# INFERRING LOW-DIMENSIONAL LATENT DESCRIPTIONS OF ANIMAL VOCALIZATIONS

**Jack Goffinet**[*†] jack.goffinet@duke.edu

**Richard Mooney**[†] mooney@neuro.duke.edu

**John Pearson**[*†‡§] john.pearson@duke.edu

October 19, 2019

## ABSTRACT

Vocalization is an essential medium for social and sexual signaling in most birds and mammals. Consequently, the analysis of vocal behavior is of great interest to fields such as neuroscience and linguistics. A standard approach to analyzing vocalization involves segmenting the sound stream into discrete vocal elements, calculating a number of handpicked acoustic features, and then using the feature values for subsequent quantitative analysis. While this approach has proven powerful, it suffers from several crucial limitations: First, handpicked acoustic features may miss important dimensions of variability that are important for communicative function. Second, many analyses assume vocalizations fall into discrete vocal categories, often without rigorous justification. Third, a syllable-level analysis requires a consistent definition of syllable boundaries, which is often difficult to maintain in practice and limits the sorts of structure one can find in the data. To address these shortcomings, we apply a data-driven approach based on the variational autoencoder (VAE), an unsupervised learning method, to the task of characterizing vocalizations in two model species: the laboratory mouse (*Mus musculus*) and the zebra finch (*Taeniopygia guttata*). We find that the VAE converges on a parsimonious representation of vocal behavior that outperforms handpicked acoustic features on a variety of common analysis tasks, including representing acoustic similarity and recovering a known effect of social context on birdsong. Additionally, we use our learned acoustic features to argue against the widespread view that mouse ultrasonic vocalizations form discrete syllable categories. Lastly, we present a novel "shotgun VAE" that can quantify moment-by-moment variability in vocalizations. In all, we show that data-derived acoustic features confirm and extend existing approaches while offering distinct advantages in several critical applications.

## 1 Introduction

Vocalization is an essential medium for social and sexual signaling in most birds and mammals, and also serves as a natural substrate for language and music in humans. Consequently, the analysis of vocal behavior is of great interest to ethologists, psychologists, linguists, and neuroscientists. A major goal of these various lines of enquiry is to develop methods for quantitative analysis of vocal behavior, efforts that have resulted in several powerful methods that enable the automatic or semi-automatic analysis of vocalizations. Key to this approach has been the existence of software packages that calculate acoustic features for each syllable within a vocalization [4, 39, 40, 7, 6]. For example, Sound Analysis Pro, focused on birdsong, calculates 14 features for each syllable, including duration, spectral entropy, and goodness of pitch, and uses these as a basis for subsequent clustering and analysis [39]. More recently, MUPET and DeepSqueak have applied a similar approach to mouse vocalizations, with a heavy focus on syllable clustering [40, 7]. Collectively, these and similar software packages have helped facilitate numerous discoveries, including the overnight consolidation of learned birdsong [9], cultural evolution among isolate zebra finches [11], and differences in ultrasonic vocalizations (USVs) between mouse strains [40].

[*]Center for Cognitive Neuroscience, Duke University
[†]Department of Neurobiology, Duke University
[‡]Department of Biostatistics & Bioinformatics, Duke University
[§]Department of Electrical and Computer Engineering, Duke University

Despite these insights, this general approach suffers from several limitations. First, handpicked acoustic features are often highly correlated, and these correlations can result in redundant characterizations of vocalization. Second, an experimenter-driven approach may exclude features that are relevant for communicative function or, conversely, may emphasize features that are not salient or capture negligible variation in the data. Third, there is no diagnostic approach to determine when enough acoustic features have been collected: Could there be important variation in the vocalizations that the chosen features simply fail to capture? Lastly and most generally, committing to a syllable-level analysis necessitates a consistent definition of syllable boundaries, which is often difficult in practice. It limits the sorts of structure one can find in the data, and is often difficult to relate to time series such as neural data, for which the relevant timescales are believed to be orders of magnitude faster than syllable rate.

Here, we address these shortcomings by applying a data-driven approach based on variational autoencoders (VAEs) [24, 32] to the task of quantifying vocal behavior in two model species: the laboratory mouse (*Mus musculus*) and the zebra finch (*Taeniopygia guttata*). The VAE is an unsupervised modeling approach that learns from data a pair of probabilistic maps, an "encoder" and a "decoder," capable of compressing the data into a small number of latent variables while attempting to preserve as much information as possible. In doing so, it discovers features that best capture variability in the data, offering a nonlinear generalization of methods like PCA and ICA that adapts well to high-dimensional data like natural images [16]. By applying this technique to collections of single syllables, encoded as time-frequency spectrograms, we looked for latent spaces underlying vocal repertoires across individuals, strains, and species, asking whether these data-dependent features might reveal aspects of vocal behavior overlooked by traditional acoustic metrics.

Our contributions are threefold: First, we show that the VAE's learned acoustic features outperform common sets of handpicked features in a variety of tasks, including capturing acoustic similarity, representing a well-studied effect of social context on zebra finch song, and comparing the USVs of different mouse strains. Second, using our learned latent features, we report new results concerning both mice and zebra finches, including the finding that mouse USV syllables do not appear to cluster into distinct subtypes, as is commonly assumed, but rather form a broad continuum. Third, we present a novel approach to characterizing stereotyped vocal behavior that does not rely on syllable boundaries, one which we find is capable of quantifying subtle changes in behavioral variability on tens-of-milliseconds timescales. In all, we show that data-derived acoustic features confirm and extend findings gained by existing approaches to vocal analysis, and offer distinct advantages over handpicked acoustic features in several critical applications.

## 2 Results

### 2.1 Autoencoders learn a low-dimensional space of vocal features

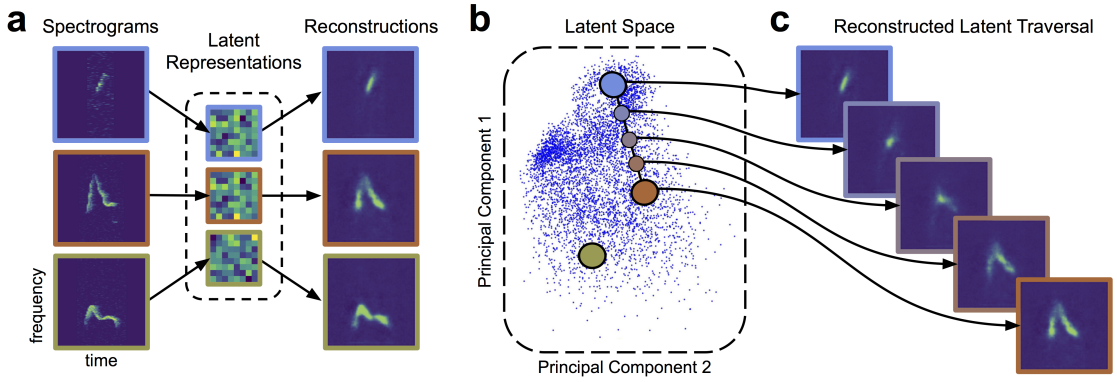


Figure 1: Autoencoders learn a latent vocal manifold. a) The VAE takes spectrograms as input (left column), maps them via a probabilistic "encoder" to a vector of latent dimensions (middle column), and reconstructs a spectrogram via a "decoder" (right column). The VAE attempts to ensure that these probabilistic maps match the original and reconstructed spectrograms as closely as possible. b) The resulting latent vectors can then be visualized via dimensionality reduction techniques like principal components analysis. c) Interpolations in latent space correspond to smooth syllable changes in spectrogram space. A series of points (dots) along a straight line in the inferred latent space is mapped, via the decoder, to a series of smoothly changing spectrograms (right). This correspondence between inferred features and realistic dimensions of variation is often observed when VAEs are applied to data like natural images [24, 32]

We trained a variational autoencoder (VAE) [24, 32] to learn a probabilistic mapping between vocalizations and a latent feature space. Specifically, we mapped single-syllable spectrogram images ($D = 16,384$ pixels) to vectors of latent features ($D = 32$) and back to the spectrogram space (Figure 1a). As with most VAE methods, we parameterized both our encoder and decoder using convolutional neural networks, with the two maps jointly trained to maximize a lower bound on the probability of the data given the model (see Methods). As visualized in Figure 1b, the result is a continuous latent space that captures the complex geometry of vocalizations. Each point in this latent space represents a single spectrogram image, and trajectories in this latent space represent sequences of spectrograms that smoothly interpolate between start and end syllables (Figure 1c). Although we cannot visualize the full 32-dimensional latent space, methods like PCA and the UMAP algorithm [27] allow us to communicate results in an informative and unsupervised way. The VAE training procedure can thus be seen as a compression algorithm that represents each spectrogram as a collection of 32 numbers describing data-derived vocal features. In what follows, we will show that these features outperform traditional handpicked features on a wide variety of analysis tasks.

### 2.2 Learned features capture and expand upon typical acoustic features

Most previous approaches to analyzing vocalizations have focused on tabulating a predetermined set of features such as syllable duration or entropy variance that are used for subsequent processing and analysis [40, 7, 39, 4]. We thus asked whether our learned feature space simply recapitulated these known features or also captured new types of information missed by traditional acoustic metrics. To address the first question, we mapped a publicly available collection of mouse USV syllables [1] into our learned latent space (31,440 total syllables) and colored the results according to three features — frequency bandwidth, maximum frequency, and duration — calculated by the analysis program MUPET [40]. As Figure 2a-c show, each acoustic feature appears to be encoded in a smooth gradient across our learned latent space, indicating that information about each has been preserved. In fact, when we quantified this pattern by asking how much variance in a wide variety of commonly used acoustic metrics could be accounted for by latent features (see Methods), we found that values ranged from 64% to 95%, indicating that most or nearly all traditional features were captured by our latent space (see Figure S1 for individual acoustic features). Furthermore, we found that, when the analysis was reversed, commonly used acoustic features were not able to explain as much variance in the VAE latent features, indicating a prediction asymmetry between the two sets (Figure 2e). That is, our learned features carry most of the information available in traditional features, as well as unique information missed by those metrics.

A potential explanation for this phenomenon lies in the latent feature disentangling properties of the VAE. Because the VAE is a Bayesian method that makes use of a prior on latent features, it benefits from an "automatic Occam's Razor" [26], pruning away unused latent dimensions and generally encouraging discovered latents to be uncorrelated [3]. By contrast, while the three software packages we tested (SAP [39], MUPET [40], DeepSqueak [7]) measure upwards of 14 acoustic features per syllable, we find that these features often exhibit high correlations (Figure 2d,S2), effectively reducing the number of independent measurements available.

