

Data-Driven Analysis of Zebra Finch Song Copying and Learning

by

Samuel Brudner

Department of Neurobiology
Duke University

Date: _____

Approved:

Richard Mooney, Supervisor

Stephen Lisberger, Chair

Henry Yin

John Pearson

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Neurobiology
in the Graduate School of Duke University
2021

ABSTRACT

Data-Driven Analysis of Zebra Finch Song Copying and
Learning

by

Samuel Brudner

Department of Neurobiology
Duke University

Date: _____

Approved:

Richard Mooney, Supervisor

Stephen Lisberger, Chair

Henry Yin

John Pearson

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Neurobiology
in the Graduate School of Duke University
2021

Copyright © 2021 by Samuel Brudner
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Children learn crucial skills like speech in part by imitating the behavior of skilled adults. Similarly, juvenile zebra finches learn to sing by learning to imitate adults. This song learning process enables laboratory study of juvenile imitative learning. But it also poses behavioral quantification challenges. Juvenile zebra finches produce hundreds of thousands of complex vocalizations as they learn. These undergo learned changes with respect to acoustic features that are relevant to the animal but experimentally unknown *a priori*. Recent developments in machine learning provide tools to reduce the dimensionality of complex behaviors like vocalizations, plausibly simplifying the challenge of characterizing vocal development. These tools have not been validated on or applied to song learning problems.

Here, I validate the use of a variational autoencoder to extract copying-relevant features from zebra finch song. Then, I develop tools to quantitatively model the developmental changes in syllable acoustic distributions over these extracted features. I also develop a method to score syllable maturity on a rendition-by-rendition basis. These developmental quantifications reveal circadian behavioral patterns that differ between normally developing and untutored juveniles. This difference suggests that tutoring affects juvenile practice behavior beyond its role in determining the target acoustics. Importantly, the tools developed here operationalize abstract behavioral concepts at the heart of a long-standing neurobiological theory of song learning, so this theory's critical components can be tested in the near future.

For my mom, dad, Rebecca, and Elizabeth

Contents

Abstract	iv
List of Figures	xi
Acknowledgements	xii
1 Introduction	1
1.1 The zebra finch behavioral model and reinforcement learning	2
1.1.1 Adult zebra finch song is a fully learned policy	3
1.1.2 Tutoring and the reward function	4
1.1.3 Development of the song policy: syllables and temporal context	5
1.1.4 Syllable policy modification	7
1.1.5 Adult learning models of auditory errors and syllable policy modification	10
1.1.6 Summary	11
1.2 The zebra finch song system and reinforcement learning	12
1.2.1 The posterior motor pathway	12
1.2.2 Insufficiencies for learning by HVC and RA may be resolved by the AFP	14
1.2.3 The AFP may receive dopaminergic prediction errors	18
1.2.4 Policy updates and the AFP	22
1.2.5 Consolidation outside the AFP	23
1.2.6 Summary and current limitations	25

1.3	Quantifying juvenile song behavior	26
1.3.1	Feature-based approaches	26
1.3.2	Nearest neighbor age classification	27
1.3.3	Variational autoencoders	28
1.3.4	Summary	29
1.4	General summary	29
2	Variational Autoencoder Learns Copied Song Features	31
2.1	Introduction	31
2.2	Results	32
2.2.1	Syllable VAE captures pupil/tutor similarity	32
2.2.2	Shotgun VAE captures pupil/tutor similarity	35
2.3	Conclusions	38
2.4	Methods	39
2.4.1	Recordings	39
2.4.2	Audio Segmenting	39
2.4.3	Spectrograms	40
2.4.4	Model Training	40
2.4.5	Shotgun Test Data	41
2.4.6	Latent Representation of Sounds	41
2.4.7	UMAP	41
2.4.8	Syllable Type Clustering	42
2.4.9	Maximum Mean Discrepancy	43
3	Developmental Forward Models: Learning the dependence of acoustic distributions on age	46
3.1	Introduction	46
3.2	Results	47

