APPROACHING AN UNKNOWN COMMUNICATION SYSTEM BY LATENT SPACE EXPLORATION AND CAUSAL INFERENCE

Gašper Beguš[1,3,+], Andrej Leban[1,3,+], and Shane Gero[2,3,4]

[1]University of California, Berkeley, Berkeley, CA, United States
[2]Department of Biology, Carleton University, Ottawa, Canada
[3]Project CETI, New York, NY, United States
[4]The Dominica Sperm Whale Project, Roseau, Dominica
[+]these authors contributed equally to this work
{andrej_leban, begus}@berkeley.edu

ABSTRACT

This paper proposes a methodology for discovering meaningful properties in data by exploring the latent space of unsupervised deep generative models. We combine manipulation of individual latent variables to extreme values outside the training range with methods inspired by causal inference into an approach we call *causal disentanglement with extreme values* (CDEV) and show that this approach yields insights for model interpretability. Using this technique, we can infer what properties of unknown data the model encodes as meaningful. We apply the methodology to test what is meaningful in the communication system of sperm whales (*Physeter macrocephalus*), one of the most intriguing and understudied animal communication systems. We train a network that has been shown to learn meaningful representations of speech and test whether we can leverage such unsupervised learning to decipher the properties of another vocal communication system for which we have no ground truth. The proposed technique suggests that sperm whales encode information using the number of clicks in a sequence, the regularity of their timing, and audio properties such as the spectral mean and the acoustic regularity of the sequences. Some of these findings are consistent with existing hypotheses, while others are proposed for the first time. We also argue that our models uncover rules that govern the structure of communication units in the sperm whale communication system and apply them while generating innovative data not shown during training. This paper suggests that an interpretation of the outputs of deep neural networks with causal methodology can be a viable strategy for approaching data about which little is known and presents another case of how deep learning can limit the hypothesis space. Finally, the proposed approach combining latent space manipulation and causal inference can be extended to other architectures and arbitrary datasets.

1 Introduction

How do we approach a communication system for which we not only don't understand what is meaningful but also do not know how to test for what is meaningful? One such case is the communication system of sperm whales (*Physeter macrocephalus*). Sperm whales communicate with sets of click vocalizations called *codas* (Watkins and Schevill, 1977). Their vocalizations are likely meaningful because they are produced in duet-like exchanges or group choruses (Schulz et al., 2008), predominantly during socialization and before dives but never while alone or when foraging at depth (Watwood et al., 2006; Whitehead, 2003; Whitehead and Weilgart, 1991), and appear to be socially learned (Rendell et al., 2012; Whitehead and Rendell, 2014). While evidence supports the use of codas as identity signals (Gero et al., 2016b; Hersh et al., 2022; Rendell and Whitehead, 2003), very little is currently known about what individual utterances mean or even what kind of properties of the communication system could have the potential to carry meaning.

In this paper, we propose an approach for discovering meaningful properties in data by using an expressive generative model as the learning mechanism. The network is trained with two objectives: (i) imitation of data and (ii) encoding of information (Figure 1). We then combine latent space exploration with methods borrowed from causal inference into a methodology we call *causal disentanglement with extreme values* (CDEV). This approach yields insights into what properties the network has learned encode uniquely relevant information for the synthetic vocalizations.

If sperm whales encode information into their vocalizations and if our model can learn to imitate those well, it is likely that the encoding in our models can reveal what might be meaningful in the sperm whale communication system. We argue that our technique reveals both properties that were posited as meaningful by human researchers as well as novel properties that have so far not yet been hypothesized as such.

One of the advantages of the proposed approach is that the fiwGAN architecture (Beguš, 2021b) used as the learning mechanism is innovative in a highly informative way. For example, when trained on human speech, the fiwGAN network generates novel words that were never part of training data (Beguš, 2021a,b). Similarly, when trained on sperm whale codas, the network not only replicates codas from the training set but also generates codas not shown during training. As such, our network discovers rules that govern the structure of codas and applies the rules learned from observed training data to novel, innovative outputs. Such inventiveness makes fiwGAN especially well-suited to the use as the learning mechanism of properties in data about which little is understood, which is the cornerstone of the approach presented in this work. In some sense, we use the network (in conjunction with CDEV) as an infinitely-flexible and information-preserving tool for decomposing the data into meaningful, observable properties without the need for, e.g., dimensionality reduction.

Causal inference is a discipline that aims to develop the assumptions, experiment designs, and statistical methodology needed to determine the *causal effect* of a particular variable that can be manipulated — the *treatment* — on some outcome(s) of interest (Hill and Stuart, 2015). While this usually involves evaluating counterfactual scenarios, for example via the *potential outcomes* framework, this is not *the* fundamental purpose of the discipline, allowing us to use the methodology in studies such as the one presented in this paper.

1.1 Sperm whale communication

In social contexts, sperm whales communicate in short (less than two seconds), socially learned, stereotyped patterned series of *clicks* called *codas* that marine biologists have grouped into several coda *types* based on the variation in the number, rhythm, and tempo of the clicks within codas (summarized in (Whitehead, 2003; Whitehead and Rendell, 2014). Fig. 2 is an illustration of such codas. Even when living in the same waters, whales only associate with other whales which use a similar dialect of coda types, creating a higher level in their social structure delineated based on culture — the vocal *clan* (Gero et al., 2016a; Hersh et al., 2022; Rendell and Whitehead, 2003). While both are broadband in the spectrum, coda clicks — used in communication — can be acoustically distinguished from echolocation clicks, which are used for navigation and foraging (Madsen et al., 2002; Whitehead, 2003).

The dataset of recordings for this study originates from The Dominica Sperm Whale Project (see Gero et al. 2003) and was collected off the coast of the island of Dominica between 2014 and 2018 from over 4000 hours in the company of sperm whales. The vast majority of the recordings are made of one sperm whale vocal clan. In the dialect of this Eastern Caribbean Clan, there are about 22 different coda types that have been defined (Gero et al., 2016a). The actual distribution and production rates of coda types are highly asymmetrical in which the two most common codas comprise 65% of all coda vocalizations recorded between 2005 and 2010 (Gero et al., 2016b).

Codas were collected from animal-borne sound and movement tags between 2014 and 2018 (*Dtag* generation 3; Johnson and Tyack 2003). *Dtags* record two-channel audio at 120 kHz with a 16-bit resolution, providing a flat (±2 dB) frequency response between 0.4 and 45 kHz. As a result, we were able to record clean signals for both the coda and echolocation clicks produced by sperm whales; and to obtain the temporal patterning and spectral properties of coda clicks used in this analysis.

Due to the above-mentioned rarity of some types in the dataset, we limit the training set used in this paper to the five most common coda types (1+1+3, 5R1, 4R2, 5R2, and 5R3, Fig. 1). This is further exacerbated by the fact we are dealing with a communication system, meaning that a significant portion of the data could not be used due to too strong of a presence of whale dialogue in which codas are often overlapped by one or more animals (Schulz et al., 2008). The residual effects of this are dealt with by our algorithms, as illustrated in Section 2. Restricting the number of coda types also helps us test whether the deep neural network can predict unobserved coda types. All in all, 2209 distinct training samples were used. GANs, contrary to other architectures, do not require extensive datasets and have been shown to learn informative properties of language with similar or substantially smaller datasets Beguš (2021b,c).

In terms of data preprocessing, a constant DC microphone bias was removed from the extracted coda recordings, which were then augmented by random zero-padding in the front to address the fact that all extracted codas would otherwise have a click at the very beginning.

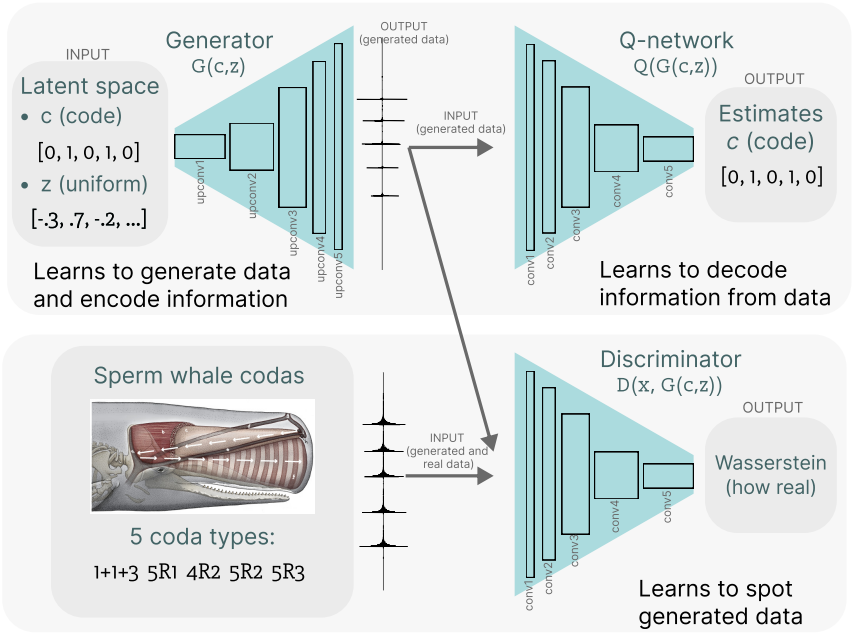


Figure 1: Overview of the fiwGAN architecture (Beguš, 2021b) and data used in training. The figure illustrates three networks: the Generator with 5 upconvolutional layers, the Q-network with 5 convolutional layers, and the Discriminator with 5 convolutional layers.

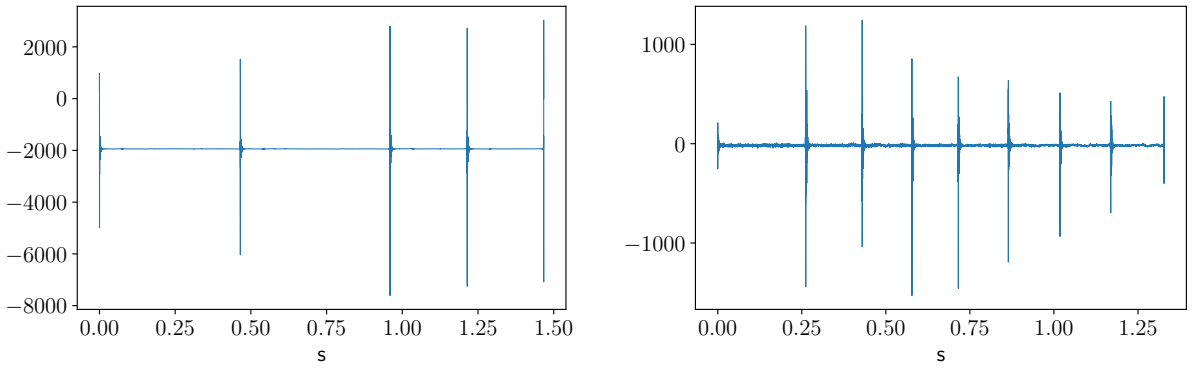


Figure 2: Examples of real codas. Left: type *1+1+3* with 5 clicks, right: type *9R* with 9 clicks. The 9R type was never part of the training data (Section 1.1), yet our network learns to generate codas with a high number of clicks that resemble this type and its regularity despite never having access to codas with more than 5 clicks (Sec. 5).

1.2 Why use deep learning?

While biologists have classified the vocalizations into *codas*, virtually nothing is known of their informational content (Andreas et al., 2022) One outstanding question about codas is whether the acoustic properties have any informational content. In this work, we aim to use the fact that the network used learns to encode informationally meaningful properties in an *unsupervised* way while being able to generate convincing synthetic whale utterances to discover what might be crucial to the communication system.

Convolutional neural networks (CNNs) are relatively unbiased compared to human researchers and can learn to encode properties that human analysis might not observe. Moreover, learning is performed on raw audio without information-losing transformations. To our knowledge, this is the first attempt to model sperm whale communication with deep learning on raw acoustic data.

Deep neural networks have recently been used to tackle some of the hard problems across sciences. Probing learned representations in deep neural networks has yielded insights in fields as diverse as drug discovery (Stokes et al., 2020), protein design (Jumper et al., 2021), or geometry (Davies et al., 2021), by shrinking the hypothesis space for each of the problems. To our knowledge, deep neural networks have not yet been used in attempts to decipher an unknown communication system.

The advantage of the proposed approach is that the discovery of meaningful properties uses as few assumptions as possible, treating the network as a black-box learning unit. The proposed approach, where individual units in the inputs of deep neural networks are manipulated to extreme values, and their effects are estimated via causal inference methods, could be applied to other architectures and data.

1.3 The architecture used

Generative adversarial networks (GANs; Goodfellow et al. 2014), in the most basic form, consist of two separate networks trained in an adversarial fashion. The *generator* outputs synthetic data with the goal of tricking the *discriminator* that the data is real. The latter's goal is the inverse: to distinguish the real data from the synthetic as well as possible.