We thus attempted to quantify the effective representational capacity of the VAE and current best approaches in terms of the dimensionalities of their respective feature spaces. We begin by noting that the VAE, although trained with a latent space of 32 dimensions, converges on a parsimonious representation that makes use of only 5 to 7 dimensions, with variance apportioned roughly equally between these (Figure S3). For the handpicked features, we normalized each feature independently by z-score to account for scale differences. For comparison purposes, we applied the same normalization step to our learned features and calculated the cumulative feature variance as a function of number of principal components (Figure 2f). In such a plot, shallow linear curves are preferred, since this indicates that variance is apportioned roughly equally among principal components and the effective dimensionality of the space is large. Equivalently, this means that the eigenvalue spectrum of the feature correlation matrix is close to the identity. As Figure 2f thus makes clear, the spaces spanned by our learned latent features have comparatively higher effective dimension than the spaces spanned by traditional features, suggesting that the learned features have a higher representational capacity.

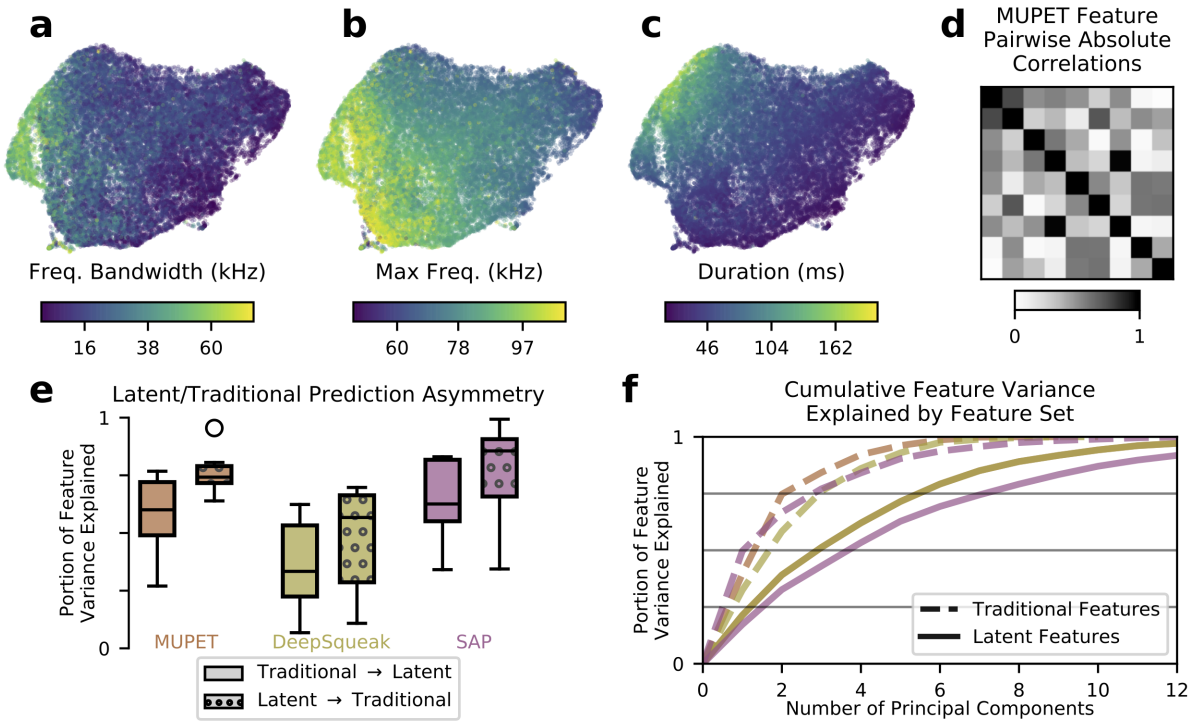


Figure 2: Learned acoustic features capture and expand upon traditional features. **a-c**) A UMAP projection of latent descriptions of mouse USVs, colored by various traditional acoustic features. The smoothly varying colors reflect that these traditional acoustic features are represented by the latent features. d) Many traditional features are highly correlated. When applied to the mouse USVs from **a-c**, many of the acoustic features compiled by the analysis program MUPET have high correlations, although an ideal representation would exhibit minimal off-diagonal correlations. e) To better understand the representational capacity of traditional and latent acoustic features, we used each set of features to predict the other and vice versa (see Methods). We find that, across software programs, our learned latent features were better able to predict the values of traditional features than vice-versa, suggesting they have a higher representational capacity. f) As a another test of representational capacity, we performed PCA on the feature vectors to determine the effective dimensionality of the space spanned by each set of features (see Methods). We find in all cases that latent features require more principal components to account for the same portion of feature variance, evidence that latent features span a higher dimensional space than traditional features applied to same datasets.

The degree to which our learned features capture novel information can also be demonstrated by considering their ability to encode a notion of spectrogram similarity, since this is a typical use to which they are put in clustering algorithms (although see [40] for an alternative approach to clustering). We tested this by selecting query spectrograms and asking for the closest spectrograms as represented in both the DeepSqueak acoustic feature space and learned latent space. As Figure 3 shows, DeepSqueak feature space often fails to return similar spectrograms, whereas the learned latent space reliably produces close matches (see Figure S4 for a representative sample). This suggest that the learned features better characterize local variation in the data by more accurately arranging nearest neighbors.

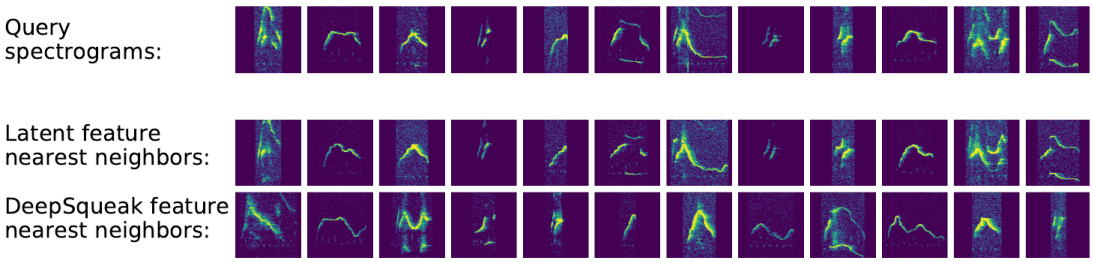


Figure 3: Latent features better represent acoustic similarity. **top row**: example spectrograms **middle row**: nearest neighbors in latent space **bottom row**: nearest neighbors in DeepSqueak feature space.

### 2.3 Latent spaces facilitate comparisons between vocal repertoires

Many experimental designs require quantifying differences between sets of vocalizations. As a result, the ability of a feature set to distinguish between syllables, individuals, and groups poses a key test of our approach. Here, we apply our VAE features to several comparison problems for which handpicked features are often used.

For example, a common comparison in birdsong research is that between female-directed and undirected song. It is well-established in the literature that directed song is more stereotyped and slightly faster than undirected song [37]. We thus asked whether our learned features can detect this effect. In Figure 4a we plot the first two principal components of named acoustic features calculated by the Sound Analysis Pro software package [39] for both directed and undirected renditions of a single zebra finch song syllable. We note a generally diffuse arrangement and a subtle leftward bias in the directed syllables compared to the undirected syllables. Figure 4b displays the same syllables with respect to the first two principal components of our learned latent features, showing a much more concentrated distribution of directed syllables relative to undirected syllables. In fact, when we quantify this reduction of variability across all feature-space dimensions and song syllables (see Methods), learned latent features consistently report greater variability reductions than SAP-generated features (Figure 4c; SAP: 0-20%, VAE: 27-37%) indicating latent features are more sensitive to this effect.

Similarly, we can ask whether latent features are able to capture differences between groups of individuals. In [40], the authors compared USVs of 12 strains of mice using a clustering-based approach. Here, we perform an alternative version of this analysis using two publicly available mouse strains (C57/BL6 and DBA/2) that were included in this earlier study. Figure 4d shows a UMAP projection of the 31,440 detected syllables, colored by mouse strain. Visualized with UMAP, clear differences between the USV distributions are apparent. In contrast to traditional acoustic features such as 'mean frequency', individual latent features (vector components) are generally less interpretable. Despite this, we note that, when taken together with an "atlas" of USV shapes derived from this visualization (Figure S6), we can develop an intuitive understanding of the differences between the USVs of the two strains: the C57 mice mostly produce noisy USVs, while the DBA mice produce a much greater variety, including many short low-frequency syllables that C57s rarely produce.

Given these results, we asked whether these strain differences are evident at the level of individual 6.5-minute recording sessions. To compare distributions of syllables without making restrictive parametric assumptions, we employed Maximum Mean Discrepancy (MMD), a difference measure between pairs of distributions [13]. We estimated MMD between the distributions of latent syllable encodings for each pair of recording sessions (see Methods) and visualized the result as a distance matrix (Figure 4e). Here, lighter values indicate more similar syllable repertoires. We note that, in general, values are brighter when comparing repertoires within strains than when comparing across strains, consistent with the hypothesis of inter-strain differences. We also note some substructure, including a well-defined cluster within the C57 block (annotated).

Finally, we used a much larger library of female-directed mouse USVs (36 individuals, 2-4 20-minute recording sessions each, 40 total hours of audio) to investigate the diversity and stability of syllable repertoires. We repeated the above procedure, estimating MMD for each pair of recording sessions (Figure S7), and then computed a t-SNE layout of the recording sessions (see Methods). In Figure 4f, each recording session is represented by a scatterpoint, and recordings of the same individual are connected and displayed in the same color. We note an overall organization of syllables into two clusters, corresponding to the genetic backgrounds of the mice. Furthermore, we note that almost all recordings of the same individuals are co-localized, indicating that within-subject differences in syllable repertoire are smaller than those between individuals. Although it has been previously shown that a deep convolutional neural network can be trained

to classify USV syllables according to mouse identity with good accuracy [18], here we see that repertoire features learned in a wholly unsupervised fashion very nearly do the same, evidence that mice emit individually-stereotyped, stable vocal repertoires.