3.2.1	Developmental changes in latent space reflect developmental spectrogram changes	47
3.2.2	Gaussian models capture within- and between-day changes in latent distributions	48
3.2.3	Gaussian model entropy is high in the morning and low in the evening	52
3.2.4	Gaussian models of isolate song	53
3.3	Conclusions	55
3.4	Methods	57
3.4.1	Recordings	57
3.4.2	VAE latent scoring	57
3.4.3	Preparing syllable type datasets	58
3.4.4	Gaussian model training	58
3.4.5	Evaluation of Gaussian models	59
3.4.6	Entropy datasets	60
3.4.7	Linear mixed effects models	60
4	Reverse Models: Scoring individual rendition maturity	64
4.1	Introduction	64
4.2	Results	65
4.2.1	Scoring syllable rendition maturity by predicting age	65
4.2.2	Daily predicted age increases depend on D1R signalling in Area X	67
4.2.3	Overnight consolidation is quantile-dependent	70
4.2.4	Developing isolate song exhibits consistent overnight reversion	73
4.3	Conclusions	74
4.4	Methods	77
4.4.1	Developmental dataset preparation	77
4.4.2	Pharmacology datasets	77

4.4.3	Predicted age model training	77
4.4.4	Predicted age model evaluation	78
4.4.5	Analysis of D1R treatment	79
4.4.6	Analysis of overnight shifts	80
5	Conclusions and future directions	83
5.1	Summary	83
5.2	Interpretation	86
5.2.1	What features of acoustic change underlie the quantile dependence of overnight predicted age shifts?	86
5.2.2	Mechanisms of variability patterns	88
5.2.3	The role of tutoring in song development	89
5.3	Testing a model for basal ganglia-based reinforcement learning	91
5.3.1	Review of an influential model	91
5.3.2	Dopaminergic reinforcement signals	92
5.3.3	Premotor contributions of the Anterior Forebrain Pathway	94
5.3.4	Consolidation outside the Anterior Forebrain Pathway	96
5.4	Limitations and extensions	97
5.4.1	Autoencoder assumptions and alternative assumptions	97
5.4.2	Time-based extensions	99
5.4.3	Modeling non-Gaussian song distributions in latent space	101
5.5	Conclusion	103
A	VAE Mechanics	104
A.1	The decoder formalizes generative constraints	104
A.2	The decoder motivates a computational and inferential problem	106
A.3	The encoder	107
A.4	Implementation in neural networks	108

B AFP experiments	110
B.1 Introduction	110
B.2 Results	111
B.2.1 Area X optogenetics	111
B.3 Discussion	114
B.3.1 Interpreting the null result	114
B.3.2 Alternative approaches for future work	117
Bibliography	119
Biography	130

List of Figures

2.1	Example pupil and tutor.	33
2.2	Group syllable VAE.	34
2.3	Group shotgun VAE.	37
2.4	Shotgun UMAP modification.	42
2.5	Shotgun MMD decomposition.	44
2.6	Autoencoder architecture.	45
3.1	Example syllable development.	48
3.2	Developmental gaussian models.	51
3.3	Daily entropy.	62
3.4	Entropy in isolates.	63
4.1	Predicting syllable age.	66
4.2	D1R signaling and predicted age.	68
4.3	Testing overnight consolidation.	71
4.4	Isolate overnight consolidation.	82
B.1	Laser offset in predicted age.	113

Acknowledgements

I want to thank Rich Mooney for allowing me to join his lab when I did not know much neuroscience at all. It has been a long, rewarding journey since then, and the whole lab team has been very supportive throughout. I especially want to recognize Michael Booze, Jiaxuan Qi, Erin Hisey, Matt Kearney, Valerie Michael, and Jonna Singh Alvarado for the help they've given along the way. Jack Goffinet has been a great collaborator and endlessly tolerant of long lists of technical questions. My committee — Steve Lisberger, Henry Yin, Henry Greenside, and John Pearson — has provided invaluable support too.