The training objective is, therefore, a *minimax game:*

$$\min_G \max_D \mathbb{E}_{X \sim P_{data}}[D(x)] + \mathbb{E}_{Z \sim P_{synth}}[1 - D(G(z))]$$

The input to the generator during training is a randomly sampled continuous vector of a given length, usually chosen to be 100. Since there is no correlation between the input selected and the loss imposed on the generator, the latter learns to associate the inputs with the outputs in such a convoluted way that the inputs can be treated as *incompressible noise*. We use this assumption to justify the use of some of the causal inference estimators used in this work. Since the elements of the input are continuous numbers, it can be assumed that the provided length of the vector gives the model enough "informational space".

The network architecture used in this work is fiwGAN (Beguš, 2021b), an InfoGAN (Chen et al., 2016) adaptation of the WaveGAN (Donahue et al., 2019) model (which itself is based on DCGAN; Radford et al. 2015). Unlike InfoGAN, the fiwGAN features a separate Q-network and a binary code instead of a one-hot vector which enables featural learning. In short, it partitions the input into an *incompressible noise* z and an additional *featural encoding* c. The latter is learned in an unsupervised way by an additional network Q during training, with the objective of ensuring consistency of output across similar values of this vector, in contrast to z. It achieves this consistency by additionally penalizing the generator for inconsistent output for a given value of c by backpropagating a loss based on mutual information between the generated output and the encoding c (Figure 1).

In our case, the generator needs to learn to generate audio that resembles sperm whale codas without accessing the actual sperm whale vocalizations directly. Learning thus occurs primarily by imitation in a fully unsupervised manner via the adversarial loss. Additionally, the network needs to encode information into its synthetic outputs so that another network — Q — is able to decode this in a game that mimics communicative intent (Figure 1).

Beguš (2020) proposes a technique to uncover individual latent variables that have linguistic meaning by setting latent variables to extreme values outside of the latent space and interpolating from extreme values of those variables. It is shown that networks trained on speech data learn to associate lexical items with code values and sublexical structure with individual bits in c (Beguš, 2021b; Beguš and Zhou, 2022).

1.4 Prior work

Combining latent space exploration with causal inference is a novel approach to the interpretability of unsupervised deep neural networks. Commonly used methods for interpretability, such as integrated gradients (Sundararajan et al., 2017), require access to the network itself, which is not the case here.

Conversely, the majority of prior work using deep learning in causal inference uses it in the opposite direction (c.f. Louizos et al. (2017)): to infer the outcome curve with missing observations rather than applying causal inference methods themselves on the outputs of deep networks.

Chockler et al. (2021) use causal methods for interpreting CNNs, but their objective is to find the subset of the input image that most influences the decision of a supervised classifier when the input is occluded rather than to uncover what an unsupervised generative model is learning in terms of derived quantities. Kocaoglu et al. (2018) present an approach to training a generative model that preserves an a-priori determined causal relationship. Bose et al. (2022a,b) use causal inference with binary treatments to quantify implicit causal relationships *between* latent space variables *within* models

(i.e., a white-box method) such as GANs via the use of proxy classifiers, in order to aid the controllable generation of outputs. In contrast, *CDEV* is a black-box method that uses continuous-treatment causal inference methodology to uncover *observable* attributes that are encoded as significant by an unsupervised learner. As such, it makes no claims about potential causality within the latent space and does not need access to the latter, simplifying the justification for the use of the methods presented (Sec. 3). Additionally, to our best knowledge, none of the related work makes use of setting the inputs to extreme values in order to disentangle the encodings.

In this work, we, therefore, analyze the generator *from the outside*: after having it learn to replicate the whale utterances and encode what it considers meaningful in an unsupervised way, we aim to uncover the latter by merely manipulating the input to the model and applying statistical methods on the generated audio: in this way, the methodology is agnostic to the type of generative process used. We present the setup of the experiment in Section 2 and argue for the use of causal inference methods in Section 3. The three particular causal methods used are presented in Section 4, which also illustrates the first encoding uncovered: the range of the number of clicks. The two succeeding sections present the results for two additional (groups of) attributes: coda regularity in Section 5, and the acoustic characteristics of the generated codas in Section 6.

2 The CDEV technique and data generation

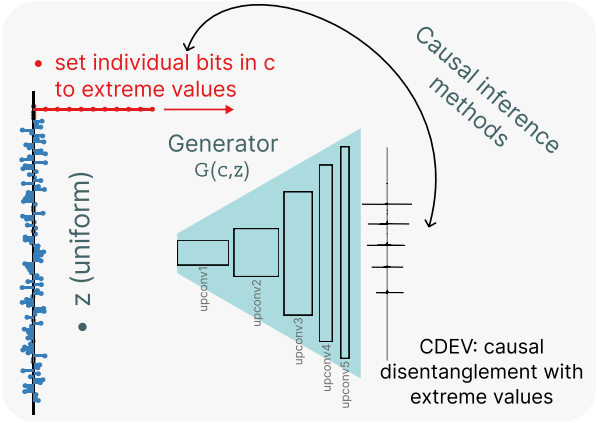


Figure 3: Overview of the *causal disentanglement with extreme values* process: individual bits in the code c are set to extreme values while z is kept within the training range.

In simple GANs, the generation is done by uncoupling the generator after training and again feeding it vectors z sampled at random; in essence, being told to generate a random outcome from its learned distribution. In our architecture, the incompressible noise - z - entries are again sampled *i.i.d* $\sim$ Unif(0, 1), while the featural encoding c is set manually to a desired value - *dosage* in causal inference terminology.

Unlike the incompressible noise z, the featural encoding c has been trained to correspond to consistent outputs; however, it is picked up in an unsupervised fashion. Therefore, this work aims to uncover whether it (and the degree it does) corresponds to observable properties.

Since the consistency of output is only enforced in a loose way, this often only becomes readily apparent when setting the numerical values outside the bounds seen in training, where the primary associated effect begins to dominate (Beguš, 2020, 2021a,c). In this work, we extend this approach into a methodology we call *causal disentanglement with extreme values* (CDEV) (Figure 3). The space available for the featural encodings is limited; hence finding the real-world attributes that map almost one-to-one with the feature space encodings suggest that the generator considers them very important to generating convincing outputs. The number of values reserved for the encoding in the network used is five, which equals the number of distinct coda types shown to the network during training.

In order to perform statistically meaningful experiments, we need to be able to measure our observables algorithmically. Using the *causal disentanglement with extreme values* approach, the output can become relatively noisy as the encoding values c move significantly out of the training range, which was within $[-1, 1]$ for all the five bits. In addition, we observe that for the network used here, the output crystallizes from noise at about the lower edge of the training range (-1), with all lower values (for any of the bits) inducing the generator to output noise, as illustrated in Figure 5.

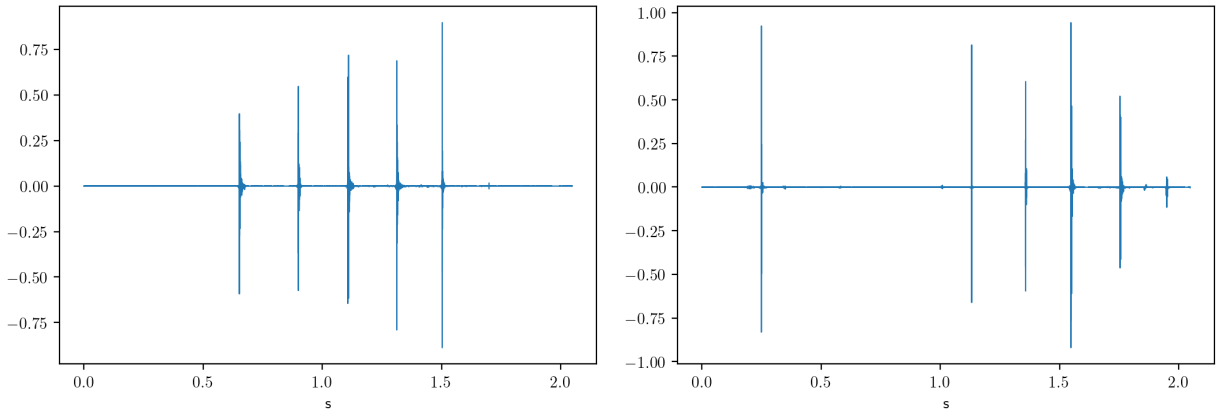


Figure 4: Examples of generated codas.

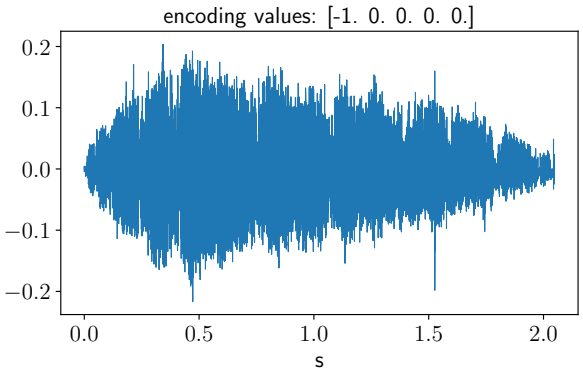


Figure 5: Example of noisy output for encoding values up to the training range limit of -1, in this case, $[-1, 0, 0, 0, 0]$.

For the first two observables considered - the number and regularity of clicks (Sections 4 and 5) — we, therefore, use an algorithmic *click detector* to find local maxima in the generated audio signal that correspond to clicks. It uses minimal thresholds for both the amplitude and the temporal separation to choose between (potentially) multiple sets of detected "clicks". The latter threshold is based on the minimum peak separation possible by the whale physiology and was set to 40 ms. The amplitude threshold is necessary to deal with the residual presence in the dataset of interwoven codas coming from other whales; it was set to 0.4 in terms of the relative amplitude to the peak click and obtained from an analysis of the amplitudes of vocalizations coming from the primary whale and secondary whales in the data.

Even so, as we move towards more extreme encoding values (in the positive direction), the output becomes progressively noisier, as shown for a relatively pathological example in Figure 6. Therefore, the detector also uses signal filtering together with a sub-algorithm that strives to “maximize entropy” by preferring well-spaced peaks over counting a single jagged peak multiple times: i.e., if multiple valid "solutions" given the minimal threshold constraints are still found, it will prefer the more spaced-out solution, which corresponds to our intuition. The final numerical range of the encodings tested: $[-1, 12.5]$ was thus also selected so that the outputs are still reasonably meaningful and the behavior of all the algorithms used for detecting the observables is predictable.

3 The experiment as a continuous-treatment causal inference problem

The methodology used in this work is inspired by causal inference, more specifically, *continuous-treatment* causal inference. In short, the latter deals with an experimental setup where the usual treatment assignment variable Z becomes a continuous variable T, often called a *dose* due to it being common in pharmacological studies. In the case of observational studies, the *propensity score* $e(x)$ thus becomes a function of two variables $r(t, x)$ (Imbens, 2000):

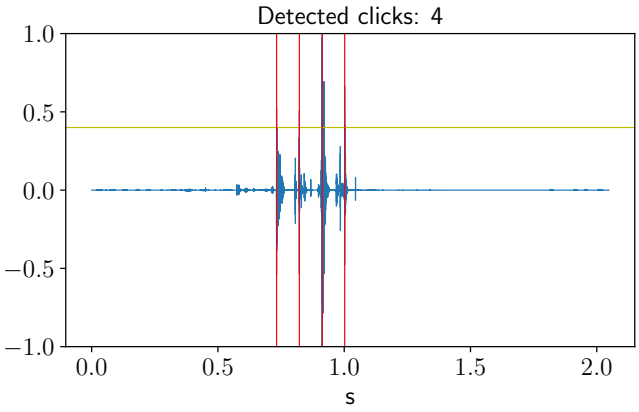


Figure 6: Example of the output of the click detector on an instance of relatively noisy generated data, which can happen when the encodings are set to extreme values. The minor peaks are artifacts picked up in training due to the data not being free of whale dialogue; the yellow horizontal line is the inferred volume level for the vocalization to be coming from the primary whale.

$$e(x) \to r(t, x) = \mathbb{P}(T = t | X = x)$$

Given the assumptions common to the discrete case, the analysis proceeds in much the same fashion, with the treatment dose being discretized at several levels t and the most common quantity of interest again being the difference in expectations of the outcome Y:

$$\mathbb{E}[Y(t)|r(t, X) = r) - \mathbb{E}[Y(s)|r(t, X) = r),$$

relative to some *baseline dose* s. The choice of the latter is natural ($s = 0$) in pharmacology, for example, where no drug being administered carries a special meaning. In our context, the choice of such a baseline is not clear, as will be addressed in this work. The ultimate goal is to obtain the full *outcome curve* for all t with regard to the s chosen.