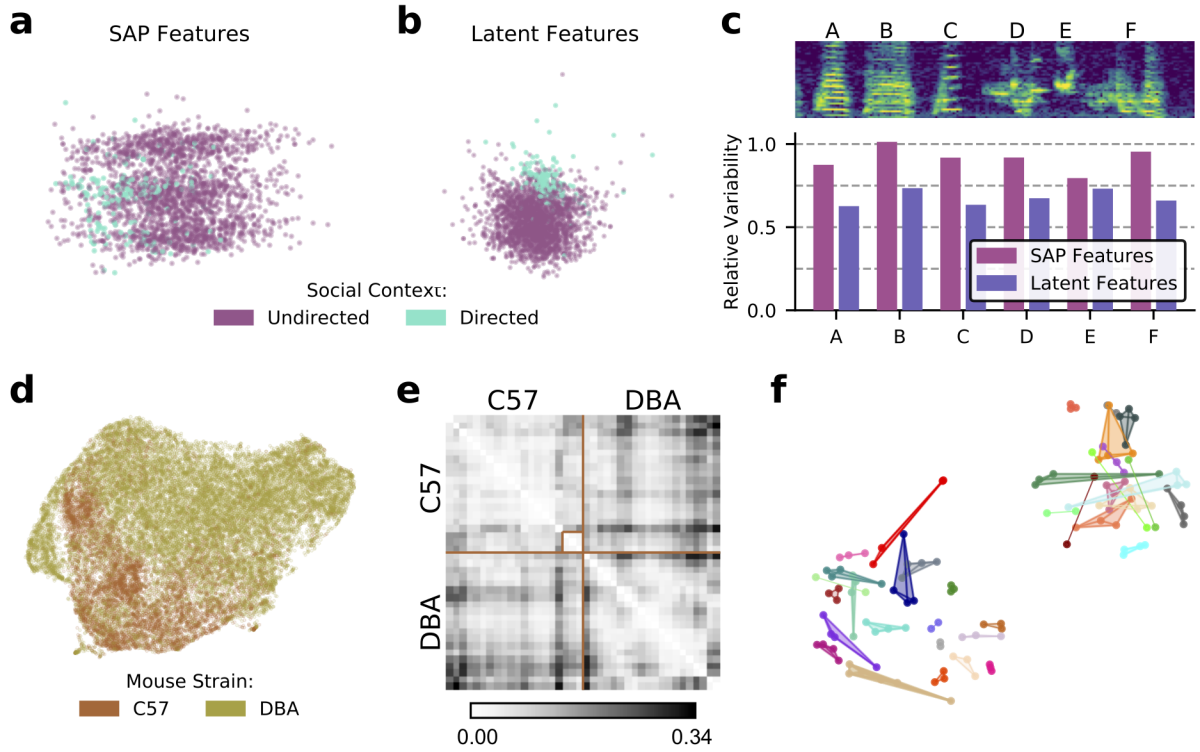


Figure 4: Latent features better capture differences in sets of vocalizations. a) The first two principal components in SAP feature space of a single zebra finch song syllable, showing differences in directed and undirected syllable distributions. b) The first two principal components of latent means, showing the same comparison. Learned latent features more clearly indicate differences between the two conditions by clustering directed syllables together. c) Acoustic variability of each song syllable as measured by SAP features and latent features (see Methods). Latent features more clearly represent the constriction of variability in the directed context. d) A UMAP projection of the latent means of USV syllables from two strains of mice, showing clear differences in their vocal repertoires. e) A matrix showing Maximum Mean Discrepancy (MMD) between syllable repertoires for each pair of the 40 recording sessions from **d** (see Methods). Lighter values correspond to more similar syllable repertoires. f) Visualization of USV repertoire variation across strains, individuals, and days. The dataset, which is distinct from that represented in **d** and **e**, contains 36 individuals, 118 recording sessions, and 156,180 total syllables. Color indicates individual mice, and scatterpoints of the same color represent repertoires recorded on different days. Distances between points represent the similarity in vocal repertoires (see Methods), with closer points more similar. We note that the major source of repertoire variability corresponds to genetic background, corresponding to the two distinct clusters. A smaller level of variability can be seen across individuals in the same clusters. Finally, we see that individual mice have repertoires with even less variability, indicated by the close proximity of most repertoires from a single mouse. The degree to which these connected points are spatially localized reflects the individuality of vocal repertoires.

### 2.4 Latent features fail to support cluster substructure in USVs

Above, we have shown that, by mapping complex sets of vocalizations to low-dimensional latent representations, autoencoders allow us to visualize the relationships among elements in mouse vocal repertoires. The same is likewise true for songbirds such as the zebra finch, *T. guttata*. Figure 5 compares the geometry of learned latent spaces for an individual of each species as visualized via UMAP. As expected, the finch latent space exhibits well-delineated clusters corresponding to song syllables (Figure 5a). However, as seen above, mouse USVs clump together in a single quasi-continuous mass (Figure 5b). This raises a puzzle, since the clustering of mouse vocalizations is often considered well-established in the literature [17, 4, 43, 6, 15] and is assumed in most other analyses of these data [40, 7]. Clusters

of mouse USVs are used to assess differences across strains [40], social contexts [6, 7, 14], and genotypes [12], and the study of transition models among clusters of syllables has given rise to models of syllable sequences that do not readily extend to the non-clustered case [17, 6].

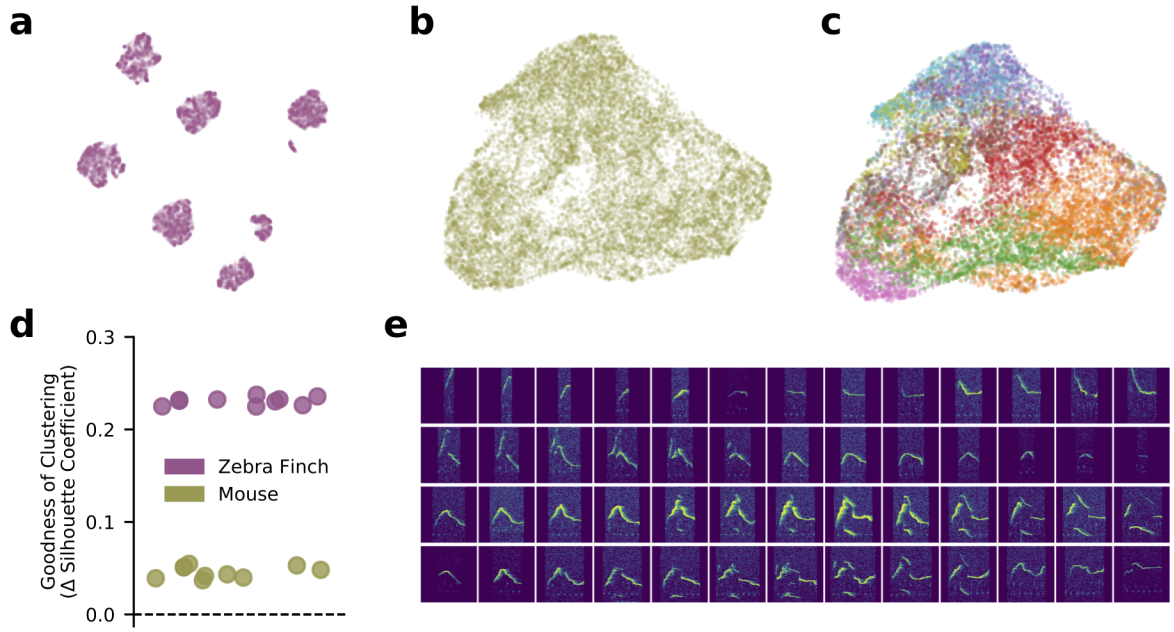


Figure 5: Bird syllables clearly cluster, but mouse USVs do not. a) UMAP projection of the song syllables of a single male zebra finch (14,270 syllables) b) UMAP projection of the USV syllables of a single male mouse (17,400 syllables) c) the same UMAP projection as in **b**, colored by MUPET-assigned labels d) Mean silhouette coefficient, an unsupervised clustering metric, applied to latent descriptions of zebra finch song syllables and mouse syllables. The dotted line indicates the null hypothesis of a single covariance-matched Gaussian noise cluster fit by the same algorithm. Each scatterpoint indicates a cross-validation fold, and scores are plotted as differences from the null model. Higher scores indicate more clustering. e) Interpolations (horizontal series) between distinct USV shapes (left and right edges) demonstrating the lack of data gaps between putative USV clusters.

We therefore asked whether mouse USVs do, in fact, cluster or whether, as our latent space projection suggests, they form a single continuum. In principle, this is impossible to answer, because, without the benefit of ground truth labels, clustering is an unsupervised learning method. Moreover, there is little consensus among researchers as to the best method for assessing clustering and where the cutoff between clustered and non-clustered data lies [19]. In practice, new clustering algorithms are held to function well when they outperform previous approaches and produce sensible results on data widely agreed on to be clustered. Thus, while it is clear that birdsong should be and is clustered by the VAE (Figure 5a), we can only ask whether clustering is a more or less satisfying account of the mouse data in Figure 5b.

To address this question, we performed a series of analyses to examine the clustering hypothesis from complementary angles. First, we asked how clusters detected by other analysis approaches correspond to regions in our latent space. As shown in Figure 5c, clusters detected by MUPET roughly correspond to regions of the UMAP projection, with some overlap between clusters (e.g., purple and blue clusters) and some non-contiguity of single clusters (red and orange clusters). That is, even though clusters do broadly label different subsets of syllables, they also appear to substantially bleed into one another, unlike the finch song syllables in Figure 5a. However, it might be objected that Figure 5b displays the UMAP projection, which only attempts to preserve local relationships between nearest neighbors and is not to be read as an accurate representation of the latent geometry. Might the lack of apparent clusters result from distortions produced by the projection to two dimensions? To test this, we calculated several unsupervised clustering metrics on full, unprojected latent descriptions of zebra finch and mouse syllables (see Methods). Both bird syllables and mouse USVs were more clustered than moment-matched samples of Gaussian noise, a simple null hypothesis, but mouse USVs were closer to the null than to birdsong on multiple goodness-of-clustering metrics (Figure 5d,S8). Finally, we tested whether the data contained noticeable gaps between syllables in different clusters. If syllable clusters are well-defined, there should not be smooths sequences of data points connecting distinct examples. However, we find

that even the most acoustically disparate syllables can be connected with a sequence of syllables exhibiting more-or-less smooth acoustic variation, in contrast to zebra finch syllables (Figure S9). Thus, even though clustering may not constitute the best account of mouse USV syllable structure, learned latent features provide useful tools to both explore and quantify the acoustic variation within and across individuals.

### 2.5 Measuring acoustic variability over tens of milliseconds

Our results above have shown that data-derived latent features represent more information about syllables than traditional metrics and can successfully capture differences within and between individuals and groups. Here, we consider how a related approach can also shed light on the short-time substructure of vocal behavior.