I would like to acknowledge the NICHD for providing funding through the F31 Fellowship Award since 2020.

On a more personal note, I want to thank my parents and sister for their love and support. Last, I could not have done this without Elizabeth.

Introduction

Children learn speech and other crucial skills by imitating adult behaviors. Constraints on these learning problems suggest they involve reinforcement learning, a collection of processes of great interest in neuroscience. Nonetheless, the mechanisms underlying imitative juvenile learning in particular are not well understood.

Research on song copying by juvenile zebra finches has the potential to address this void. But without control over the natural learning process, experimenters must infer relevant behavioral parameters from the behavior itself. These inferences are especially challenging because vocal behavior is high-dimensional and complex. Past efforts to quantify vocal behavior have yielded important insights, but also have limitations and in some cases conflict.

Recently, I participated in a collaboration — led by Jack Goffinet and supervised by John Pearson — to use variational autoencoders (Kingma and Welling (2013), Rezende et al. (2014)), an unsupervised dimensionality-reduction technique, to describe the complex acoustics of individual zebra finch song segments using only a small number of descriptive variables per segment (Goffinet et al. (2021)). In this thesis, I build on this tool by designing methods to describe song development processes

in terms of these descriptive variables. I focus on methods to describe aspects of behavior that may rely on basal ganglia-dependent reinforcement learning, although the descriptions themselves are agnostic to mechanisms generating the behavioral patterns. In particular, the research in this thesis is aimed to (1) address whether variational autoencoder song descriptions effectively capture copied song features; (2) develop a tool to capture developmental changes in the distribution of song syllable acoustics as juvenile birds practice; (3) develop a tool to score the evaluative quality of individual renditions of juvenile practice song.

1.1 The zebra finch behavioral model and reinforcement learning

Juvenile imitative learning involves practicing in order to adapt behavioral output towards an imitation target. More generally, the problem of adapting behavior to achieve a goal through practice arises in studies of animal behavior, neuroscience, computer science, and engineering, and is referred to as reinforcement learning. Reinforcement learning problems share a few defining characteristics that will be useful concepts throughout this thesis. These problems involve an agent that acts in accordance with a policy, essentially a context-dependent action plan. The agent’s actions influence a quantity called “reward” through mechanisms that are not completely known to the agent. Reward has different specific meanings in different reinforcement learning problems; in general it serves to formalize the agent’s goal as the production of behavior that maximizes reward. Solutions to reinforcement learning problems are procedures that allow the agent to acquire policies that maximize reward. Solving a reinforcement learning problem requires a stochastic policy, so the agent can observe the differential reward outcomes of different actions taken in similar contexts (Barto et al. (1983)).

These concepts help to frame practice behaviors, including in the case of juvenile imitative learning. I will connect these concepts specifically to a powerful

laboratory model: juvenile zebra finches learning to imitate adult tutors. My goal in this thesis is to develop methods that leverage experimentally accessible vocal data in order to quantitatively describe the policies and rewards of practicing juvenile zebra finches. I hope this work enables experimental tests of the neurobiological mechanisms of reinforcement learning in this model system. In this section, I review the model behavior through the lens of reinforcement learning.

1.1.1 Adult zebra finch song is a fully learned policy

Adult male zebra finches produce hierarchically structured song as a component of courtship displays to female birds (Morris (1954)). At the highest organizational level, song occurs in bouts that often last several seconds. Bouts typically begin with a series of innate introductory notes, after which they are composed of multiple repetitions of a core song “motif,” with multiple motif repetitions usually concatenated with variable numbers of intervening connective sounds. Although song motifs differ across individuals, any particular adult produces a single motif with a highly stereotyped structure from rendition to rendition. Zebra finch song motifs are composed of acoustically distinct syllables (usually 2 to 5) produced in a fixed order. Syllables are produced during expiration, and syllable boundaries are defined by ~ 50 ms silent gaps, during which adults typically take a “minibreath” of inspiration (Franz and Goller (2002)).