In terms of causal inference terminology, we thus have the following variables: the *incompressible noise* z part of the output is treated as the 95-dimensional covariate vector X, while the *featural encoding* is considered to be the treatment t and is a five-dimensional vector. The observables considered, such as the number of clicks (output by using the click detector on the generated audio data) for a given X and t, are thus the outcome variable Y_i. From here on, we will prefer the latter causal inference notation in equations over the usual one used when discussing GANs, i.e., z and c.

Each audio output was generated by sampling $X \overset{\text{iid}}{\sim}_{1\ldots95} Unif[-1, 1]$. Such vectors can be treated as unique within the generated sample of audio outputs used for the estimation of a particular quantity and hence correspond to a separate *unit* indexed by i. These were then fed to the generator at each level of the treatment $t \in [-1, 12.5]$ set at each of the five featural bits, meaning the covariates were kept the same across treatment levels. The generated audio was then run through the appropriate detection algorithm, such as the click detector, and the outcome $Y_i(t)$ was observed. The total number of units was kept at $N = 2500$.

In other words, we perform a completely randomized experiment with the added bonus that the outcome is observed at each treatment dose for each unit:

$$\mathbb{P}(T_i = t | X_i = x_i) = 1, \ \forall i$$

This means that there is no *fundamental problem of causal inference* (Holland, 1986) at play here, simplifying the estimation. This relies on the assumption that we *can "re-set" the unit for each dose*, corresponding to the assumption that X and t have enough informational capacity to encode the vast majority of the latent space salient to our outcome - in other words, given X and t, the observable of interest is *consistent*.

Our primary motivation for approaching the problem from the point of view of causal inference is due to being primarily interested in estimating the *amount of an observed effect* in a very complex system without making additional

assumptions or needing access to the internals of the generative process, for which we can borrow several interesting methods of estimation from causal inference. Therefore, their (potential) (dis)agreement is also of interest. Likewise, the main goal of this section is to argue for the applicability of such methods.

The assumption of consistency of the observables has been borne out in similar studies: in a work that generated canary bird vocalizations with GANs (Pagliarini et al., 2021), the authors found that the total variation of the dataset was sufficiently captured with a latent space length of only 3. Since the training dataset in our case was restricted to the few most popular coda types, we have no reason to severely doubt the consistency assumption, given the total available latent space length of 100. This is also apparent by visually and aurally examining the generated data for fixed (X, t) - the differences, if any, are imperceptible. To summarize - given the whole input (X, t), the observable becomes quite deterministic, while, on the other hand, only a fraction of X is actually correlated with the outcome of interest, and that in a highly convoluted way.

It is thus how we square the seemingly opposing concepts of X being *incompressible noise* only serving as an index to the *units* and the fact that conditioning on the whole of X gives us *ignorability.* In other words, the working assumption is that the relation of X to the outcome is so complex that sampling it uniformly at random does not produce a noticeable additional effect besides the treatment t. This can be likened to the concept of *deterministic chaos*, where a non-linear (in X) mapping produces essentially random output, allowing us to perform *completely randomized experiments*. On the other hand, the observables derived from an output generated given a specific (X, t) is consistent, giving us *ignorability* when conditioning on X.

The main part that remains potentially problematic for using causal inference approaches here is that there is less of a guarantee of a *non-interaction* between the treatment and the covariate X. Even though the training enforces t to correspond to consistent output while X can vary without constraints, the process by no means enforces total separation; moreover, what is optimized is actually the mutual information's lower bound (Chen et al., 2016). This corresponds to potential *treatment effect heterogeneity induced by the covariates*. In cases when the true outcome function is linear, this is not an issue if the covariates X are centered $\bar{X} = 0$, which is the case here. However, in our case, the function is non-linear (and highly complex); therefore, a necessary condition for the simpler estimators examined here would be that any possible interaction of the covariates X with the treatment t is an approximately odd function since X are sampled uniformly on $[-1, 1]$.

4 An introduction to the methodology: the number of clicks

The first observable we're interested in is the *number of clicks* in the generated audio. As this quantity is the easiest to visualize, this section will also introduce the methodological approaches used.

4.1 Continuous average treatment effect

The most basic estimator is the usual average treatment effect (ATE), applied separately at each treatment dose value t. Formally, we're interested in the effect relative to some *baseline dose* t':

$$\mathbb{E}[Y(t)] - \mathbb{E}[Y(t')] = \mathbb{E}[\mathbb{E}[Y|X, T = t]] - \mathbb{E}[\mathbb{E}[Y|X, T = t']] \tag{1}$$

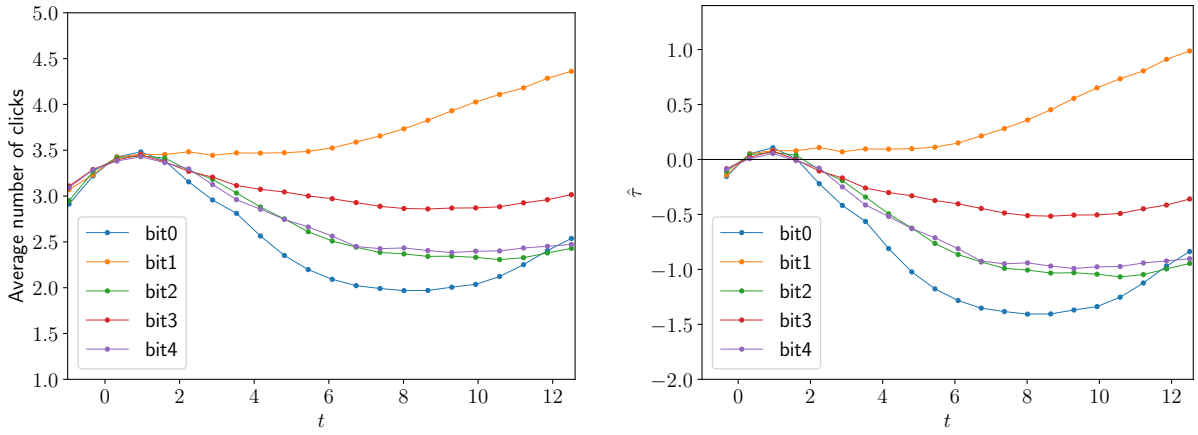
Since the units i are defined by their randomly drawn X, and we observe every $Y_i(t)$, this simply corresponds to the difference in sample averages:

$$\hat{\tau}(t) = \frac{1}{N}\sum_{i=1}^{N} Y_i(t) - \frac{1}{N}\sum_{i=1}^{N} Y_i(t'), \tag{2}$$

This is estimable since both ignorability and overlap hold due to this being a completely randomized experiment, as discussed in Section 3.

At first, the "baseline dose" will be chosen as $t' = [0, 0, 0, 0, 0]$. The treatments correspond to setting the values of single bits in the encoding while keeping the others at zero. However, as mentioned in Section3, this baseline *has no special meaning* since every encoding is meaningful; this will be addressed shortly in Section 4.1.1

The mean number of clicks per treatment level is presented for all bits in Figure 7. The results indicate that bit 1 is the most influential on the number of clicks in the outcome. Using the outcome at 0 as the baseline, the average treatment effect is presented on the right in Figure 7. In addition to the effect of bit 1, we also note a persistence of an effect in bit



(a) The mean number of clicks per dose t level.

(b) The ATE per bit with the baseline at $t = 0$.

Figure 7: The mean number of clicks and the ATE with regard to the baseline $t' = [0, 0, 0, 0, 0]$.

3, which stabilizes at a different "stationary" value. In the following sections, we will examine whether this persists when using different estimators.

We get a fuller picture by examining the standard deviation in the number of clicks, as shown in Figure 8. We again observe the same phenomenon of *disentanglement* of bits other than 1, while the latter takes on the encoding as before. Its real effect is thus the encoding of the *range of clicks* output by the generator.

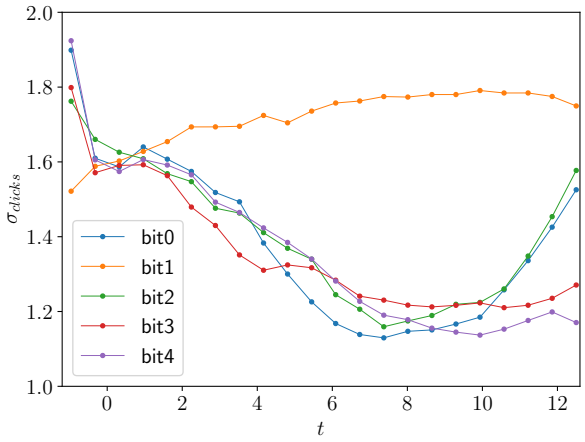


Figure 8: Standard deviation in the number of clicks with increasing value of a particular bit.

4.1.1 Choosing a better baseline

We observe the issue already mentioned in Beguš (2021b) of high entanglement of the learned encodings within the range seen in training ($[-1, 1]$), prompting us to disentangle them by setting their value above that range. In tandem with this, the other bits stabilize at roughly a constant effect, which lends further credence to interpreting this as a *process of disentanglement*: with higher values, the primary encoded effect starts to dominate, while the other bits lose their previous, entangled effect on the number of clicks, which is not their primary associated encoding.

This leads us to consider setting different, more natural values for the baseline. The limits of the training range: -1, where the output coalesces out of pure noise (cf. Figure 5) and +1, where the process of disentanglement begins, serve as logical choices and are shown in Figure 9.

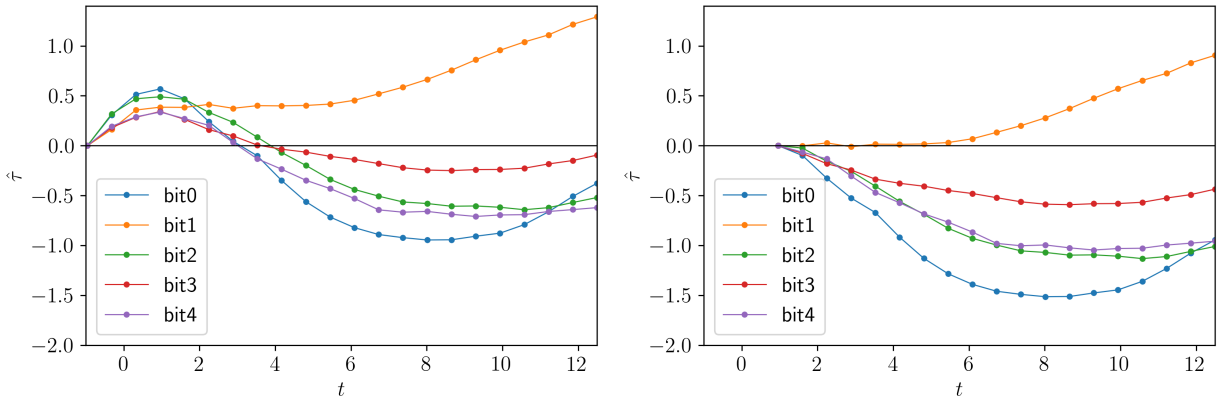


Figure 9: The ATE for the number of clicks with more natural choices of the baseline dose; left: at $t' = -1$, right: at $t' = 1$.

4.2 Incremental causal effect

The question of selecting the appropriate baseline dose is elegantly avoided by using the *incremental causal effect* proposed by Rothenhäusler and Yu (2019) — the effect due to an infinitesimal shift in dosage:

$$\tau_{ICE} = \mathbb{E}[Y(T + \delta)] - \mathbb{E}[Y(T)] \tag{3}$$

Under *local ignorability* and *local overlap,* we have:

$$\mathbb{E}[Y'(t)|T = t, X = x] = \partial_t \mathbb{E}[Y|T = t, X = x], \tag{4}$$

meaning that the right-hand side - the true incremental causal effect - is identifiable by the expectation of the derivative - the left-hand side.

In our case, both assumptions hold trivially: (i) local ignorability holds because strong ignorability holds, and (ii) local overlap holds since the probability of treatment being assigned is a constant — 1. Conceptually, we are again interested in this estimator due to its being agnostic with regard to the generative model, with the added benefit of it being invariant to any choice of a baseline.

In interpreting the results, we are interested in whether the incremental effect for a single bit is prominent and consistent across the whole range under consideration while the others are not; in such a case, we can justifiably argue that that bit encodes the number of clicks in a vocalization.

A potential formal drawback of using this estimator in this setup is the discontinuity of the outcome: the number of clicks is a discrete variable, while the authors assume $Y(t)$ to be continuous, meaning that we are formally estimating a discontinuous derivative. However, the finite differences used are naturally always defined. Furthermore, this does not apply to the other observables evaluated in this work.

