrony, and the evolutionary functions of rhythm. *Front. Psychol.*, 5:1118, October 2014.

Tyler M Schulz, Hal Whitehead, Shane Gero, and Luke Rendell. Overlapping and matching of codas in vocal interactions between sperm whales: insights into communication function. *Anim. Behav.*, 76(6):1977–1988, December 2008.

Robert C. Berwick, Kazuo Okanoya, Gabriel J.L. Beckers, and Johan J. Bolhuis. Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Sciences*, 15(3):113–121, 2011. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2011.01.002. URL https://www.sciencedirect.com/science/article/pii/S1364661311000039.

A Sperm Whale Language Model

A.1 Tokenization scheme

In natural language processing, *tokenization* refers to the process of representing a text corpus in terms of a finite collection of atomic text units called *tokens*. To build a sequence model of whale exchanges, we apply an analogous tokenization procedure to represent call sequences in terms of a finite set of atomic elements. Following the phonetic alphabet features defined by **?**, we represent each coda using rhythm, tempo, rubato, and ornamentation (see Table **??** for definitions). Past work identifies 18 rhythm types, 5 tempo types, 3 rubato types, a binary ornamentation feature. To accurately model the structure of whale exchanges, our tokenization scheme also includes information about turn-taking behaviors, which account for speaker changes and the timing of calls. There are three types of turn-taking: 1) Self-response: A whale follows up its own call after a pause. 2) Response by another whale: A different whale responds after a pause. 3) Overlapping call: A whale produces a call that overlaps with another's. Each unique combination of rhythm, tempo, rubato, ornament, and turn-taking behavior is assigned a unique token. Not all possible combinations are realized across different data splits. Any combination not present in the training split but appearing in the test split is mapped to the same `<unk>` (unknown) token.

A.2 Details on cross-validation

To ensure robust performance estimates, we conduct each experiment on 10 different dataset splits. In each split, whale calls from a single day are held out for testing, while recordings from the remaining days form the training and validation set. There is no overlap in days between the training and test recordings in any split. The size of the train-val-test datasets is different for different splits of the dataset. This dataset splitting ensures we measure the model's ability to generalize to exchanges from a new day. Due to variability in exchanges across days, the model's performance varies across different splits.

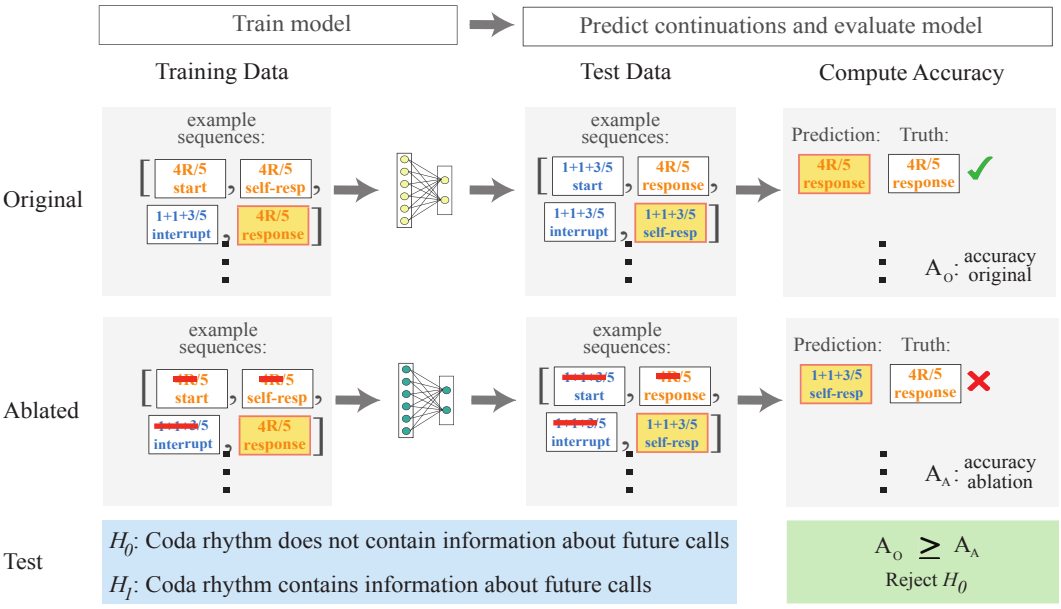

Figure 4: **Schematic of our method:** Our proposed approach uses sequence models to test hypotheses about the information content and structure of whale calls. Here, we illustrate our method with an example of verifying if rhythm impacts the prediction of future calls. Two models, one with information about rhythm in its input and the other without, are trained and then tested to predict all the call features i.e., the same output. If the model without information of the rubato loses predictive power on the test set, then we reject the null hypothesis "H_0: *Coda rhythm does not contain information about future calls*" in favour of the alternate hypothesis.