Although zebra finch song is a courtship display, zebra finches sing readily by themselves, often over a thousand times a day. Song produced in the absence of a female target is called “undirected” song. It can be distinguished from female-directed (or just “directed”) song by its slightly increased rendition-to-rendition acoustic variability (Kao et al. (2005)) and slightly slower pace (Sossinka and Böhner (1980)). The subtle differences between directed and undirected song are readily detected by female birds who prefer the directed song to undirected song (Woolley and Doupe

(2008)).

In the reinforcement learning framework, the various vocal sounds a zebra finch can produce form the set of possible vocal motor actions he can take. The animal’s goal is to produce the right sounds in the right order, with respect to his individual preference. In other words, this goal sequence defines reward-maximizing behavior. Achieving this appropriately sequenced behavior requires a map from intramotif times to vocal motor actions. So, adult song is a fully learned policy that maps intramotif times to vocal motor actions. This intuitive characterization is additionally motivated by underlying neurobiology, as I will discuss in Section 1.2.

1.1.2 Tutoring and the reward function

In the wild, zebra finches live in large colonies, and juvenile birds have many opportunities to interact with singing adult males, including their fathers (Zann (1990)). The song they hear in these interactions affects their learning goal, in the sense that they learn to produce similar song themselves (Morris (1954)). Zebra finches will even learn to copy syllables with species-atypical spectral content if those sounds are produced by a heterospecific tutor (Clayton (1989)). As with speech acquisition (Curtiss et al. (1974)), tutor model exposure must occur within a critical developmental window (~ 20 to ~ 60 days post hatch (dph); Immelmann (1969)), or animals will develop abnormal ‘isolate’ songs (Konishi (1965), Immelmann (1969), Marler (1970), Price (1979)).

Although tutor song acoustics clearly influence the song template that guides juvenile learning, careful observation has revealed the influence of experience-independent influences on this template. Exposure to unusually simple zebra finch song models increases the likelihood that pupils will incorporate improvised song elements into their motif (Tchernichovski et al. (2021)). These pupil/tutor mismatches are difficult to explain on the basis of production constraints in the pupil. Rather, they suggest

a model where prior expectations for the template interact with the experience of a tutor to produce an updated template.

After tutoring, juveniles must hear their own vocalizations in order to learn to match their internal template (Konishi (1965); Marler and Waser (1977)). In fact, Konishi (1965) observed that birds deafened as juveniles produced more abnormal song than juveniles reared without tutors. He concluded that even without exposure to an external example song, birds use auditory feedback and an ‘innate template’ to guide their development, consistent with previously mentioned evidence for experience-independent template priors.

In reinforcement learning, parameters of the reward function define the agent’s goal. Thus, in the case of juvenile zebra finch song learning, the internal template is formalized as these reward function parameters. In particular, the template defines a reward function that operates at least in part on auditory feedback of the juvenile’s own vocalizations. For the most part, reward increases with increasing similarity between the juvenile vocalization and the tutor acoustics, so that a reward-maximizing learning procedure increases this similarity.

1.1.3 Development of the song policy: syllables and temporal context

Above, I described adult song policies as maps from motif times to vocal motor gestures. To extend this policy framework to juvenile song, it will be useful to reframe the description of song policies in terms of syllables. Here I describe patterns of syllable formation in juvenile song and formulate a syllable-based description of song policy.

Juveniles begin to produce song around 35dph (Price (1979)). This early production mode is termed subsong. Subsong syllable durations have characteristic exponential decay distributions (Aronov et al. (2011), Veit et al. (2011)), consistent with syllable termination occurring randomly at a fixed rate after syllable initiation.

Subsong syllable and gap production mechanisms differ from adult mechanisms as well. In particular, breathing is relatively uncoordinated with subsong vocalizations unlike its tight coupling to syllable onsets and offsets in adults (Veit et al. (2011)).