Its estimator — $\hat{\tau}_{ICE}$ — is the usual sample mean of the numeric differences in the outcomes for single units. It is presented for two bits in Figure 10. We again observe the phenomenon of disentanglement outside the training range, where the effect of bit 1 remains consistently positive. In contrast, that of bit 4, for example, returns to an oscillation around zero.

Given the expectation of the derivative curve, we can define the finite sample *expected effect of an infinitesimal increase:*

$$\hat{\theta}_{fs} := \frac{1}{N_t} \sum_{i_t=1}^{N_t} \mathbb{E}[Y'(t_{i_t})|T = t_{i_t}, X = x_i] = \frac{1}{N_t} \frac{1}{N} \sum_{i_t=1}^{N_t} \sum_{i=1}^{N} \frac{\Delta Y_i(t_{i_t})}{\Delta t_{i_t}}, \tag{5}$$

where N_t is the number of examined treatment levels with the corresponding t_{i_t}, and N is the number of units with the corresponding x_i, each receiving each treatment level.

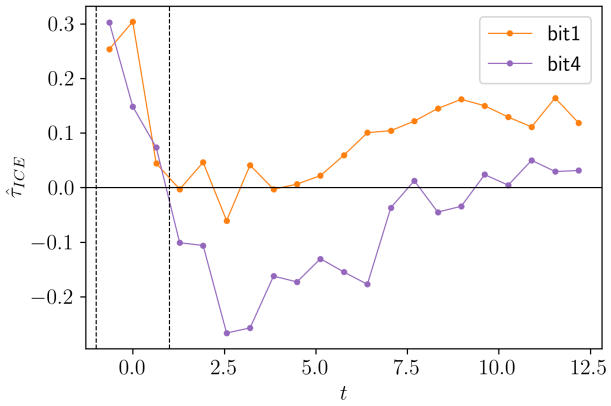


Figure 10: The incremental causal effect on the number of clicks across the range for two bits; the training range is delineated by dashed lines.

In short, this is the *expected overall effect* if all the current treatments were infinitesimally shifted (in the positive direction) for the given sample. While formally different from $\mathbb{E}[Y'(t)]$, this meaning is well-defined and very helpful to the problem at hand.

Moreover, it is a single metric and can be interpreted as an analog to the *average treatment effect* for binary treatments without the need for a baseline dose as a reference. Its values are presented in Table 1, corroborating the results obtained with the simple average treatment effect.

Table 1: The expected effect of an infinitesimal increase in the treatment on the number of clicks.

bit	0	1	2	3	4
$\hat{\theta}_{fs}$	-0.028	0.096	-0.039	-0.007	-0.046
$\hat{\theta}_{fs} \vert t \geq 1$	-0.082	0.079	-0.087	-0.038	-0.083

We again observe the *disentangling* phenomenon where bit 1 stands out in its effect while the others are grouped together in a smaller, opposite effect. We thus have another indication that bit 1 primarily encodes the number of clicks. As before, bit 3 retains a measure of influence on the outcome when using this estimator, as well, as evidenced by its relative distance from bits 0, 2, and 4.

4.3 Regressing the outcomes directly

An additional approach is to estimate the outcome with machine learning methods, as done for observational studies in Athey and Imbens (2015). There, the first step is the estimation of the propensity score, which we can skip here since the latter is always 1. What remains is the regression of the outcome on the treatment and covariate for each unit.

This means we regress the *individual* observed outcomes $Y_i(t)$ (in this section, the number of clicks) against the full input vector for each unit at each level t, i.e., for bit 1: $x_i = [0, t, 0, 0, 0 \vert\, z_i]$. The *outcome curve* in this setting corresponds to the *mean* of the individual inferred outcomes at each treatment level.

For this task, we've chosen boosted tree regression as implemented in the `lightgbm` package (Ke et al., 2017). This method combines extraordinary flexibility with a greater degree of interpretability as compared to, e.g., using an additional neural net to regress the outcome curve. `lightgbm` trees differ from other gradient-boosted trees in that the trees are grown by leaves, as opposed to depth-first. Hence the main parameter corresponding to model complexity is the (maximum allowed) number of leaves in the base learner.

We intentionally use no sparsity-inducing regularization to prevent overfitting: we wish to check if the hypothesized encodings are consistently picked up by a completely unrelated non-parametric method that is (i) able to approximate any function arbitrarily closely with sufficient model complexity (ii) when unregularized, is prone to overfitting. Hence we are looking for *consistency of explanation* across varying model complexity, *despite all odds*.

For determining feature importance, we use values obtained from SHAP (Lundberg and Lee, 2017), which uses a game-theoretic concept called the *Shapley value* to distribute attribution to the outcome to the input features. Note that while one could, in theory, use SHAP on the generative model itself, it is much more computationally efficient when

used with trees (Lundberg et al., 2018). As mentioned, we additionally wish to test our findings via an unrelated method as opposed to disassembling the generative network.

Most of the hyperparameters except the main one - the number of leaves - were chosen by an early stopping of the training as determined by a separate validation set; the MSE on this set for trees with different numbers of leaves is presented on the right in Figure 11 and serves as the final determinant of what we consider the optimal model.

We would like to re-emphasize that the goal here is the inference of the effect of particular parts of the input by way of letting an unrelated, expressive model "test out the competing hypotheses" itself. Specifically, we would like to see such models of adequate but varying complexity *consistently* assign the credit for the outcome to the encodings suggested by the other methods, despite not being prohibited from picking up spurious relationships. This differs somewhat from other uses of such methods in causal inference, where correctly accounting for unobserved outcomes takes precedence (Athey and Imbens, 2016; Künzel et al., 2019).

The figure on the left in Figure 11 shows the inferred *outcome curve* as predicted by the model with a maximum of 13 leaves per tree (the best model in terms of the MSE) using the whole of the inputs (i.e., including the validation set, which had only been used as a stopping criterion). We can reasonably say that the model is able to approximate the real outcomes sufficiently well.

The SHAP plot in Figure 12 corroborates the results obtained with the other estimators, picking up bit 1 as the most salient feature in its effect on the number of clicks for each unit, with the same relationship. This confirms that the uncovered relationship between the encoding bit and the outcome is not spurious — i.e., the outcome being just as related to the *incompressible noise* part of the input. Lower values are "bunched" together in a slightly negative effect due to residual entanglement at the lower end of the treatment values. Additionally, the consistency of explanation can be somewhat observed here by the homogeneity of the coloring — i.e., the lack of dots representing high values in the region of low impact.

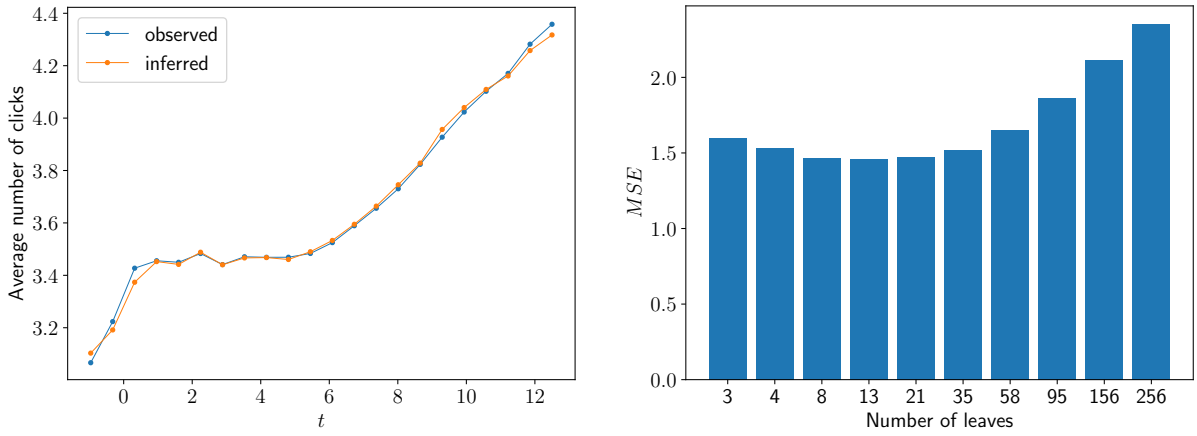


Figure 11: Right: MSE by model complexity as determined by the maximum number of leaves allowed. Left: the outcome curve derived from the data (the mean number of clicks, cf. Fig. 7) and the one inferred by the best fitting model in terms of the MSE.

The consistency of explanation can be better observed in Figure 13, which presents a *heatmap* plot as produced by the `shap` package, with the units being ordered on the x-axis in terms of increasing treatment value. The y-axis displays the features ordered in terms of their overall importance as measured by SHAP. The plot above the central heatmap plot is the output of the model centered around the explanation's mean value, and the bars on the right display the feature's cumulative contribution.

The features for individual units are colored in terms of their contribution to the outcome. For bits encoding the observable, we expect to see *local consistency* in the assigned effects since the units are ordered by increasing values of t. Conversely, we should observe much less consistently assigned effects for the other bits once the process of disentanglement has reached stationary values. A point we'd like to raise here is that the *CDEV* process does not necessarily start at exactly the same values of treatment for different quantities, so we might not observe full disentanglement for all observables presented.

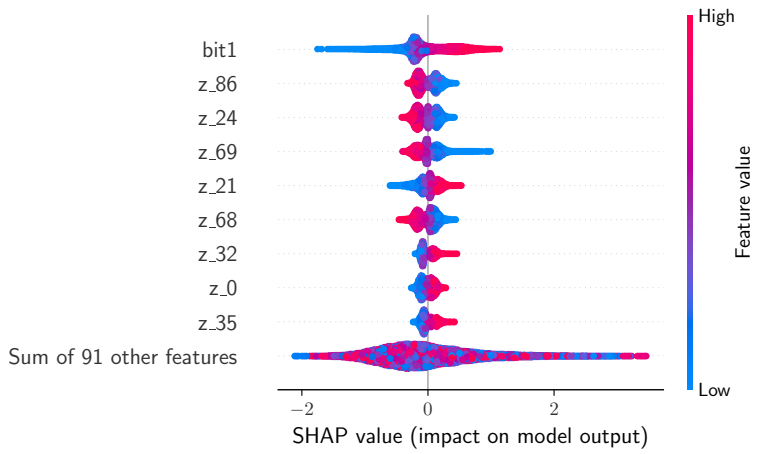


Figure 12: SHAP values for the best fitting model for bit 1.

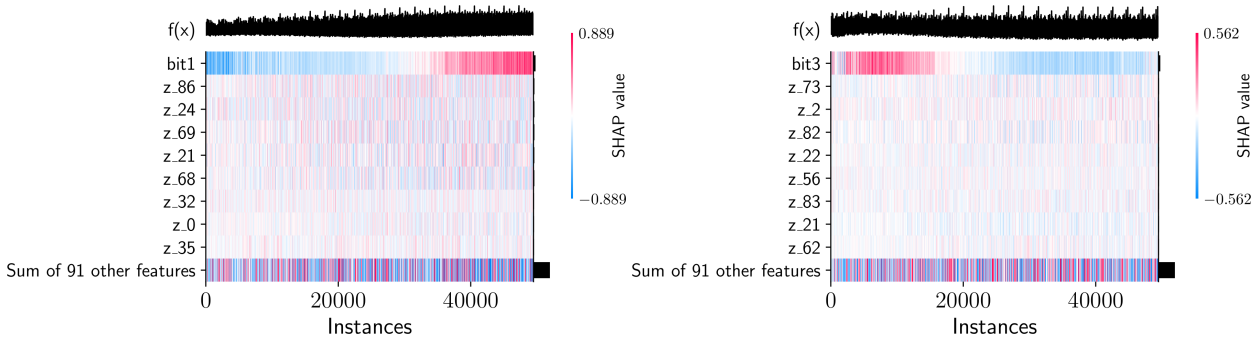


Figure 13: SHAP *heatmap* for best-fitting models for bit 1 (left) and bit 3 (right). The instances are ordered in terms of increasing t from left to right.

In Figure 13, we thus observe a greater degree of local consistency in the high-end of the treatment value of the bit encoding the property — bit 1 — as opposed to bit 3, which has reached the stationary state of disentanglement.

5 Click spacing and regularity

In this and the following section, we will apply this methodology to uncover additional properties of the communication system that the network encodes as meaningful. In behavior, the encoding works similarly to the one presented for the number of clicks: one bit is assigned an encoding and has an observable effect on the quantity, while the rest move in relative unison to a *stationary value*.

The observables we are interested in in this section are the *click spacing*, as measured by the mean *inter-click interval* (ICI) of a coda, and the *coda regularity*, which we measure by the standard deviation of the inter-click intervals within a coda. Since these quantities are not completely independent of the overall coda length, we stratify the results by the number of clicks observed.

5.1 Average treatment effect

Figure 14 shows the ATE with regard to the baseline at -1 and 1 for the mean ICI for bits 1 and 2. We again observe a consistent effect in the value of bit 1, while bit 2 (and others not shown in the figure) do not display any consistency across varying numbers of clicks. Since the former also encodes the range of the number of clicks, as shown in Section 4, the data for codas with high numbers of clicks is almost impossible to get while keeping the value of bit 1 at 0.