A.3 Architectural details

To evaluate how expressive sequence models need to be to capture the long-range dependencies and structure of whale calls, we train a collection of models with different inductive biases. Each model is trained on the same input sequence length (sequence length of 6) and optimized for the same objective: predicting the next call. We outline the architectural details of the models below.

n-gram model: An n-gram model is a probabilistic language model that predicts the probability of the next item in a sequence based on the previous $n-1$ items. This is done by computing frequency-based estimates of the conditional probability of the next token given the previous $n-1$ tokens on the training set.

The paper uses the implementation of the kenLM repository for training the n-gram models **?** with its default settings. Like with any count-based model, one challenge with an n-gram model is modeling the probability of occurrence of unseen n-grams. To obtain better probability estimates for unseen and less frequent n-grams, Kneser-Ney smoothing is used. Kneser-Ney smoothing starts by discounting from the counts of observed n-grams and redistributes the probability mass to better handle rare and unseen n-grams. Further, the n-gram model is discounted with backoff penalties. Backoff penalties in n-gram models adjust probability estimates when the model has to rely on lower-order n-grams due to the absence of higher-order ones. These penalties help balance the model's reliance on different levels of context, ensuring more accurate and realistic probability estimates across different sequences.

Linear model: A linear model assumes that the output can be expressed as a linear combination of the input features. Here we learn a linear model that outputs the probability distribution over the next token given the previous 6 calls. The number of parameters in this model is a product of the context window times the output.

Multi-layer perceptron: A multi-layer perceptron model contains multiple linear layer layers arranged in a feed-forward fashion with non-linearities between the layers. For this experiment we train a two-layer neural network with a hidden dimension of 64 with a ReLU non-linearity in between.

LSTM: An LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) architecture. The LSTM cell contains a more complex unit structure with a specialized gating mechanism that regulates the flow of information in the network, thereby giving it a much more powerful inductive bias to effectively model sequential data. For our experiments, we use an encoder-decoder LSTM, where the encoder LSTM encodes the input sequence into a context vector and the decoder LSTM decodes this vector into an output sequence. We use a bi-directional encoder LSTM cell and a uni-directional decoder LSTM cell, both with 64-dimensional hidden state.

LSTM with attention: An LSTM with Attention is an enhanced version of the Long Short-Term Memory (LSTM) architecture. It incorporates an attention mechanism to improve the model's ability to focus on specific parts of the input sequence when generating each element of the output sequence. This architecture is especially useful for tasks when certain parts of the input sequence are more relevant to the output than others. We modify the architecture of the encoder-decoder LSTM to add to this computation.

Implementational details: The parameters of the model were trained with stochastic gradient descent (SGD) using the Adam optimizer with a learning rate of $1e^{-4}$, weight decay of $1e^{-5}$ and batch size of 32. We early-stopped the training of the models based on their performance on a held-out validation set to prevent the models from over-fitting on the small training sets. This usually resulted in the models being trained up to 50 epochs in practice.

A.4 Additional Ablations: Rubato and Ornamentation

Ablating ornamentation and rubato information does not affect the model's ability to predict the next call with statistical significance. This may be partly because ornaments are rare, making up only 4% of the dataset, and the dataset is too small to capture the precise dynamics of changing rubato.

B Behavior Prediction

B.1 Details on annotating behavior phases

The different behavioral phases—sleep, shallow dives, and foraging dives—are annotated both automatically and with input from expert humans. Below we outline the procedure used to identify each of the different behavioral phases.

Foraging dives: The start and end points of a whale's foraging dives are moments when the whale starts a sharp descent into the ocean to forage and when the whale first arrives at the ocean's surface by ascending post-foraging. These are automatically detected using the accelerometer and depth data from the DTAG. Foraging dives typically show a steep, uninterrupted descent and ascent profile. A foraging dive is identified when the rate of depth change is nearly constant before and after the start and end points and when the whale reaches a depth of over 500m. This method correctly identifies all the dive start and end points from the collected DTAG data, which are thereafter verified by a human for accuracy.