Between 40 and 50dph, in the transition to “plastic song,” an overrepresented syllable duration appears (Aronov et al. (2011), Veit et al. (2011)). Syllables that share this characteristic modal duration are often acoustically similar to one another, and are produced as a rhythmic stream of repeating sounds called “protosyllabic song” (Tchernichovski et al. (2001)). In other cases, several different syllable “types” that presage future adult syllable types may appear together (Liu et al. (2004)). In the case that a “protosyllable” with unimodal duration and acoustic distribution appears first, it can be used to generate novel syllable types through a process called “sound differentiation *in situ*” (Tchernichovski et al. (2001)). In this process, the unimodal protosyllable acoustic distribution gradually splits into a multimodal distribution with modes corresponding to different syllable types. Differentiation often results in bouts composed of sounds from one mode repeated multiple times, followed by repetitions from another mode, a practice behavior called the “serial repetition strategy” (Liu et al. (2004)). Renditions of the syllable near bout onsets can also be “fused” to short bout-initiating sounds to generate a novel syllable type (bout onset differentiation; Okubo et al. (2015)).

These behavioral patterns complicate the identification of motifs with clear starts and ends, and in turn complicate the idea that song policy is a map from motif times to vocal motor actions. In the case of juvenile singing, syllable repetitions suggest that the animal is repeatedly entering the same song context multiple times in the same “motif.” These observations lead to a modified view of the song policy. We can view the juvenile policy as hierarchical: one high-level policy stochastically specifies syllable sequence. The syllable type specified by this high-level policy provides overarching context to a downstream stochastic policy from intrasyllabic times to vocal

motor actions. The analysis developed in this thesis is concerned with modeling learning in the downstream, intrasyllable policy. I will use the term policy to refer to this downstream policy, not the processes that sequence different syllable types. (The syllable-based framing of song policies also permits extending this framework to other songbird species, which produce syllables according to syntactic rules rather than in fixed sequences; see, e.g., Cohen et al. (2020a) for work in canaries.)

1.1.4 Syllable policy modification

By around 60dph, juvenile songs mostly contain identifiable syllable types that are precursors to syllables appearing in the animal’s final song. As described above, this behavior constitutes a policy that stochastically maps intrasyllabic times to vocal motor actions. These policies are modified during plastic song so they maximize the output of a reward function parameterized by an internal template. My central aim in this thesis is to develop methods to quantitatively describe these syllable policies as they develop, as well as this reward function. My efforts build on previous characterizations of syllable development that are summarized below.

The role of sleep and circadian patterns in syllable modification

Syllable policies are regulated by circadian or sleep-dependent processes (Derégnaucourt et al. (2005), Miller et al. (2010), Ravbar et al. (2012), Kollmorgen et al. (2020)). In an initial landmark study, Derégnaucourt et al. (2005) suggested that the approach of juvenile syllable acoustics to target acoustics exhibited a non-monotonic circadian pattern. In particular, they suggested that the changes in syllable policies acquired through daytime practice partially “revert” during sleep. Birds exhibiting greater sleep reversion ultimately produce better tutor copies. More recently, Kollmorgen et al. (2020) demonstrated that the tendency of sleep to “revert” prior day syllable changes depends on the acoustic feature chosen for analysis. Instead of picking

specific acoustic features to analyze, these authors developed a holistic syllable maturity score based on the spectrogram similarity of renditions produced at different ages. (This procedure has similarities to the independently developed ‘predicted age’ metric I present in Chapter 4.) These authors leverage this maturity score to infer that, with respect to systematically and slowly changing features, most within-day changes to syllable distributions consolidate rather than revert overnight; only the population of syllable renditions that are immature for their age undergo a reversion in maturity overnight. Instead, these authors conclude that “sleep reversion” occurs principally in learning-orthogonal acoustic dimensions that simply oscillate in place at long timescales by changing each day and resetting overnight without any long-term movement.