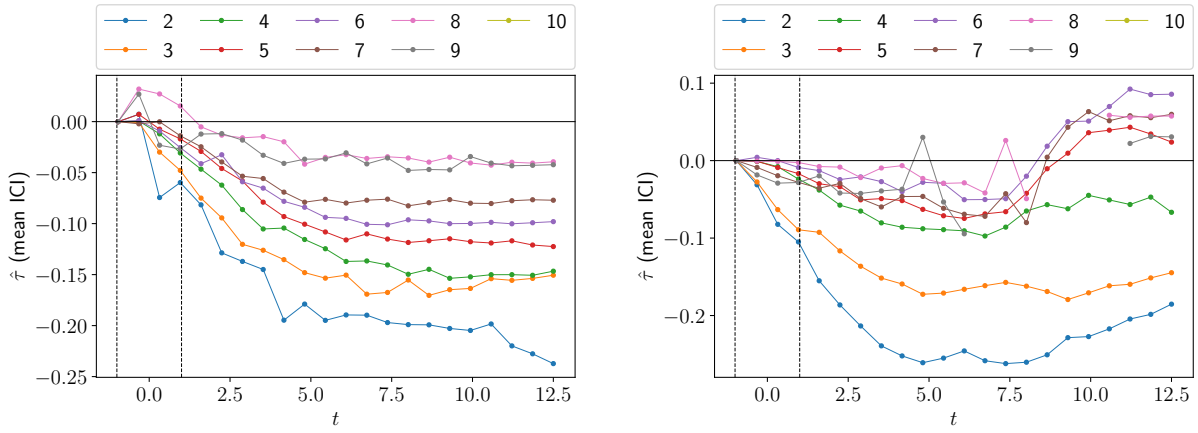


Figure 14: **Mean ICI ATE**, stratified by the number of clicks. Left: bit 1, right: bit 2. We observe a consistent, monotonic effect for bit 1 regardless of the number of clicks observed, while for, e.g., bit 2, the behavior is inconsistent. The missing high number of clicks data in the latter is simply due to the value of bit 1 being too low.

Note that after stratifying by the number of clicks, the overall coda length should not act as a constraint on the mean of the intervals - the network architecture supported outputs up to 2s in coda duration, which was not reached by any of the codas generated; or, indeed, any in the dataset (see Figure S1 in (Gero et al., 2016a).

We present the ATE for the inter-click distance regularity (i.e., coda regularity) in Figure 15 for both baselines.

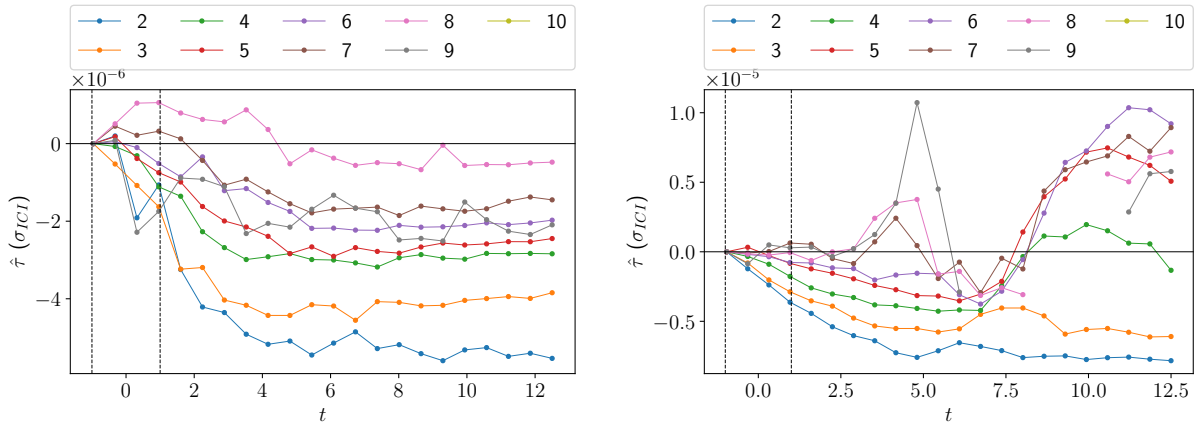


Figure 15: **Coda regularity ATE**, stratified by the number of clicks. Left: bit 1, right: bit 2. In terms of coda regularity, as measured by the standard deviation of the inter-click intervals in a coda, we again observe a consistent, monotonic relationship with the encoding of bit 1 across codas with varying numbers of clicks, while other bits carry a much more indeterminate behavior.

We observe that bit 1 additionally encodes an increasing *coda regularity* (i.e., decreasing variance in the spacings between clicks) across all coda types (proxied here by the number of clicks). Since it also encodes the number of clicks, this implies that the generator has learned to connect these two properties: the codas with a higher number of clicks are more regular, with the clicks being more closely spaced together. This is especially poignant since the same property holds for actual whale codas. Gero et al. (Gero et al., 2016a) results suggest that codas reach a limit in duration around 2s long and that as coda length in clicks increases, mean ICI decreases to fit clicks within this duration limit, which appears to be the result of avoiding overlapping with the next coda within an exchange between whales. Furthermore, codas with more clicks are often more regular in their ICIs regardless of their duration. The generator has inferred this connection *without the codas with a high (>5) number of clicks even being present in the dataset* due to data quality limitations (cf. Section 1) and encoded it in the limited space (5 bits) it has reserved for encodings.

In this, we again observe the remarkable propensity of generative adversarial networks to discover hidden structures of the data and innovate in semantically meaningful ways, as shown, for instance, in Beguš (2021b).

5.2 Expected effect of an infinitesimal increase

Table 2 shows the expected effect of an infinitesimal increase (Eq. 5) in bit values on the mean ICI, stratified by the overall number of clicks in the coda.

# clicks bit	2	3	4	5	6	7	8	9	10	11
0	-0.013	-0.002	0.007	0.007	0.005	0.004	0.004	0.009	0.008	N/A
1	-0.018	-0.011	-0.011	-0.009	-0.007	-0.006	-0.003	-0.003	-0.003	-0.007
2	-0.014	-0.011	-0.005	0.002	0.006	0.004	-0.005	-0.010	-0.027	-0.008
3	-0.017	-0.012	-0.006	-0.004	0.000	0.001	0.003	-0.002	-0.008	N/A
4	-0.022	-0.017	-0.011	-0.002	0.004	0.005	-0.001	-0.006	-0.001	0.000

Table 2: The expected effect of an infinitesimal increase in the treatment on the mean inter-click distance.

Since comparing the effects across differing numbers of clicks can be somewhat harder to summarize, we can get an overall picture of the effect of a particular bit by simply looking at whether the effect is positive or negative for generated codas with a particular number of clicks. This is presented as the *sign score* in Table 3.

bit	0	1	2	3	4
sign score	4	-10	-4	-4	-4

Table 3: Overall *sign score* of an infinitesimal increase in the treatment on the mean inter-click distance.

Similarly, we show the results for coda regularity in Table 4 and the corresponding *sign score* in Table 5.

# clicks bit	2	3	4	5	6	7	8	9	10	11
0	-0.004	0.005	0.023	0.023	0.019	0.013	0.016	0.037	0.019	N/A
1	-0.013	-0.009	-0.007	-0.006	-0.005	-0.003	-0.001	-0.005	-0.007	-0.005
2	-0.019	-0.014	-0.003	0.012	0.022	0.021	-0.004	0	-0.072	-0.074
3	-0.014	-0.009	-0.006	-0.003	-0.003	-0.001	0.001	-0.013	-0.004	N/A
4	-0.017	-0.012	-0.005	0.012	0.016	0.016	-0.01	-0.006	-0.014	0.015

Table 4: The expected effect on an infinitesimal increase in the treatment on the standard deviation of the ICIs.

bit	0	1	2	3	4
sign score	6	-10	-2	-8	-2

Table 5: Overall *sign score* of an infinitesimal increase in the treatment on the inter-click interval standard deviation.

This estimator confirms that bit 1 is consistent in its encoding of both click spacing and regularity, regardless of the overall number of clicks in the generated codas. While bit 3 again remains slightly entangled for the latter quantity, its corresponding values of the expected effects are smaller than those of bit 1.

5.3 Direct regression

We again apply the methodology discussed in Section 4.3. The observations and corresponding inputs are again stratified by the number of clicks and regressed separately with boosted tree ensembles of varying complexity. We

are investigating whether such models corroborate the encodings discovered by the preceding estimators. Due to this separate regression, it is somewhat more challenging to visually present consistency across multiple numbers of clicks, as we could in the prior section. For the two bits - 1 and 2 - compared in Figure 14, we show the corresponding out-of-sample MSEs for the coda regularity regression as a function of model complexity in Figure 16. The figure demonstrates that more complex models are increasingly finding spurious relationships for bit 2, while this is not the case for bit 1, which the preceding estimators suggest encodes the coda regularity. This is another way of saying that the relationship is consistent in bit 1, while not in bit 2.

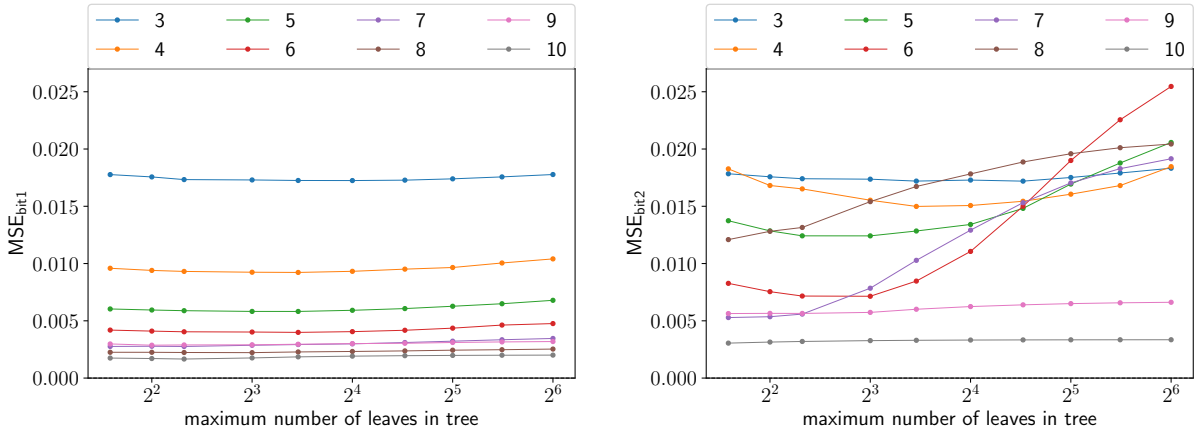


Figure 16: Validation set MSEs for left: bit 1, right: bit 2 for the coda regularity regression (cf. Fig. 15).

Figure 17 displays the associated SHAP values for the best-fitting models, restricted to codas with five clicks due to the above-mentioned presentation issue. We can again make use of the *heatmap* plots to evaluate the degree the *disentanglement* process has taken place within the range examined. Figure 17 presents the results for the same two bits as before for the mean ICI, while Figure 18 does the same for coda regularity.

For the mean ICIs, shown in Figure 17, we observe the effect suggested by the preceding estimators. We additionally observe greater *consistency* with regard to the expected value in local neighborhoods for bit 1 but not for bit 2, as evidenced by interchanging positive and negative contributions for units with the same treatment value.

Similarly, for coda regularity shown in Figure 18, we again confirm the suggested effect and observe a greater *consistency* of the effect with regard to the expected value in local neighborhoods for bit 1.

Overall, the disentanglement process (as seen in the *heatmap* plots) seems to have progressed further for the mean ICIs than for the coda regularity. The *local inconsistency* of the assigned SHAP values in the bits *not* primarily encoding the quantity is a sign of the bit being almost completely disentangled; we can still, however, pronounce the bit as *not* encoding the quantity in question if it (i) displays the observed behavior, i.e., moving in unison with other bits, and (ii) is inconsistent across codas with a varying number of clicks, as shown in the two preceding sections, as well as with varying model complexity, as illustrated in Figure 16.