The analysis of syllables and other discrete segments of time is limited in at least two ways: First, timing information, such as the lengths of gaps between syllables, is ignored. Second, experimenters must choose the unit of analysis (syllable, song motif, bout), which has a significant impact on the sorts of structure that can be identified [21]. In an attempt to avoid these limitations, we pursued a complementary approach, using the VAE to infer latent descriptions of fixed duration audio segments, irrespective of syllable boundaries. Similar to the shotgun approach to gene sequencing [41] (and a related method of neural connectivity inference [38]), we trained the VAE on randomly sampled segments of audio, requiring that it learn latent descriptions sufficient to characterize any given time window during the recording. That is, this "shotgun-VAE" approach encouraged the autoencoder to find latent features sufficient to "glue" continuous sequences back together from random audio snippets.

Figure 6a shows a UMAP projection of fixed-duration segments from a subset of the mouse USVs shown in Figure 5b. While this projection does reveal some structure (silence on the right, shorter to longer syllables in a gradient from right to left), there is no evidence of stereotyped sequential structure. In contrast, Figure 6b shows the same technique applied to bouts of zebra finch song, with the song represented as a single well-defined strand coursing clockwise from the bottom to the top left of the projection. Other notable features are the loop on the left containing repeated, highly variable introductory notes that precede and often join song renditions and a 'linking note' that sometimes joins song motifs. Most importantly, such a view of the data clearly illustrates not only stereotypy but variability: introductory notes are highly variable, but so are particular syllables (B, E) in contrast to others (C, F).

Following this, we asked whether our shotgun VAE method could be used to assess the phenomenon of reduced variability in directed birdsong [37]. We examined the song portion of Figure 6b (Fig. 4c) in both directed and undirected conditions, warping each in time to account for well-documented differences in rhythm and tempo. We then trained a VAE on the warped spectrograms. As a plot of the first principle component of the latent embedding shows (Fig. 4d), the VAE is able to recover the expected reduction in directed song variability on a tens-of-milliseconds timescale relevant to the hypothesized neural underpinnings of the effect [10]. This result recapitulates similar analyses that have focused on harmonic and tonal syllables like A and B in Figure 4c [20], but our method is applicable to all syllables, yielding a continuous estimate of song variability (Fig. 4e). Thus, not only do VAE-derived latent variables capture structural properties of syllable repertoires, our shotgun VAE approach serves to characterize vocal dynamics as well.

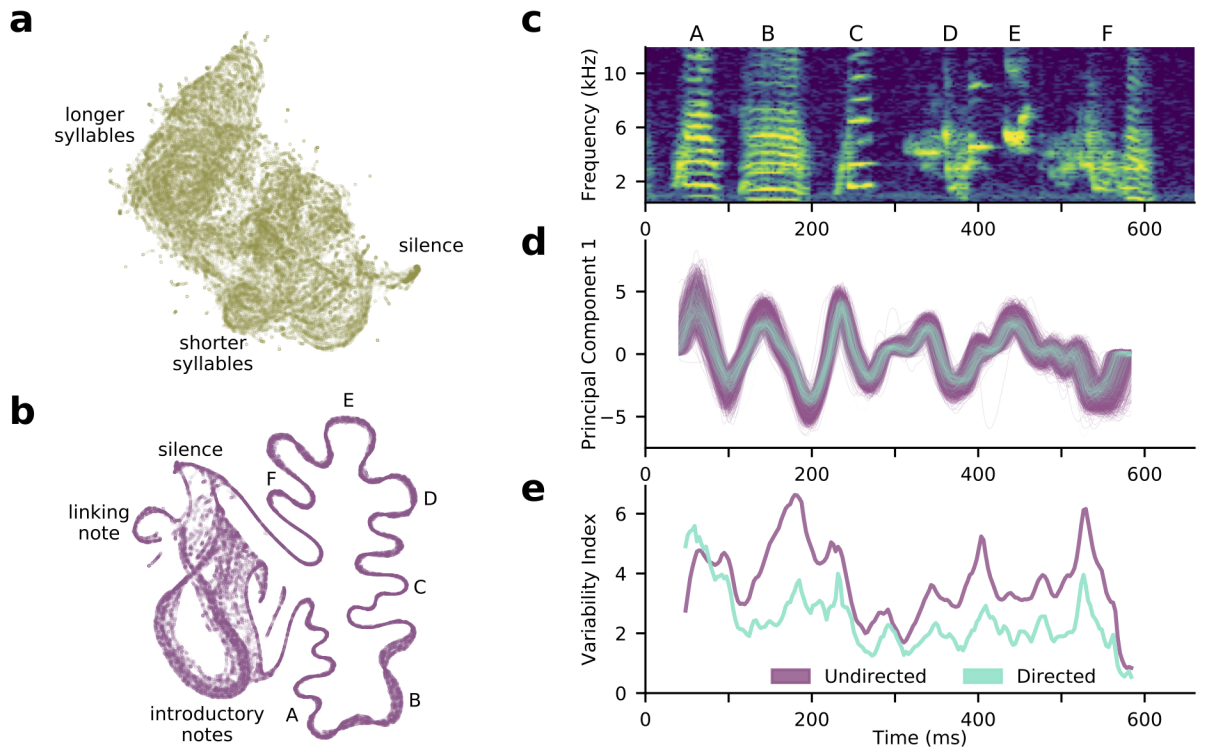


Figure 6: Our shotgun VAE approach learns to compress arbitrary fixed-duration spectrograms of vocal behavior and can be used to represent short-timescale variability and global trends in behavior. a) A latent mean UMAP projection of 100,000 200ms windows of USVs (cp. Figure 4a). b) A latent mean UMAP projection of 100,000 120ms windows of zebra finch song (compare to Figure 4b) Stereotyped song progresses counterclockwise on the right side, while more more variable, repeated introductory notes form a loop on the right side. c) A rendition of the song in b). d) The song's first principal component in latent space, showing both directed (blue) and undirected (purple) renditions. e) In contrast to a syllable-level analysis, we can measure zebra finch song variability using the shotgun VAE in continuous time. Song variability in both directed (purple) and undirected (cyan) contexts are plotted (see Methods).

## Discussion

The complexity and high dimensionality of vocal behavior have posed a persistent challenge to the scientific study of animal vocalization. In particular, comparisons of vocalizations across across time, individuals, groups, and experimental conditions require some means of characterizing the similarity of selected groups of vocal behaviors. Feature vector-based approaches and widespread software tools have gone a long way toward addressing this challenge and providing meaningful scientific insights, but the reliance of these methods on handpicked features leaves open the question of whether other feature sets might better characterize vocal variability.

Here, by adopting a a data-driven approach, we have shown that features learned by the variational autoencoder (VAE), an unsupervised learning method, outperform frequently used acoustic features across a variety of common analysis tasks. As we have shown, these learned features are both more parsimonious (Figure S2), capture more variability in the data (Figure 2e,f), and better characterize vocalizations as judged by nearest neighbor similarity (Figure 3). Moreover, these features easily facilitate comparisons across sessions (Figure 4f), social contexts (Figure 4a-c), and individuals (Figure 4d-f), quantifying not only differences in mean vocal behavior (Figure 4d), but also in vocal variability (Figure 4c).

Moreover, we have argued above that, despite conventional wisdom, clustering is not the best account of the diversity of mouse vocal behavior. We argued this on the basis of multiple converging lines of evidence, but note three important qualifications: First, the huge variety of vocal behavior among rodents [2, 17, 29, 34, 35, 28] suggests the possibility of clustered vocal behavior in some mouse strains not included in our data. Second, it is possible that the difference in clustered and non-clustered data depends crucially on data set size. If real syllables even occasionally fall between well-defined clusters, a large enough data set might lightly "fill in" true gaps. Conversely, even highly clustered data

may look more or less continuous given an insufficient number of samples per cluster. While this does not appear to be the case in Figure 5, it is difficult to rule this out in general. Finally, our purely signal-level analysis of vocal behavior cannot address the possibility that a continuous distribution of syllables could nevertheless be perceived categorically. For example, swamp sparrows exhibit categorical neural and behavioral responses to changes in syllable duration [31]. Nonetheless, we argue that, without empirical evidence to this effect in rodents, caution is in order when interpreting the apparent continuum of USV syllables in categorical terms.

Lastly, we showed how a novel "shotgun VAE" approach can be used to extend our approach to the quantification of moment-by-moment vocal variability. In previous studies, syllable variability has only been quantified for certain well-characterized syllables like harmonic stacks in zebra finch song [20]. Our method, by contrast, provides a continuous variability measure for all syllables in Figure 6c. This is particularly useful for studies of the neural basis of this vocal variability, which is hypothesized to operate on millisecond to tens-of-milliseconds timescales [10].

Nonetheless, as a data-driven method, our approach carries some drawbacks. Most notably, the VAE must be trained on a per-dataset basis. This is more computationally intensive than calculating typical acoustic features ($\approx$1 hour training times on a GPU) and also prevents direct comparisons across datasets unless they are trained together in a single model. Additionally, the resulting learned features, representing nonlinear, non-separable acoustic effects, are somewhat less interpretable than named acoustic features like duration and spectral entropy. However, several recent studies in the VAE literature have attempted to address this issue by focusing on the introduction of covariates [36, 25, 22] and "disentangling" approaches that attempt to learn independent sources of variation in the data [16, 3], which we consider to be promising future directions.

Finally, we note that while our focus in this work is vocal behavior, our training data are simply syllable spectrogram images. Similar VAE approaches could also be applied to other kinds of data summarizable as images or vectors. Our "shotgun VAE" approach could likewise be applied to sequences of such vectors, potentially revealing structures like those in Figure 6b. More broadly, our results suggest that data-driven dimensionality reduction methods, particularly modern nonlinear, overparameterized methods, and the latent spaces that come with them, offer a promising avenue for the study of many types of complex behavior.

# 3 Methods

### Animal Statement

All experiments were conducted according to protocols approved by the Duke University Institutional Animal Care and Use Committee.

### Recordings

Recordings of C57BL/6 and DBA/2 mice were obtained from the MUPET Wiki [1]. These recordings are used in Figures 2a-f, 3, 4d-e, and 5e.