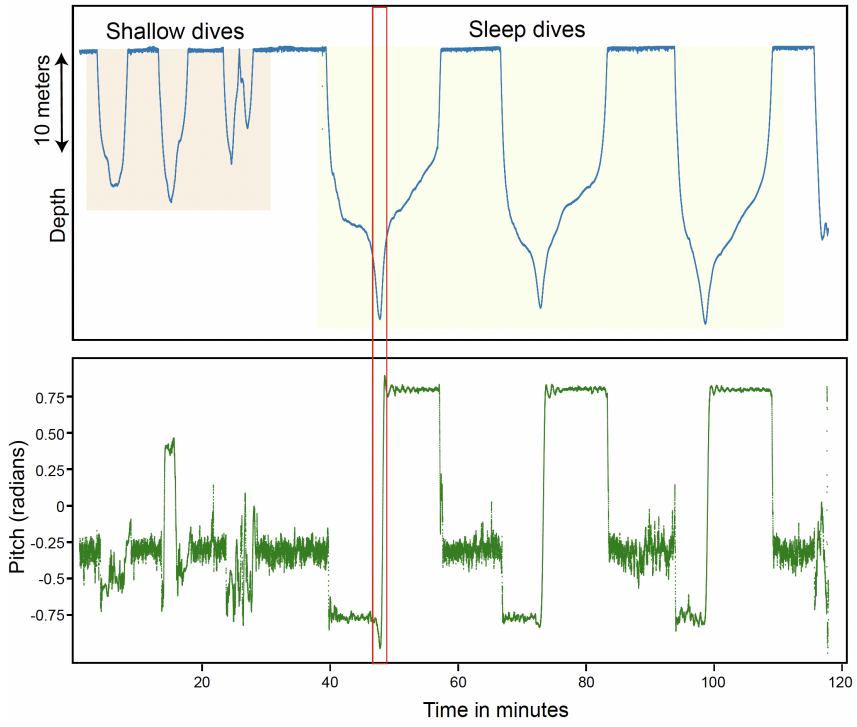

Figure 5: **Pitch and depth data for shallow and sleep dives:** Shallow and sleep dives are identified using motion data collected using the DTAG. Sperm whale sleep dives have a characteristic change in the depth and pitch data (indicated by the red box) where the whale goes from a position parallel to the surface of the ocean to one where it becomes perpendicular to the ocean.

Shallow and sleep dives: Sleep and shallow dives are identified using accelerometer data and are then annotated and verified by expert human annotators. Examples of the depth and accelerometer data for these dives are shown in Fig. **??**. Sleep dives exhibit a distinctive change in the accelerometer reading, indicating the whale's shift from a horizontal position, parallel to the ocean's surface, to a position that is vertical and perpendicular to the ocean's surface. This is highlighted in Fig. **??**. In contrast, no such change in the accelerometer reading is observed in other shallow dives. A dive is classified as shallow if the maximum depth of the whale in the course of the dive is less than 100m.

B.2 Behavioral-context baselines for future-behavior prediction.

Our main experiments show that vocalizations contain information about both vocalizing whales' present and future behavior. However, these two prediction targets are correlated with each other: for example, 83% of foraging dives are followed by another foraging dive, rather than a shallow dives. Thus it is possible to obtain non-trivial accuracy at the future-behavior-prediction task using *only* information about a whale's current state, and not its vocalizations.

In this section, we present an additional analysis showing that these behavioral correlations do not fully explain model accuracy at the future prediction task: that is, vocalizations contain information about future behavior *even after accounting for the information they contain about present behavior.* In particular, we compare the difference in the performance of the future-behavior-prediction model with a model that predicts the most common next-turn behavior *conditioned* on a whale's current behavioral state as predicted by the current-behavior-prediction model. Across the different cross-validation splits, the average difference in the performance between the future-behavior-prediction model and this suggested baseline model is 21.92% (test: Wilcoxon Sign-Ranked Test, sum of ranks $= 36$, p-value $= 0.006$). This indicates that future-behavior predictions are not fully explained by correlations between future and current behaviors: some vocalization features are directly predictive of future behavior.

Interestingly, the characteristic "pre-dive" calls identified in the main text are produced not only during the ascent phase but also at the end of social exchanges produced at the surface; however, not all calls produced during ascent follow this "pre-dive" pattern.

B.3 Dataset

We study coda exchanges in a manually annotated coda dataset from The Dominica Sperm Whale Project (DSWP). This includes recordings of the Eastern Caribbean clan (EC1) collected between 2014 and 2018 from bio-logging tags (Dtags, **?**) deployed on known individuals off the island of Dominica. This dataset contains manually annotated coda clicks and extracted inter-click intervals comprising 3948 codas **?**. The dataset also contains the accelerometer, gyroscope, and magnetometer readings from the tags. This allows us to compute the position of the tagged whale over time. The EC1 clan has a membership of fewer than 300 individuals **?**. A total of 41 tags were deployed on 25 different individuals in 11 different social units. We conservatively estimate that at least 60 distinct whales are recorded in our dataset. An example sequence of coda exchanges between two whales and the depth profile of the tagged whale is shown in Fig. **??**A.