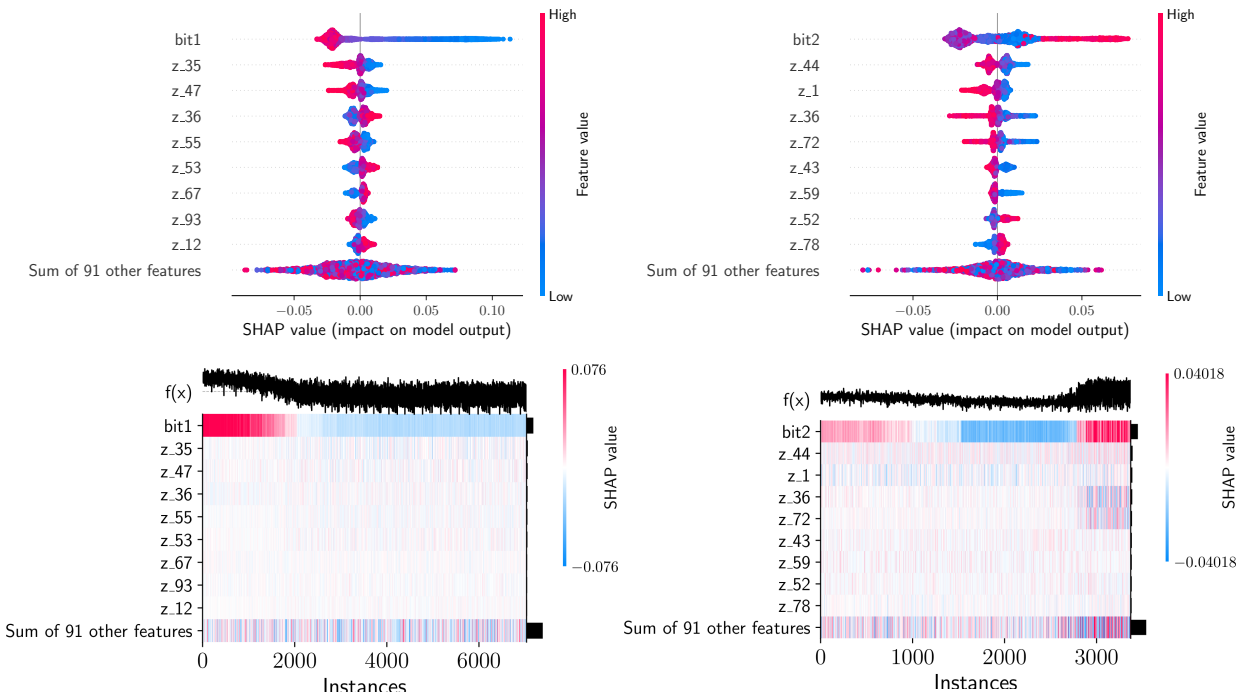


Figure 17: **Coda mean ICI**: SHAP values for codas with **5 clicks** for the best-fitting models for *left*: bit 1 and *right:* bit 2. Bottom: heatmap plots show the effects of individual units ordered in terms of increasing t from left to right. Note that the number of outcomes with 5 clicks is different across the two experiments.

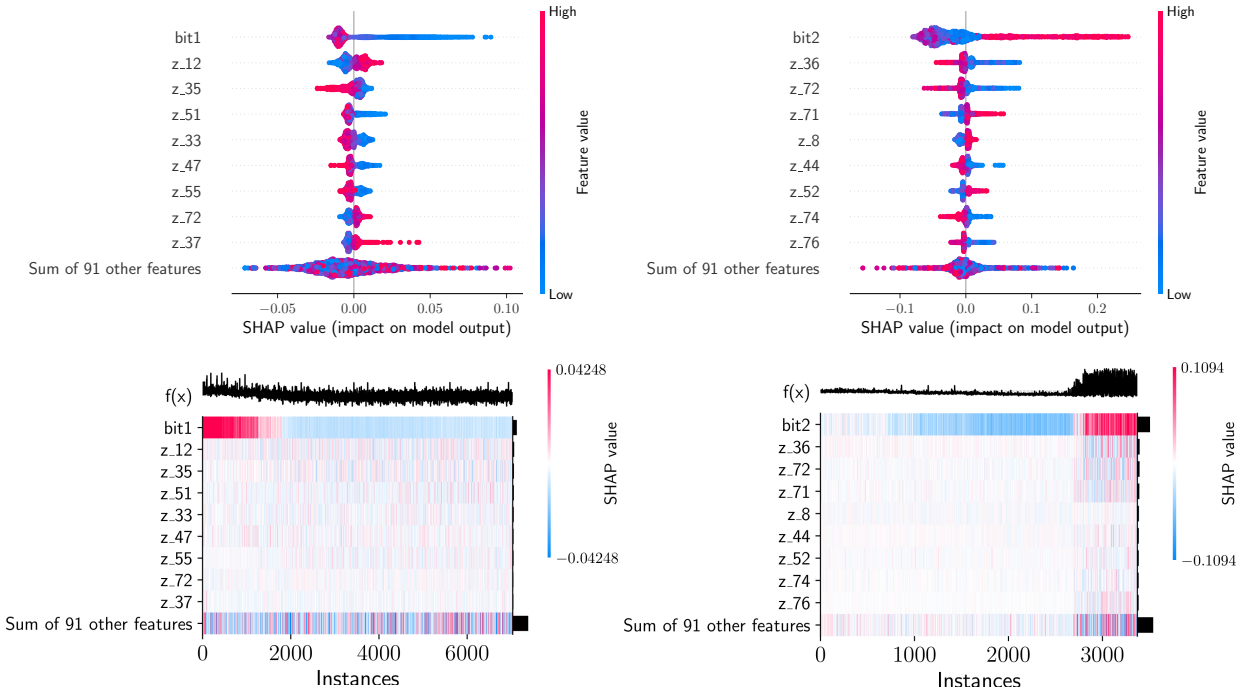


Figure 18: **Coda regularity**: SHAP values for codas with five clicks for the best-fitting models for *left*: bit 1 and *right:* bit 2. Bottom: heatmap plots show the effects of individual units ordered in terms of increasing t from left to right. Note that the number of outcomes with 5 clicks is different across the two experiments.

6 Acoustic properties

We now apply the methodology to acoustic quantities, as captured by the spectra at either the coda or click level. So far, little is known (or has been hypothesized) about the informational content of the acoustic properties of whale communication.

Before calculating the spectra, we apply a slight high-pass filter to the generated data, similar to what Bermant et al. (2019) applied to the real data.

6.1 Average treatment effect

6.1.1 Mean spectral frequency

The first quantity we consider is the mean spectral frequency with regard to the treatment value, either at the coda or click level. In the case of the former, this is the mean frequency of the coda-level periodogram, computed with an added Hamming window. At the click level, we first isolated the clicks with our click detector algorithm, then applied the same on the extracted audio slices. The quantity considered in this case is the average spectral mean across all the per-click means in a particular generated coda. The results and corresponding ATEs are displayed in Figure 19.

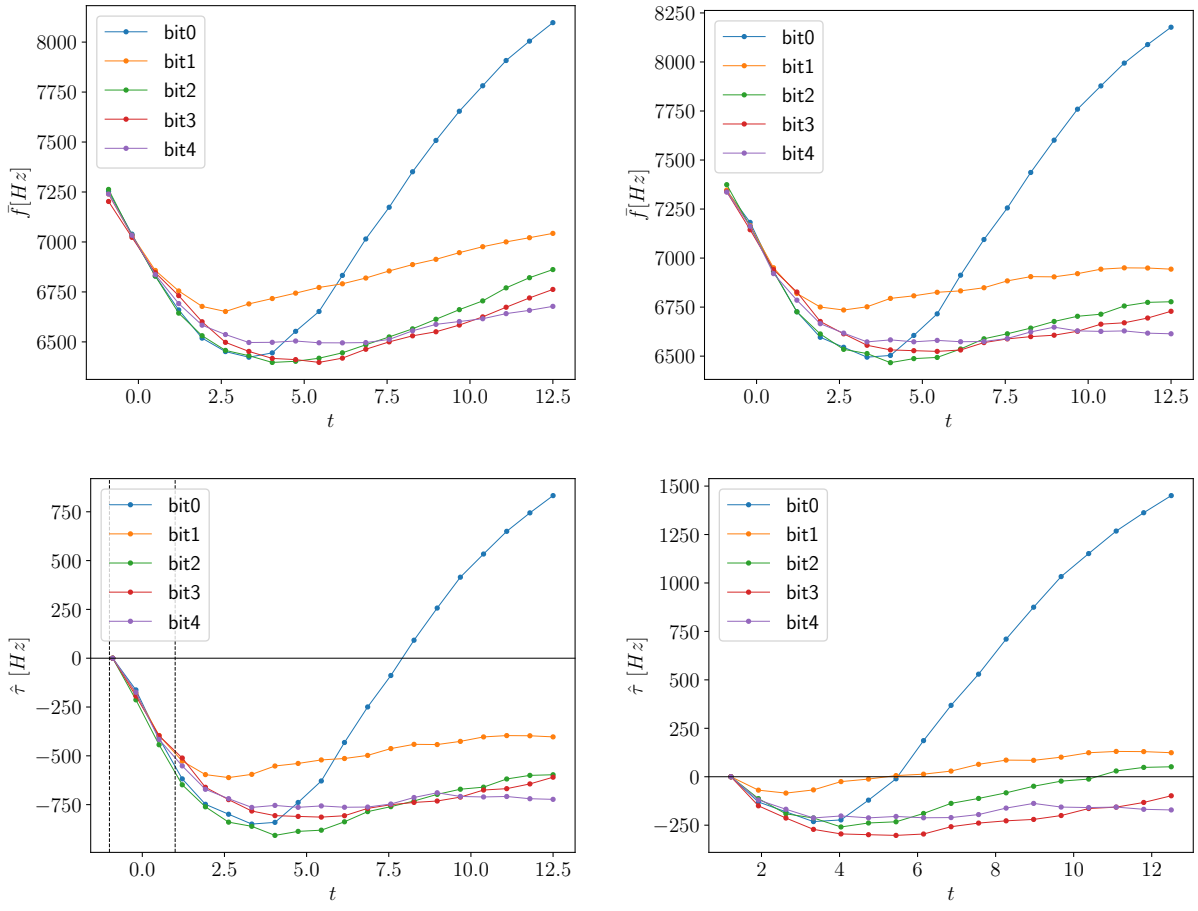


Figure 19: Top: left — coda spectral mean, right — average click spectral mean. Bottom: The corresponding ATE at the click level (i.e., top-right) for the baseline at -1 (left) and $+1$ (right). The encoding seems to be picked up by bit 0.

We again observe the familiar pattern corresponding to *CDEV* — one bit has a positive effect on the observable quantity while all the rest have a common, less pronounced negative effect. As before, we thus hypothesize that *bit 0 encodes the spectral mean*. The difference in the coda- and average click-level means is negligible, as is to be expected, leaving us to concentrate on the average click-level quantities for estimation. As before, we will give further credence to this hypothesis by applying the other estimation approaches.

The potential for the spectral means of codas to be meaningful has not been hypothesized in previous research, but the hypothesis is not completely ungrounded. In a recent paper, Madsen et al. (2023) suggest that odontocetes can vocalize in different registers and that their articulators are sufficiently flexible for such differences to be possible.

6.1.2 Acoustic regularity

We are also interested in *acoustic regularity*, which we measure by the standard deviation of the click spectral means *within* each generated coda. The per-click spectra were estimated as discussed in Section 6.1.1. The results are presented in Figure 20.