Additional recordings of female-directed mouse USVs are analyzed in Figure 4f. These recordings comprise 36 male mice from various genetic backgrounds over 118 recording sessions of roughly 20 minutes each ($\approx$40 total hours, 156,180 total syllables). USVs were recorded with an ultrasonic microphone (Avisoft, CMPA/CM16), amplified (Presonus TubePreV2), and digitized at 300 kHz (Spike 7, CED). A subset of these recordings corresponding to a single individual (17,400 syllables) are further analyzed in Figure 5b-d, and 6a. Because these recordings contained more noise than the first set of C57/DBA recordings, we removed false positive syllables by training the VAE on all detected syllables, projecting latent syllables to two dimensions, and then removing syllables contained within the resulting cluster of noise syllables (see Figure S10).

A single male zebra finch was recorded over two-day period (153-154 days post-hatch) in both female-directed and undirected contexts (14,270 total syllables). Songs were recorded with Sound Analysis Pro 2011.104 [39] in a soundproof box.

### Software Comparisons

We compared our VAE method to three widely used vocal analysis packages: MUPET [40], DeepSqueak [7] (for mouse USVs) and SAP [39] (for birdsong). We used MUPET 2.0, DeepSqueak 2.0, and SAP 2011.104, each with default parameter settings. MUPET clusters were found using the combined C57/DBA data set [1], using the minimum number of clusters (10).

### Audio Segmentation

Individual USV syllable onsets and offsets were detected using MUPET with default parameter settings. DeepSqueak features were generated using the DeepSqueak "import from MUPET" feature (Figure [**?** ]E,G). Sliding window analysis was restricted to manually-defined regions (bouts) of vocalization. Zebra finch songs were segmented semi-automatically: first we selected four representative song renditions from each individual. Then we converted these to spectrograms using a Short Time Fourier Transform with Hann windows of length 512 and overlap of 256, averaged these spectrograms, and blurred the reusult using a gaussian filter with 0.5 pixel standard deviation. The result was a song template used to match against the remaining data. Specifically, we looked for local maxima in the normalized cross-correlation between the template and each audio file. Matches corresponded to local maxima with cross-correlations above 1.8 median absolute deviations from the median, calculated on a per-audio-file basis. A spectrogram was then computed for each match. All match spectrograms were then projected to two dimensions using UMAP [27], from which a single well-defined cluster, containing stereotyped song, was retained. Zebra finch syllable onsets and offsets were then detected using SAP on this collection of song renditions. We used the following hand-picked SAP parameters: frequency range=11,025Hz; FFT window: 9.27ms; advance window: 1ms; contour threshold: 10; pitch calculation: 'simple pitch average, mean frequency'; amplitude segmentation with syllables defined when amplitude $> 42.4$; minimum stop duration: 7ms; bout ends when stop $> 100$ms; minimum syllable duration: 15ms. After segmentation, syllable spectrograms were projected to two dimensions using UMAP, and eight well-defined clusters of incorrectly segmented syllables were removed, leaving six well-defined clusters of song syllables.

### Spectrograms

Spectrograms were computed using the log modulus of a signal's Short Time Fourier Transform, computed using Hann windows of length 512 and overlap of 256 for bird vocalization, and length 1024 and overlap 512 for mouse vocalization. Sample rates were 32kHz for bird vocalization and 250kHz for mouse vocalization, except for the recordings in Figure 3f, which were sampled at a rate of 300kHz. The resulting time/frequency representation was then interpolated at desired frequencies and times. Frequencies were mel-spaced from 0.4 to 10kHz for bird recordings and linearly spaced from 30 to 110kHz for mouse recordings. For both species, syllables longer than $t_{\text{max}} = 200$ms were discarded. Additionally, short syllables were stretched in time in a way that preserved relative duration, but encouraged the VAE to represent fine temporal details. Specifically, a syllable of duration $t$ was stretched by a factor of $\sqrt{\frac{t_{\text{max}}}{t}}$. The resulting spectrograms were then clipped to manually tuned minimum and maximum values: mice from Figure 2, Figures 3, 4d-e, and 5c-e (2.0,6.0); mice from Figures 4f, 5a, 6a (-6.5,-2.5); zebra finch (2.0,6.5). The values were then linearly stretched to lie in the interval [0,1]. The resulting spectrograms were 128 x 128 = 16384 pixels, with syllables shorter than $t_{\text{max}}$ zero-padded symmetrically.

### Model Training

Our variational autoencoder is implemented in PyTorch (v1.1.0) and trained to maximize the standard evidence lower bound (ELBO) objective using the reparameterization trick and ADAM optimization [24, 32, 30, 23]. The encoder and decoder are deep convolutional neural networks with fixed architecture diagrammed in Figure S11. The latent dimension was fixed to 32, which was found to be sufficient for all training runs. The approximate posterior was parameterized as a normal distribution with low rank plus diagonal covariance: $q(z) = \mathcal{N}(z; \mu, uu^\top + \text{diag}(d))$ where $\mu$ is the latent mean, $u$ is a 32x1 covariance factor, and $d$ was the latent diagonal, a vector of length 32. The observation distribution was parameterized as $\mathcal{N}(\mu, 0.1)$ where $\mu$ was the output of the decoder. All activation functions were Rectified Linear Units. Learning rate was set to $10^{-3}$ and batch size was set to 64.

### Comparison of VAE and handpicked features

For each analysis tool (MUPET, DeepSqueak, SAP), we assembled two feature sets: one calculated by the comparison tool (e.g., MUPET featurs) and one a matched VAE set. For the first set, each feature calculated by the program was z-scored and all components with non-zero variance were retained (9/9, 10/10, and 13/14 components for MUPET, DeepSqueak, and SAP, respectively). For the second set, we trained a VAE on all syllables, computed latent means of these via the VAE encoder, and removed principal components containing less than 1% of the total feature variance (7, 5, and 5 out of 32 components retained for MUPET, DeepSqueak, and SAP syllables, respectively). Each feature set was used as a basis for predicting the features in the other set using $k$-nearest neighbors regression with $k$ set to 10 and nearest neighbors determined using Euclidean distance in the assembled feature spaces. The variance-explained value reported is the average over 5 shuffled train/test folds (Figure 2e).

Unlike latent features, traditional features do not come equipped with a natural scaling. For this reason, we z-scored traditional features to avoid tethering our analyses to the identities of particular acoustic features involved. Then, to fairly compare the effective dimensionalities of traditional and acoustic features, we thus also z-scored the latent features as well, thereby disregarding the natural scaling of the latent features. PCA was then performed on the resulting scaled feature set (Figure 2f).

### Birdsong Variability Index

For Figure 4c, given a set $\{z_i | i = 1 \dots n\}$ of feature vectors of $n$ syllables, we defined a variability index for the data as follows:

$$\text{V.I.} = \min_{z_i} \rho(z_i) \tag{1}$$

where $\rho(z)$ is proportional to a robust estimator of the variance of the data around $z$:

$$\rho(z) = \underset{z_j}{\text{median}} \, ||z - z_j||_2^2 \tag{2}$$

We calculate above metric for every combination of syllable (A-F), feature set (SAP-generated vs. VAE-generated), and social context (directed vs. undirected) and report the variability index of the directed condition relative to the variability index of the undirected condition (Figure 4c).

For Figure 6e, we would ideally use the variability index defined above, but $\rho(z)$ is expensive to compute for each data point, as required in (1). Thus, we use an approximate center point defined by the median along each *coordinate*: $\hat{z}^i \equiv \text{median}(z^i)$, where the superscript here represents the $i$th coordinate of $z$. That is, $\hat{z}$ contains the medians of the marginal distributions. This value is calculated for each combination of timepoint and social context (directed vs. undirected) and plotted in Figure 6e.

### Maximum Mean Discrepancy

We used the Maximum Mean Discrepancy (MMD) integral probability metric to quantify differences in sets of syllables [13]. Given random variables $x$ and $y$, MMD is defined over a function class $\mathcal{F}$ as $\sup_{f \in \mathcal{F}} \mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)]$. Here, $\mathcal{F}$ was taken to the set of functions on the unit ball in a reproducing kernel Hilbert space with fixed spherical Gaussian kernel. The kernel width $\sigma$ was chosen to be the median distance between points in the aggregate sample $\{x_i\} \cup \{y_i\}$, a common heuristic [13]. In our application (Figure 3e), we obtained 20 approximately 6.5 minute recordings of male C57BL/6 mice and 20 approximately 6.5 minute recordings of male DBA/2 mice (see Recordings). Latent means of USVs from a single recording were treated as independent and identically distributed draws from a recording-specific USV distribution, and MMD was estimated using these latent means. In Figure 4e, these MMD values are represented as a matrix with darker values representing more distinct repertoires. The order of rows was obtained by agglomerative clustering. In Figure 4f, a t-distributed Stochastic Neighbor Embedding (t-SNE) was computed for each recording session, with the distance between recording sessions taken to be the estimated MMD between them.

### Unsupervised Clustering Metrics

We used three unsupervised clustering metrics to assess the quality of clustering for both zebra finch and mouse syllables: the mean silhouette coefficient [33], the Calinski-Harabasz Index [5], and the Davies-Bouldin Index [8]. For each species (zebra finch and mouse) we partitioned the data for tenfold cross-validation (train on 9/10, test on 1/10 held out). For a null comparison, for each 10% subset of the data, we created a synthetic Gaussian noise dataset matched for covariance and number of samples. These synthetic noise data sets were then used to produce the dotted line in Figure 5d.

For each data split, we clustered using a Gaussian Mixture Model (GMM) with full covariance using Expectation Maximization on the training set. We then evaluated each clustering metric on the test set. The number of clusters, $k$, was set to 6 in Figure 5d, but qualitatively similar results were obtained when $k$ was allowed to vary between 2 and 12 (Figure S8). Reported values in Figure 5d and Figure S8 are the differences in unsupervised metrics on real data and Gaussian noise for each cross-validation fold, with a possible sign change to indicate higher values as more clustered.

### Shotgun VAE

To perform the analysis in Figure 6a-b, regions of active vocalization were defined manually for both species (22 minutes of mouse recordings, 2 minutes of zebra finch recordings). Zebra finch bouts containing only calls and no song motifs were excluded. For both species, the duration of audio chunks was chosen to be roughly as long as the

longest syllables (zebra finch: 120ms; mouse: 200ms). No explicit training set was made. Rather, onsets and offsets were drawn uniformly at random from the set of fixed-duration segments and the corresponding spectrograms were computed on a per-datapoint basis. Thus, the VAE likely never encountered the same spectrogram twice, encouraging it to learn the underlying time series.