Notation	Description
Coda:	A short burst of clicks with varying inter-click intervals generally less than two seconds in duration.
Inter Click Interval (ICI):	The time difference between two consecutive clicks within a coda.
Coda duration:	The sum of a coda's absolute ICIs.
Rhythm type:	The discrete category a coda is assigned to based on its characteristic sequence of standardized ICIs.
Tempo type:	The discrete category a coda is assigned to based on its characteristic duration.
Exchange / Chorus:	Period of time where codas are made by more than a single whale (as in **?**).
Single-Whale Call Sequence:	A sequence of calls made by a given whale where every consecutive pair of calls occur within 8 seconds (twice the average response time) of each other.
Turn-taking:	An exchange of codas involving alternating coda production. Also referred to as 'adjacent' codas, these are defined as next-in-sequence codas whose onset occurred within two seconds, but after the termination, of the initial coda (as in **?**).
Overlapping Codas:	An exchange of codas such that the next-in-sequence coda's onset occurs after the onset, but before the termination, of the previous coda (as in **?**).
Ornament:	"Extra click" appended to the end of a coda in a group of shorter codas. (For further details on the identification criterion, see Ornamentation section in the manuscript.)
Rubato:	Gradual variation in duration across adjacent codas made by the same whale within the same rhythm and tempo type.
Descent:	The initial period of a foraging dive where there is a steady increase in the depth the whale is located at. This is the period of time starting where the whale is at the surface of the water and makes a plunge to start its foraging dive to the point it reaches a depth at which it can start feeding.
Ascent:	The terminal period of a foraging dive where there is a steady decrease in the depth the whale is located at. This is the time period starting where the whale is returning from feeding in deep waters to the point of time it reaches the surface of the water.
Social (Socializing on the surface):	The period of time when multiple whales remain at the surface or make shallow dives (< 300 meters).
Foraging dives:	Deep dives typically involve whales diving to a depth of over four hundred meters. Deep dives almost always have buzzes which are evidence of foraging.
Pre-dive calls:	The set of codas made fifteen minutes less before the onet of a foraging dive.
Behavioral Contexts:	Groups of behaviors exhibited by whales motivated by a set of goals (diving, socializing, pre-dive etc)
Context specific calls:	The set of calls prototypically associated with a unique behavioral context.

Table 1: Glossary: Definitions of previously used and newly introduced terminology.

Notation	Description
Language:	Any possible set of strings over some (usually finite) alphabet of words. **?**
Syntax:	The rules for arranging items (sounds, words, word parts or phrases) into their possible permissible combinations in a language. **?**
Likelihood:	Likelihood is a statistical concept that measures how probable a particular set of observations is, given a specific model and its parameters.
Neural network:	A neural network is a computational model consisting of interconnected nodes (neurons) organized in layers. It is designed to recognize patterns and learn from data through training by adjusting the connections (weights) between nodes to improve predictions or classifications.
Multi-Layer Perceptron:	A multi-layer perceptron (MLP) is a type of neural network consisting of an input layer, one or more hidden layers, and an output layer. Each layer is made up of neurons that use activation functions to process inputs and produce outputs.
Neural Sequence Model:	A neural sequence model is a type of machine learning model designed to handle sequential data, where the order of the data points is significant. Sequence models, such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), are used to predict the next item in a sequence or to understand dependencies within the sequence.
Sequence to sequence models:	A sequence-to-sequence (seq2seq) model is a type of neural network architecture designed to transform one sequence into another. It consists of an encoder that processes the input sequence and a decoder that generates the output sequence.
Language Model:	A language model is a type of sequence model typically trained on the next token prediction object. It learns the probabilities of sequences of tokens, enabling it to generate coherent text, autocomplete sentences, or predict the next word in a sentence.
LSTM:	A type of recurrent neural network (RNN) architecture designed to effectively capture long-term dependencies in sequential data.
n-gram models:	Statistical language models that predict the next item in a sequence based on the preceding $n-1$ items. They represent sequences as contiguous sequences of n items (words or characters) and estimate the probability of each sequence based on the observed frequencies of such sequences in the training data.
Perplexity:	Perplexity is a metric used to evaluate the performance of a language model. It is simply the exponentiated average log-likelihood per token. It measures how well the model predicts a sequence, with lower perplexity indicating better performance.

Table 2: Glossary 2: Definitions of important linguistics and ML concepts