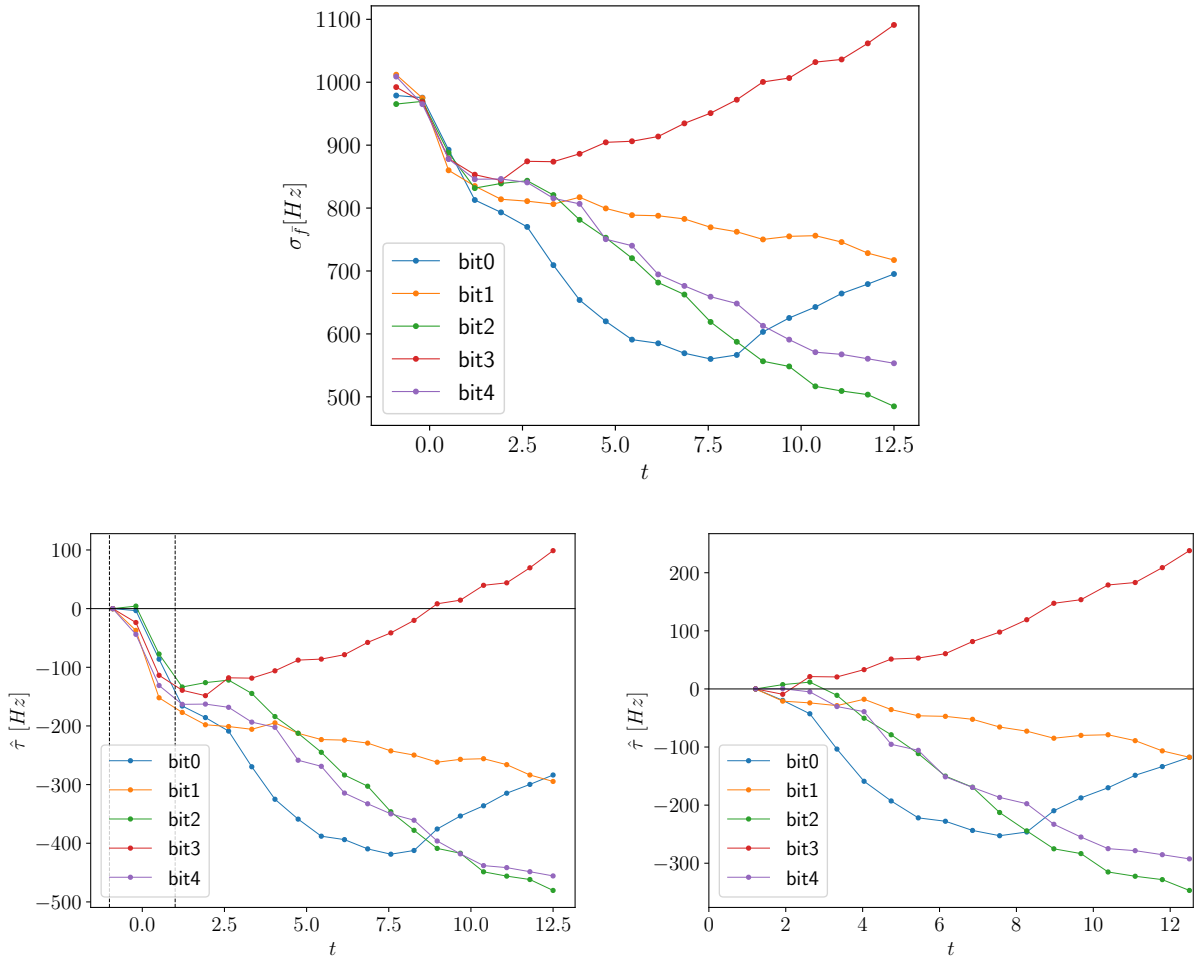


Figure 20: Top: Average click spectral mean (per coda) standard deviation. Bottom: The corresponding ATE for the baseline at -1 (left) and $+1$ (right). The encoding seems to be picked up by bit 3.

The results suggest an additional meaningful encoding: bit 3 appears to encode the within-coda *acoustic regularity* of the output. It can, additionally, be suspected that bits 1, 2, and 4 have not yet reached stationary values of their disentanglement. Unfortunately, we are limited in setting the upper limit of our treatment since, much like on the negative side (cf. Fig. 5), the output becomes too noisy above a certain level for meaningful estimation.

6.2 Wasserstein mean spectral distances

Since the spectra are distributions themselves, we can measure the effect of applying treatments in specific bits of the encoding on the *overall acoustic output*, as captured by the average coda-level spectrum (across N units with random X). The Wasserstein distances from the baseline at -1 (where the output coalesces from noise) and at $+1$ (the upper

limit of the training range) of the average coda-level spectrum for the output generated by setting the corresponding bit to the treatment value t are presented in Figure 21.

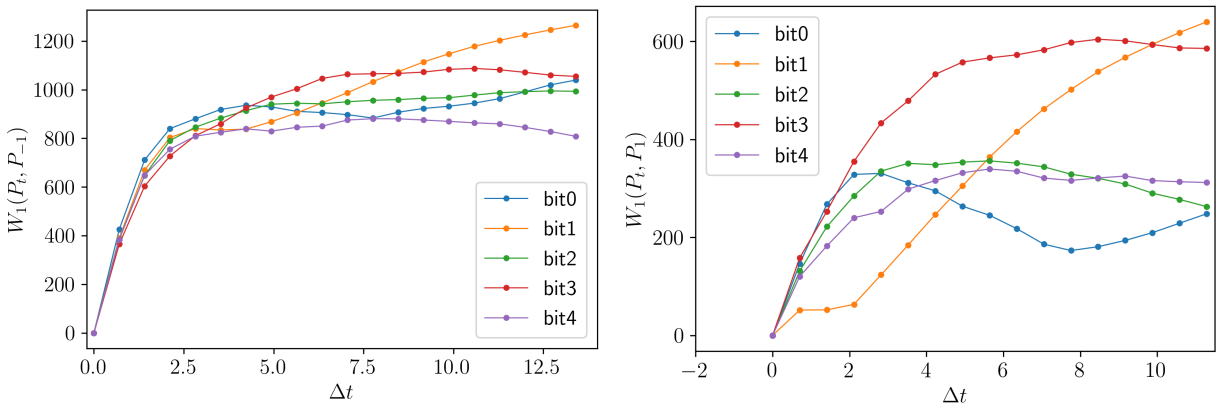


Figure 21: Average spectrum Wasserstein distance to the spectrum at the value (-1) (left) and $+1$ (right) of the corresponding bit.

The result shows a relative spectral stabilization slightly above the limit of the training range for bits 2 and 4. Very notably, bits 0, 1, and 3, for which we have uncovered observable effects - spectral mean, click number (range) and regularity, and spectral regularity, respectively, continue to evolve their average spectra with further treatment. The fact that the start of the disentanglement process is somewhat shifted for acoustic properties was evident in the ATE estimation in the preceding section, as well. This is also the reason that the average spectra for increasing bit 0 are generally more similar to the average spectrum at $t = +1$ (right figure) (cf. Fig. 19).

6.3 Expected effect of an infinitesimal increase

Table 6 presents the results for the expected effect of an infinitesimal increase on the coda and click mean frequency, both for the overall range of treatments and outside the training range, where the encoding becomes disentangled. The results corroborate the results obtained via the ATE estimator since the effect of bit 0 dominates the others.

Table 6: The expected effect of an infinitesimal increase in the treatment on the average coda/click mean frequency

bit	coda $\hat{\theta}_{fs}(\bar{f})$	click $\hat{\theta}_{fs}(\bar{f})$	coda $\hat{\theta}_{fs}(\bar{f})\|t \geq 1$	click $\hat{\theta}_{fs}(\bar{f})\|t \geq 1$
0	63.07	62.15	127.37	128.58
1	-14.73	-30.07	25.52	10.99
2	-29.93	-44.55	19.31	4.56
3	-32.85	-45.48	2.75	-8.71
4	-41.93	-53.95	-1.24	-0.35

Similarly, Table 7 shows bit 3 to be the outlier, with bits 0 an 1 (for which we have uncovered associated encodings, hence having an "unintended" effect on the spectral mean standard deviation) also being relative outliers to the baseline of *disentanglement* evident in bits 2 and 4.

Table 7: The expected effect of an infinitesimal increase in the treatment on the average click mean frequency standard deviation within a coda.

bit	$\hat{\theta}_{fs}(\bar{\sigma}_f)$	$\hat{\theta}_{fs}(\bar{\sigma}_f)\|t \geq 1$
0	-21.17	-10.42
1	-21.99	-10.42
2	-35.85	-30.74
3	7.37	21.09
4	-34.01	-28.64

The clicks matched to an observable effect also mostly show a more pronounced infinitesimal causal effect on the *Wasserstein* distance between the average spectra relative to the point where the respective bit value is set to -1 and +1, as

shown in Table 8. The result for bit 0 is an outlier due to the fact that for the click mean frequency, the disentanglement only picks up with relatively higher values of t (cf. Fig. 19).

Table 8: The expected effect of an infinitesimal increase in the treatment on the Wasserstein distance between average spectra relative to $t = -1$ and $t = +1$

bit	$\hat{\theta}_{fs}(W_1(P_t, P_{-1}))$	$\hat{\theta}_{fs}(W_1(P_t, P_1))$
0	77.57	22.00
1	94.37	94.37
2	74.14	74.14
3	78.70	78.70
4	60.32	60.32

6.4 Direct regression

Similarly, we can regress all the observed outcomes Y_i against all the inputs X_i using gradient-boosted trees. In Figure 22, we again observe the expected patterns when comparing the SHAP values obtained from the best-performing models for the bit that picks up the encoding of spectral mean — bit 0 — and bit 4, which does not. As usual, the inferred effects move in opposite directions with regard to the expected (in terms of SHAP) outcome, with the effects becoming random in bit 4 once the process of *disentanglement* is complete.

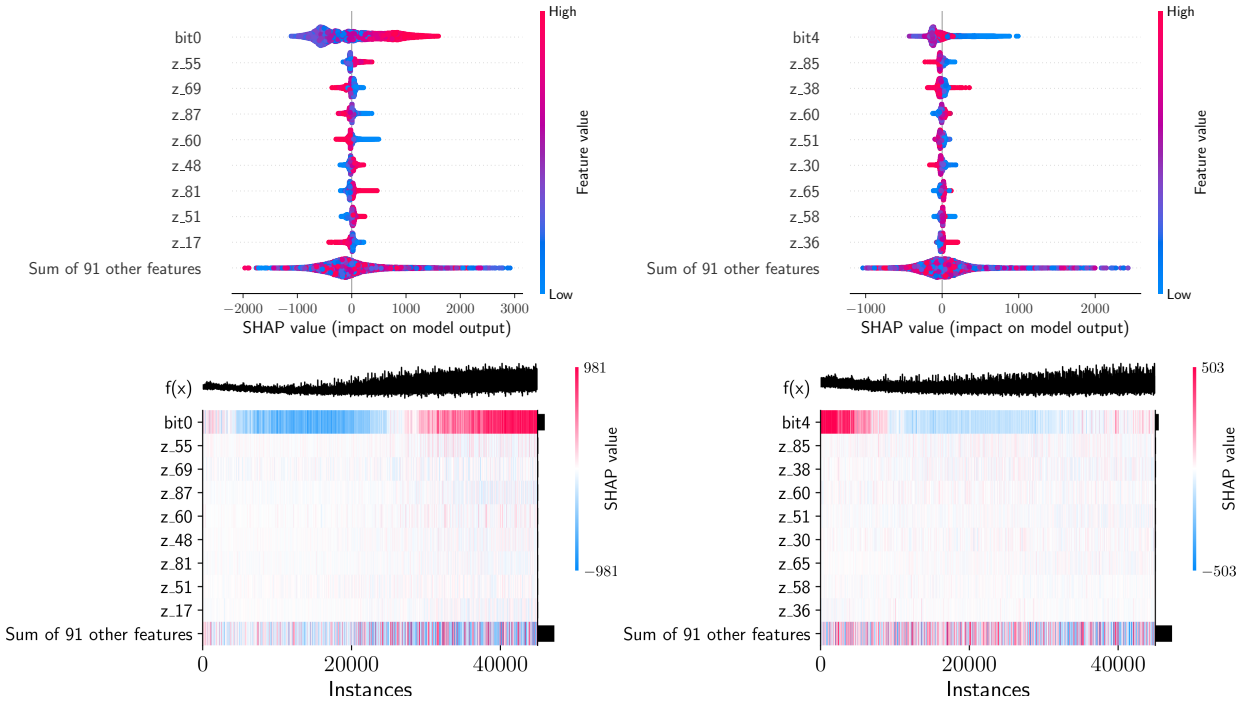


Figure 22: **Spectral mean**: comparison in SHAP values between *left*: bit 0, and *right* bit 4. Bottom: *heatmap* shows individual units in terms of their increasing value of t from left to right.

Conversely, the results for the spectral regularity shown in Figure 23 do not display the common randomness of assigned effect for values with high treatment in bit 4 — the bit that does not pick up the encoding, meaning that the process of disentanglement has not reached the stationary value yet, as we suspected when discussing the ATE results (Sec. 6.1.1). Nonetheless, this does not affect the conclusion that bit 3 seems to encode the spectral regularity of a coda, as evidenced by the agreement of all three estimation approaches.