To perform the variability reduction analysis in Figure 6d-e, song renditions were collected (see Audio Segmenting) and a spectrogram was computed for each. The whole collection of spectrograms was then jointly warped using piecewise-linear time warping [42]. Fixed-duration training spectrograms were made by interpolating normal spectrograms (as described in Spectrograms) at linearly spaced time points in warped time, generally corresponding to non-linearly spaced time points in real time. As above, spectrograms were made during training on a per-datapoint basis. After training the VAE on these spectrograms, latent means were collected for 200 spectrograms for each song rendition, linearly spaced in warped time from the the beginning to the end of the song bout. Lastly, for each combination of condition (directed vs. undirected song) and timepoint, the variability index described above was calculated. A total of 186 directed and 2227 undirected song renditions were collected and analyzed.

### Software Availability

The latest version of Autoencoded Vocal Analysis, the Python package used to generate, plot, and analyze latent features, is freely available online: `https://github.com/jackgoffinet/autoencoded-vocal-analysis`.

## References

[1] Mupet wiki. `https://github.com/mvansegbroeck/mupet/wiki/MUPET-wiki`. Accessed: 2019-09-07.

[2] Julia C Berryman. Guinea-pig vocalizations: their structure, causation and function. *Zeitschrift für Tierpsychologie*, 41(1):80–106, 1976.

[3] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $\beta$-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[4] Zachary D Burkett, Nancy F Day, Olga Peñagarikano, Daniel H Geschwind, and Stephanie A White. Voice: A semi-automated pipeline for standardizing vocal analysis across models. *Scientific reports*, 5:10237, 2015.

[5] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[6] Jonathan Chabout, Abhra Sarkar, David B Dunson, and Erich D Jarvis. Male mice song syntax depends on social contexts and influences female preferences. *Frontiers in behavioral neuroscience*, 9:76, 2015.

[7] Kevin R Coffey, Russell G Marx, and John F Neumaier. Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology*, 44(5):859, 2019.

[8] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[9] Sébastien Derégnaucourt, Partha P Mitra, Olga Fehér, Carolyn Pytte, and Ofer Tchernichovski. How sleep affects the developmental learning of bird song. *Nature*, 433(7027):710, 2005.

[10] Michale S Fee and Jesse H Goldberg. A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. *Neuroscience*, 198:152–170, 2011.

[11] Olga Fehér, Haibin Wang, Sigal Saar, Partha P Mitra, and Ofer Tchernichovski. De novo establishment of wild-type song culture in the zebra finch. *Nature*, 459(7246):564, 2009.

[12] Simone Gaub, Matthias Groszer, Simon E Fisher, and Günter Ehret. The structure of innate vocalizations in foxp2-deficient mouse pups. *Genes, Brain and Behavior*, 9(4):390–401, 2010.

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[14] Kurt Hammerschmidt, Konstantin Radyushkin, Hannelore Ehrenreich, and Julia Fischer. The structure and usage of female and male mouse ultrasonic vocalizations reveal only minor differences. *PloS one*, 7(7):e41133, 2012.

[15] Stav Hertz, Benjamin Weiner, Nisim Perets, and Michael London. High order structure in mouse courtship vocalizations. *bioRxiv*, 2019.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

[17] Timothy E Holy and Zhongsheng Guo. Ultrasonic songs of male mice. *PLoS biology*, 3(12):e386, 2005.

[18] Aleksandr Ivanenko, Paul Watkins, MAJ van Gerven, Kurt Hammerschmidt, and Bernhard Englitz. Classification of mouse ultrasonic vocalizations using deep learning. *bioRxiv*, page 358143, 2018.

[19] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[20] Mimi H Kao and Michael S Brainard. Lesions of an avian basal ganglia circuit prevent context-dependent changes to song variability. *Journal of neurophysiology*, 96(3):1441–1455, 2006.

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

[22] Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *arXiv preprint arXiv:1907.04809*, 2019.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[26] David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

[27] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[28] Jacqueline R Miller and Mark D Engstrom. Vocal stereotypy and singing behavior in baiomyine mice. *Journal of Mammalogy*, 88(6):1447–1465, 2007.

[29] Nicolas Stephen Novakowski. The influence of vocalization on the behavior of beaver, castor canadensis kuhl. *American Midland Naturalist*, pages 198–204, 1969.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] Jonathan F Prather, Stephen Nowicki, Rindy C Anderson, Susan Peters, and Richard Mooney. Neural correlates of categorical perception in learned vocal communication. *Nature neuroscience*, 12(2):221, 2009.

[32] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. *arXiv preprint arXiv:1401.4082*, 2014.

[33] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[34] Monika Sadananda, Markus Wöhr, and Rainer KW Schwarting. Playback of 22-khz and 50-khz ultrasonic vocalizations induces differential c-fos expression in rat brain. *Neuroscience letters*, 435(1):17–23, 2008.

[35] W John Smith, Sharon L Smith, Elizabeth C Oppenheimer, and Jill G Devilla. Vocalizations of the black-tailed prairie dog, cynomys ludovicianus. *Animal behaviour*, 25:152–164, 1977.

[36] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

[37] Roland Sossinka and Jörg Böhner. Song types in the zebra finch poephila guttata castanotis. *Zeitschrift für Tierpsychologie*, 53(2):123–132, 1980.

[38] Daniel Soudry, Suraj Keshri, Patrick Stinson, Min-hwan Oh, Garud Iyengar, and Liam Paninski. Efficient" shotgun" inference of neural connectivity from highly sub-sampled activity data. *PLoS computational biology*, 11(10):e1004464, 2015.

[39] O Tchernichovski and PP Mitra. Sound analysis pro user manual. *CCNY, New York*, 2004.

[40] Maarten Van Segbroeck, Allison T Knoll, Pat Levitt, and Shrikanth Narayanan. Mupet—mouse ultrasonic profile extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations. *Neuron*, 94(3):465–485, 2017.

[41] J Craig Venter, Mark D Adams, Granger G Sutton, Anthony R Kerlavage, Hamilton O Smith, and Michael Hunkapiller. Shotgun sequencing of the human genome, 1998.

[42] Alex H Williams, Ben Poole, Niru Maheswaranathan, Ashesh K Dhawale, Tucker Fisher, Christopher D Wilson, David H Brann, Eric Trautmann, Stephen Ryu, Roman Shusterman, et al. Discovering precise temporal patterns in large-scale neural recordings through robust and interpretable time warping. *BioRxiv*, page 661165, 2019.

[43] Markus Woehr. Ultrasonic vocalizations in shank mouse models for autism spectrum disorders: detailed spectrographic analyses and developmental profiles. *Neuroscience & Biobehavioral Reviews*, 43:199–212, 2014.

# 4 Supplementary Figures

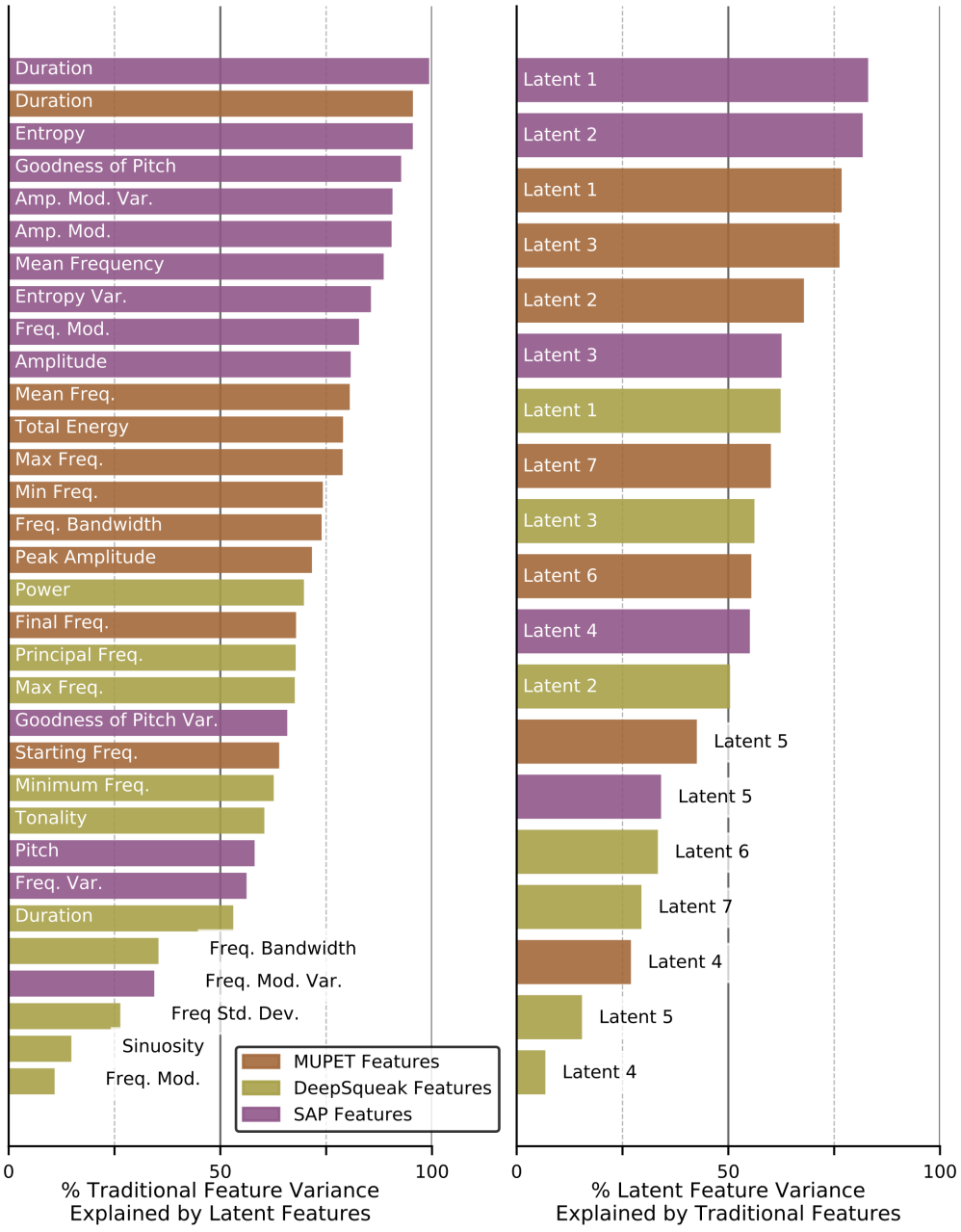


Figure S1: Left column: Named acoustic feature variance explained by latent features. Right column: Latent acoustic feature variance explained by named acoustic features.

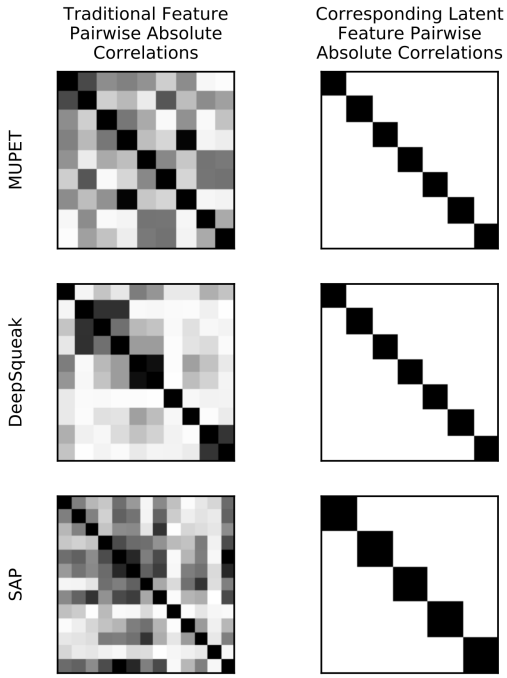


Figure S2: Traditional acoustic features are highly correlated. Left column: pairwise absolute correlations between named acoustic features when applied to the datasets in Figure 2. Right column: pairwise absolute correlations of latent features for the same datasets.

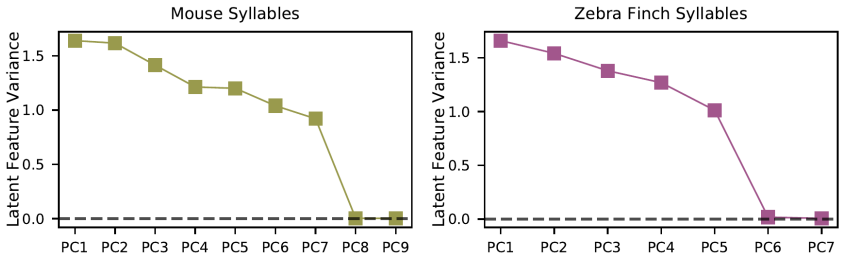


Figure S3: The VAE learns a parsimonious representation of acoustic features. When trained on mouse syllables (from Figure 2a), the VAE makes use of only 7 of 32 latent dimensions. When trained on zebra finch syllables (from Figure 5a), the VAE makes use of only 5 of 32 latent dimensions.

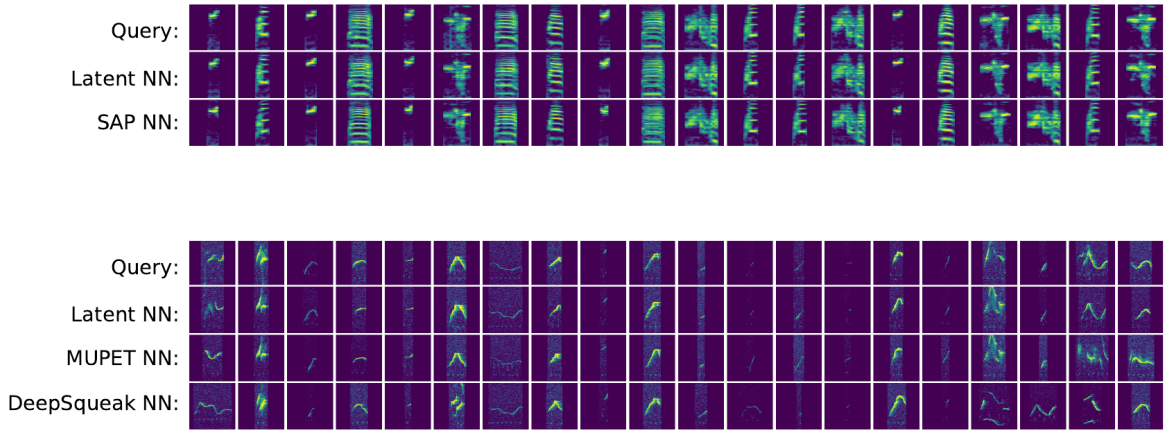


Figure S4: Representative sample of nearest neighbors returned by several feature spaces. Top block: Given 20 random zebra finch syllable spectrograms, both Sound Analysis Pro and latent acoustic features consistently find nearest neighbors of the same syllable type. Bottom block: Given 20 random mouse syllable spectrograms, latent, MUPET, and DeepSqueak feature spaces generally find acoustically similar nearest neighbors. However, latent features consistently return better matches than either MUPET or DeepSqueak features.

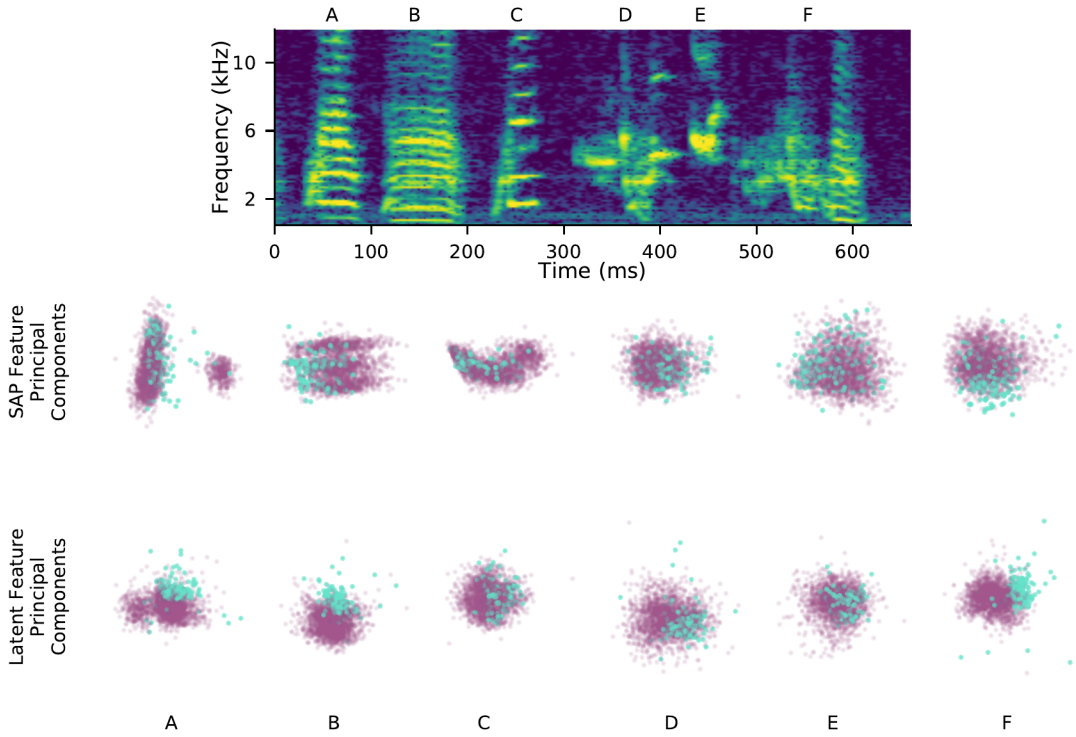


Figure S5: Latent features better represent constricted variability of female-directed zebra finch song. At top is a single rendition of a male zebra finch's song motif, with individual syllables labeled A-F. The top row of scatterplots shows each syllable over many directed (blue) and undirected (purple) renditions, plotted with respect to the first two principal components of the Sound Analysis Pro acoustic feature space. The bottom row of scatterplots shows the same syllables plotted with respect to the first two principal components of latent feature space. The difference in distributions between the two social contexts is displayed more clearly in the latent feature space, especially for non-harmonic syllables (D,E,F).

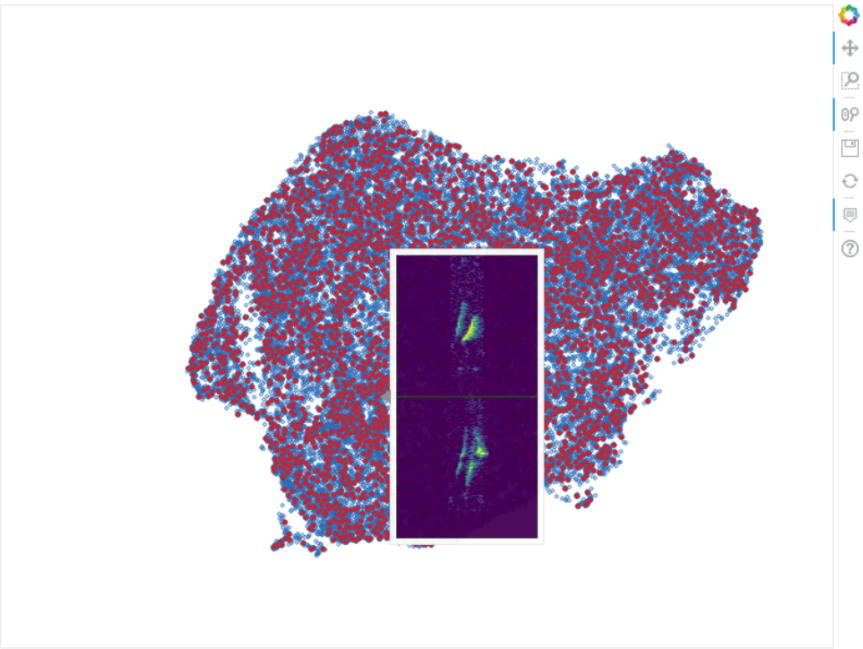

Figure S6: "Atlas" of mouse USVs. This screenshot shows an interactive version of Figure 4d in which example spectrograms are displayed as tooltips when a cursor hovers over the plot. This plot is hosted at the following web address: `https://pearsonlab.github.io/research.html#mouse_tooltip`

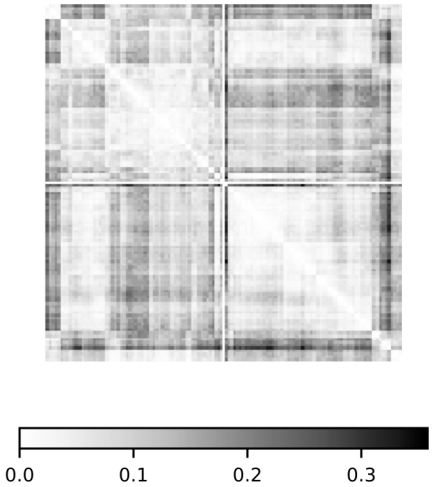


Figure S7: a) Maximum Mean Discrepancy for each pair of recording sessions from Figure 4f. Compare to 4e.