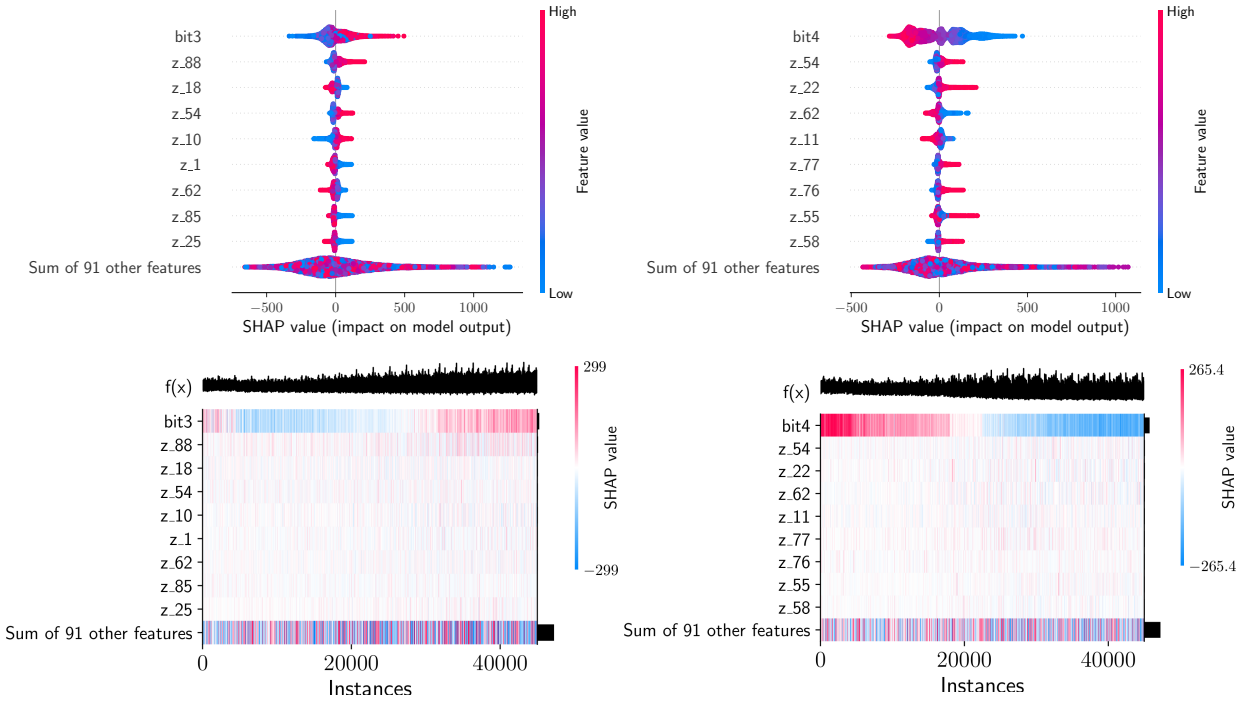


Figure 23: **Spectral regularity**: comparison in SHAP values between *left*: bit 3, and *right* bit 4. Bottom: *heatmap* shows individual units in terms of their increasing value of t from left to right.

7 Conclusion

In this paper, we have presented a model-agnostic approach to uncover the observable properties a deep generative model encodes as meaningful as a way of gleaning information from data that is alien to us in the true sense of the word: the vocalizations of sperm whales. In this, we leverage the power of information-theoretic GANs to encode semantically meaningful properties in a completely unsupervised fashion. Since the model is constrained in the number of such encodings it can learn, we can argue that it must consider these critical to its ability to generate data.

To uncover these properties, we consider the trained model as an experiment and use a technique we call *causal disentanglement with extreme values* to facilitate the discovery of the encodings. We present three independent methods inspired by causal inference that enable us to consistently pair up particular bits of the encoding with a physically observable property of the communication system. The agreement between the methods gives further credence to the results.

With this setup, we confirm that the number of clicks, which is what the existing coda classification developed by marine biologists is primarily based upon, indeed seems to be a fundamental property of the communication system. This can be seen as a good grounding point with regard to the credibility of the approach.

Since generative adversarial networks are well known for their innovative generation of examples not seen in the data, we discover that the network correctly associates synthetic codas with a high number of clicks with their increasing regularity by encoding both properties simultaneously, despite scarcely having had access to such codas in the training data. Thus, it correctly infers a property of unseen real-life codas, illustrating its ability to learn the hidden structure of the data.

Using the proposed technique, we also uncover that two acoustic properties might be meaningful in the sperm whale communication system: (i) the coda spectral mean and (ii) the regularity in the spectra across the clicks within a coda (coda spectral regularity). These two properties have not been considered significant by previous research. This could serve as an indicator that the sperm whale communication system is not merely a Morse-like code where only the number of clicks and the intervals between them are meaningful, but that whales actively control the acoustic properties of their vocalizations and encode meaning in those acoustic properties.

Finally, the methodology presented could be applied to any problem for which one would like to leverage the immense expressiveness of models such as GANs as a way of consistently discovering what properties of the data are semantically meaningful.

Acknowledgments

This study was funded by Project CETI via grants from Dalio Philanthropies and Ocean X; Sea Grape Foundation; Rosamund Zander/Hansjorg Wyss, Chris Anderson/Jacqueline Novogratz through The Audacious Project: a collaborative funding initiative housed at TED.

Data was collected through fieldwork for The Dominica Sperm Whale Project undertaken through permits from Fisheries Division of the Government of Dominica. Research was funded through a FNU fellowship for the Danish Council for Independent Research supplemented by a Sapere Aude Research Talent Award, a Carlsberg Foundation expedition grant, a grant from Focused on Nature, two Explorer Grants from the National Geographic Society, and supplementary grants from the Arizona Center for Nature Conservation, and Quarters For Conservation all to S.G. Further funding was provided by Discovery and Equipment grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Hal Whitehead of Dalhousie University, and a FNU large frame grant and a Villum Foundation Grant to Peter Madsen of Aarhus University; and small grants from the Dansk Akustisks Selskab, Oticon Foundation, and the Dansk Tennis Fond to Pernille Tønnesen of Aarhus University. Through 2014-2019, we were grateful for collaborative access to DTAGs and use of custom analytical tag code from Mark Johnson, Peter Tyack, and Peter Madsen.

A.L. would like to thank Prof. Peng Ding (UC Berkeley) for suggesting the incremental causal effect estimator.

References

J. Andreas, G. Beguš, M. M. Bronstein, R. Diamant, D. Delaney, S. Gero, S. Goldwasser, D. F. Gruber, S. de Haas, P. Malkin, N. Pavlov, R. Payne, G. Petri, D. Rus, P. Sharma, D. Tchernov, P. Tønnesen, A. Torralba, D. Vogt, and R. J. Wood. Toward understanding the communication in sperm whales. *iScience*, 25(6):104393, 2022. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2022.104393. URL https://www.sciencedirect.com/science/article/pii/S2589004222006642.

S. Athey and G. Imbens. Machine learning for estimating heterogeneous causal effects. Research papers, Stanford University, Graduate School of Business, 2015. URL https://EconPapers.repec.org/RePEc:ecl:stabus:3350.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1510489113.

G. Beguš. Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. *Frontiers in Artificial Intelligence*, 3:44, 2020. ISSN 2624-8212. doi: 10.3389/frai.2020.00044. URL https://www.frontiersin.org/article/10.3389/frai.2020.00044.

G. Beguš. Identity-based patterns in deep convolutional networks: Generative adversarial phonology and reduplication. *Transactions of the Association for Computational Linguistics*, 9:1180–1196, 2021a. doi: 10.1162/tacl_a_00421. URL https://aclanthology.org/2021.tacl-1.70.

G. Beguš. CiwGAN and fiwGAN: Encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. *Neural Networks*, 139:305–325, 2021b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.03.017. URL https://www.sciencedirect.com/science/article/pii/S0893608021001052.

G. Beguš. Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. *Computer Speech & Language*, page 101244, 2021c. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2021.101244. URL https://www.sciencedirect.com/science/article/pii/S0885230821000516.

G. Beguš and A. Zhou. Modeling speech recognition and synthesis simultaneously: Encoding and decoding lexical and sublexical semantic information into speech with no direct access to speech data. In *Proc. Interspeech 2022*, pages 5298–5302, 2022. doi: 10.21437/Interspeech.2022-11219.

P. Bermant, M. Bronstein, R. Wood, S. Gero, and D. Gruber. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific Reports*, 9(1), Dec. 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-48909-4. Publisher Copyright: © 2019, The Author(s).

J. Bose, R. P. Monti, and A. Grover. CAGE: Probing causal relationships in deep generative models, 2022a. URL https://openreview.net/forum?id=VCD05OEn7r.

J. Bose, R. P. Monti, and A. Grover. Controllable generative modeling via causal reasoning. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=Z44YAcLaGw.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6399-infogan-interpretable-representation-learning-by-information-maximizing-generative-adversarial-nets.pdf.

H. Chockler, D. Kroening, and Y. Sun. Explanations for occluded images, 2021. URL https://arxiv.org/abs/2103.03622.

A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis, and P. Kohli. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74, 2021. doi: 10.1038/s41586-021-04086-x. URL https://doi.org/10.1038/s41586-021-04086-x.

C. Donahue, J. J. McAuley, and M. S. Puckette. Adversarial audio synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ByMVTsR5KQ.

S. Gero, M. Milligan, C. Rinaldi, P. Francis, J. Gordon, C. Carlson, A. Steffen, P. Tyack, P. Evans, , and H. Whitehead. Behavior and social structure of the sperm whales of dominica, west indies. *Marine Mammal Science*, 30:905–922,, 2003.

S. Gero, A. Bøttcher, H. Whitehead, , and P. T. Madsen. Socially segregated, sympatric sperm whale clans in the atlantic ocean. *Royal Society Open Science*, 3:160061,, 2016a.

S. Gero, H. Whitehead, and L. Rendell. Individual, unit and vocal clan level identity cues in sperm whale codas. *Royal Society Open Science*, 3(1):150372, 2016b. doi: 10.1098/rsos.150372.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

T. A. Hersh, S. Gero, L. Rendell, M. Cantor, L. Weilgart, M. Amano, S. M. Dawson, E. Slooten, C. M. Johnson, I. Kerr, R. Payne, A. Rogan, R. Antunes, O. Andrews, E. L. Ferguson, C. A. Hom-Weaver, T. F. Norris, Y. M. Barkley, K. P. Merkens, E. M. Oleson, T. Doniol-Valcroze, J. F. Pilkington, J. Gordon, M. Fernandes, M. Guerra, L. Hickmott, and H. Whitehead. Evidence from sperm whale clans of symbolic marking in non-human cultures. *Proceedings of the National Academy of Sciences*, 119:e2201692119,, 2022.

J. Hill and E. A. Stuart. Causal inference: Overview. In J. D. Wright, editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 255–260. Elsevier, Oxford, second edition edition, 2015. ISBN 978-0-08-097087-5. doi: https://doi.org/10.1016/B978-0-08-097086-8.42095-7. URL https://www.sciencedirect.com/science/article/pii/B9780080970868420957.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289064.

G. W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000. ISSN 00063444. URL http://www.jstor.org/stable/2673642.

M. P. Johnson and P. L. Tyack. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE Journal of Oceanic Engineering*, 28:3–12,, 2003.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1804597116.

C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles, 2018. URL https://arxiv.org/abs/1802.03888.

P. T. Madsen, R. Payne, N. Kristiansen, I. Kerr, and B. Møhl. Sperm whale sound production studied with ultrasound-time-depth-recording tags. *Journal of Experimental Biology*, 205:1899–1906,, 2002.

P. T. Madsen, U. Siebert, and C. P. H. Elemans. Toothed whales use distinct vocal registers for echolocation and communication. *Science*, 379(6635):928–933, 2023. doi: 10.1126/science.adc9570. URL https://www.science.org/doi/abs/10.1126/science.adc9570.

S. Pagliarini, N. Trouvain, A. Leblois, and X. Hinaut. What does the Canary Say? Low-Dimensional GAN Applied to Birdsong. working paper or preprint, Nov. 2021. URL https://hal.inria.fr/hal-03244723.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv*, 2015.

L. Rendell and H. Whitehead. Vocal clans in sperm whales (physeter macrocephalus). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270:225–231,, 2003.

L. E. Rendell, S. L. Mesnick, M. L. Dalebout, J. Burtenshaw, and H. Whitehead. Can genetic differences explain vocal dialect variation in sperm whales, physeter macrocephalus? *Behavior Genetics*, 42:332–343,, 2012.

D. Rothenhäusler and B. Yu. Incremental causal effects, 2019. URL https://arxiv.org/abs/1907.13258.

T. Schulz, H. Whitehead, S. Gero, and L. Rendell. Overlapping and matching of codas in vocal interactions between sperm whales: Insights into communication function. *Animal Behaviour*, 76:1977–1988, 2008.

J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay, and J. J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.01.021. URL https://www.sciencedirect.com/science/article/pii/S0092867420301021.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.

W. Watkins and W. Schevill. Sperm whale codas. *Animal Behaviour*, 62:1485–1490,, 1977.

S. L. Watwood, P. J. O. Miller, M. Johnson, P. T. Madsen, and P. L. Tyack. Deep-diving foraging behaviour of sperm whales (physeter macrocephalus). *Journal of Animal Ecology*, 75(3):814–825, 2006. doi: https://doi.org/10.1111/j.1365-2656.2006.01101.x. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2656.2006.01101.x.

H. Whitehead. *Sperm whales : social evolution in the ocean*. University of Chicago Press, Chicago, 2003. ISBN 0226895173.

H. Whitehead and L. Rendell. *The cultural lives of whales and dolphins*. University of Chicago Press, Chicago, 2014.

H. Whitehead and L. Weilgart. Patterns of visually observable behavior and vocalizations in groups of female sperm whales. *Behaviour*, 118:332–343,, 1991.