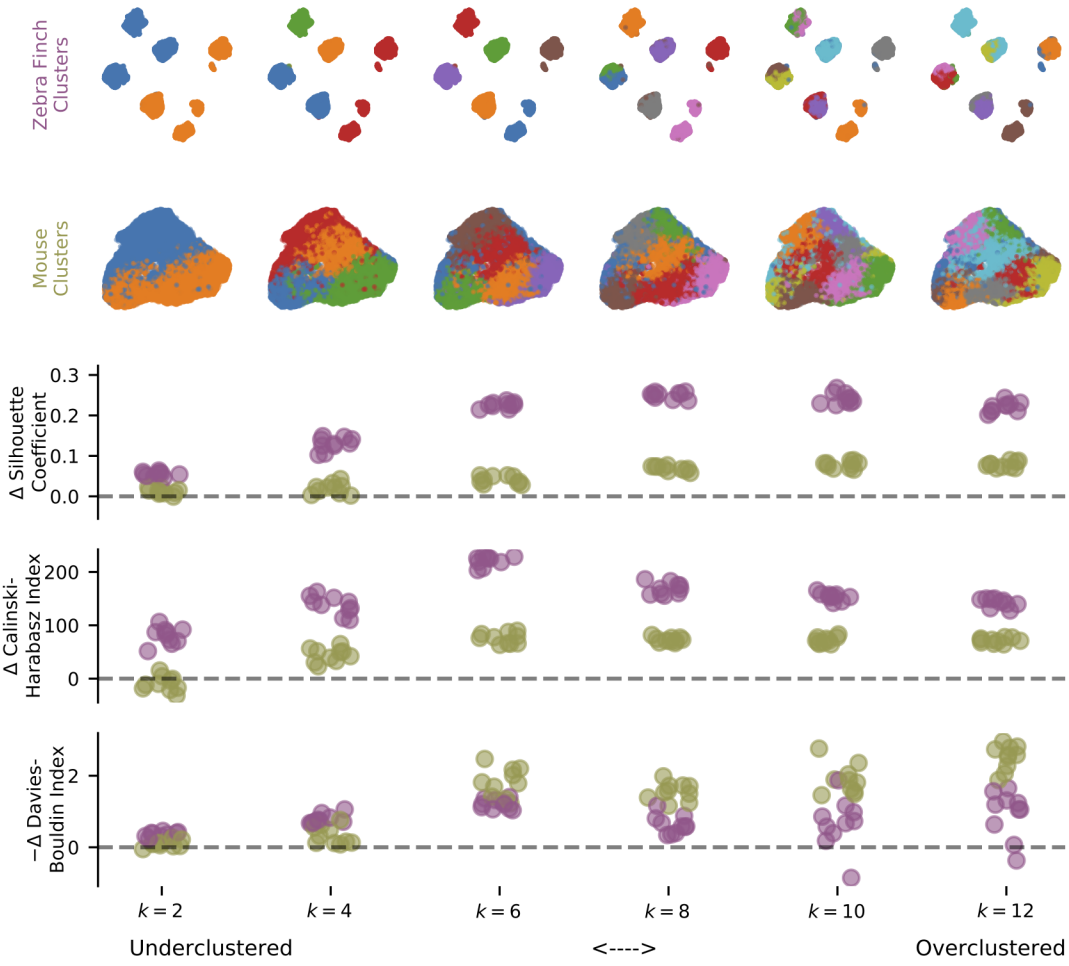


Figure S8: Three unsupervised clustering metrics evaluated on the latent description of zebra finch song syllables (Figure 5a) and mouse USV syllables (Figure 5b) as the number of components, $k$, varies from 2 to 12. Clustering metrics are reported relative to moment-matched Gaussian noise (see Methods) with a possible sign change so that higher scores indicate more clustering.

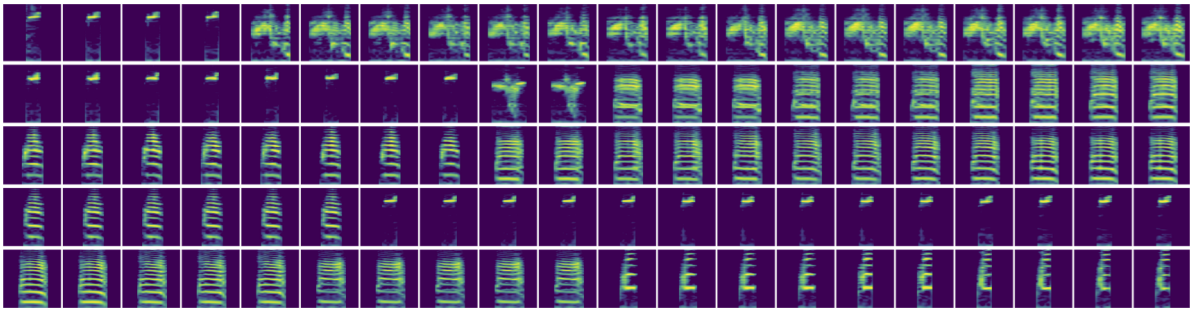

Figure S9: Absence of continuous interpolations between zebra finch song syllables. Each row displays two random zebra finch syllables of different syllable types at either end and an attempted smooth interpolation between the two. Interpolating spectrograms are those with the closest latent features along a linear interpolation in latent space. Note the discontinuous jump in each attempted interpolation. Compare with Figure 5e.

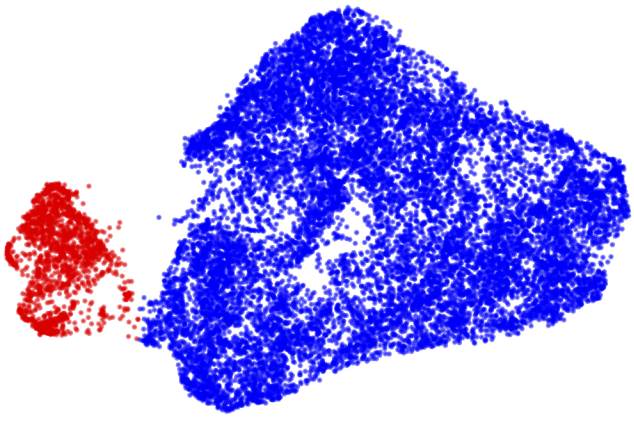

Figure S10: Removing noise from single mouse recordings (see Recordings). Above is a UMAP projection of all detected USV syllables. The false positives (red) cluster fairly well, so they were removed from further analysis.

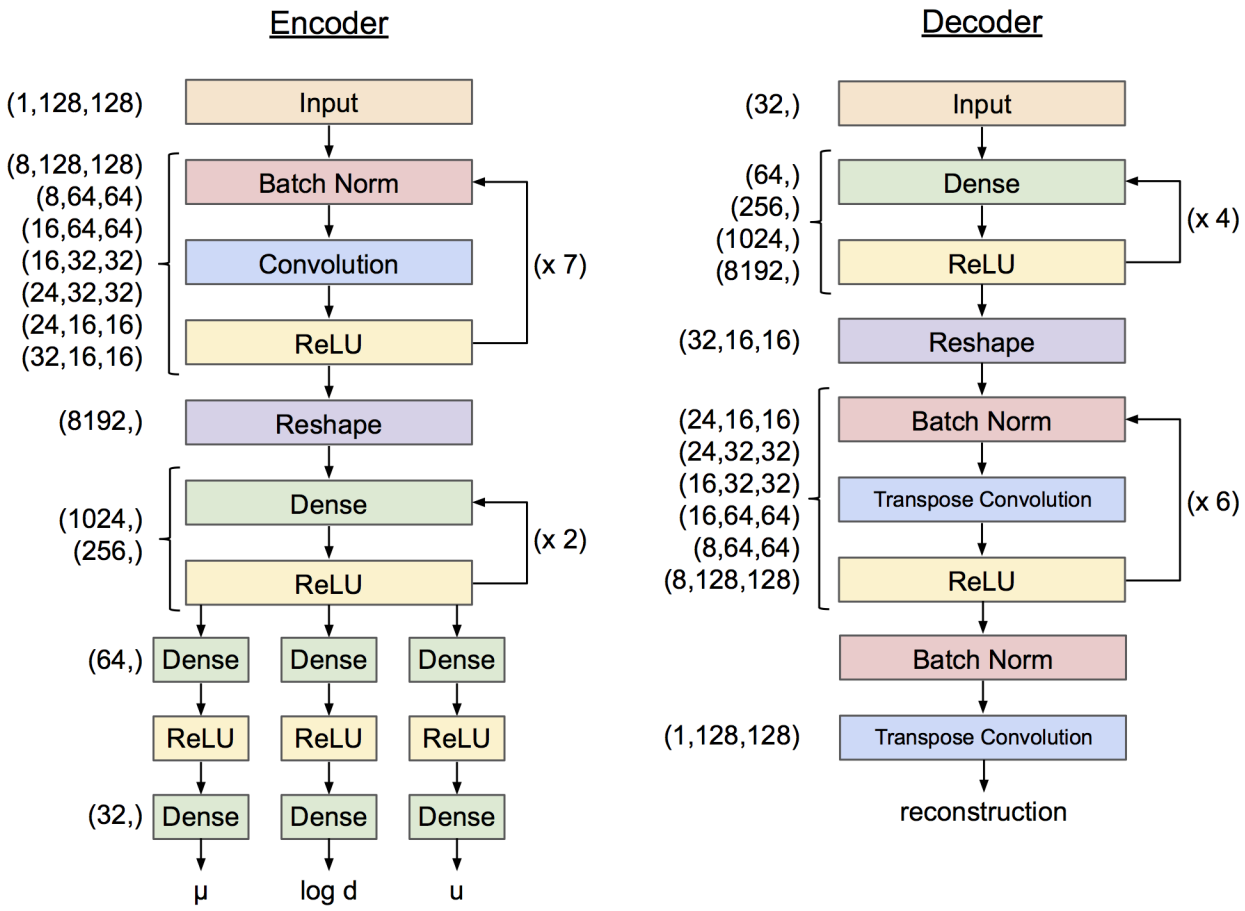


Figure S11: VAE network architecture. The architecture outlined above was used for all training runs. The looping arrows at the right of the encoder and decoder denote repeated sequences of layer types, not recurrent connections. For training details see Methods. For implementation details, see: `https://github.com/jackgoffinet/autoencoded-vocal-analysis`