At present these discrepancies are unresolved. On the one hand, Kollmorgen et al. (2020) identify legitimate interpretative challenges that arise from hand-selecting acoustic features to analyze among many possibilities. On the other hand, Derégnaucourt et al. (2005) specifically analyzed syllables in terms of an acoustic feature that exhibited systematic changes over long time scales, discarding syllable types for which the feature did not exhibit a systematic trend. In this way, their analysis appears to contradict the claims of Kollmorgen et al. (2020) because their chosen acoustic feature both systematically changes over development and reverts during sleep. Moreover, the multidimensional space that Kollmorgen et al. (2020) infer from their one-dimensional maturity score has multiple interpretations, including some that may be consistent with the conclusions of Derégnaucourt et al. (2005). In particular, a point in the multidimensional space developed by Kollmorgen et al. (2020) represents an entire collection of many song renditions. Proximity between two points reflects the overlap of maturity score distributions from two collections. Intuitively, rendition collections with maturity score distributions that are centered far from each other will map to distant points. Less intuitively, rendition collections with

maturity score distributions of different width will also map to relatively distant points. This consideration complicates interpretations of the ‘directions’ that Kollmorgen et al. (2020) identify. In particular, the learning-orthogonal ‘directions’ may relate to circadian patterns in rendition-to-rendition variability, rather than learning-orthogonal translations of acoustic distributions. Certainly, the impact of changes in song variability on the geometry of the space developed by Kollmorgen et al. (2020) complicates interpretations of their methods.

Policy variability

To support learning, policies must be stochastic maps onto vocal motor actions, i.e., context-conditioned probability distributions over possible actions. The shapes of these distributions can influence learning (Zhou et al. (2018)). In juvenile song learning, the maturation of plastic song syllables is associated with a decrease in the randomness of syllable policies. In fact, subsyllabic regions appear to decrease in rendition-to-rendition acoustic variability as they approach their crystallized endpoint, even while other subsyllabic regions of the same syllable are still relatively immature in both similarity to adult crystallized form and variability (Ravbar et al. (2012)).

In addition to these multiday changes in policy randomness, multiple reports indicate that acoustic variability oscillates on a daily timescale. However, one study reports that variability is low in the morning and high in the afternoon (Miller et al. (2010)), and another reports that variability is low in the afternoon and high in the morning (Ravbar et al. (2012)). The first result may owe to daily oscillations in the acoustic heterogeneity within each individual syllable rendition, rather than owing to rendition-to-rendition variation in the execution of a syllable policy (Derégnaucourt et al. (2005), Ravbar et al. (2012)), but this discrepancy is not fully resolved.

1.1.5 Adult learning models of auditory errors and syllable policy modification

As previously reviewed, a variety of behavioral processes occur during juvenile development: template updating during tutoring, syllable formation and sequencing, and syllable policy modification. During syllable policy modification, the animal listens to his own vocalizations to assess their reward value, and leverages these assessments to improve his performance. Because of the complexity of this process in juveniles, researchers have attempted to model aspects of this reinforcement learning process in adults to gain greater experimenter control. In particular, the detection of low-reward syllable quality, and downstream adaptation of vocal output, have been modeled in adult birds. These models have enabled testing aspects of a neural hypothesis for syllable policy reinforcement learning (see Section 1.2), so I review these behavioral models here.

In the primary adult model for detection of low-reward syllable quality, the experimenter plays an external sound over an ongoing syllable vocalization. In this way, the acoustic feedback available to the animal consists of his regularly vocalized syllable and an overlaid sound deviating from the target output — an acoustic “error” (Tumer and Brainard (2007), Andalman et al. (2009)). In a second, related paradigm, birds are equipped with headphones whose output largely replaces natural airborne auditory feedback, enabling experimenters to substitute pitch-shifted distortions of ongoing singing for veridical auditory information (Sober and Brainard (2009)). These distortions can also be interpreted as acoustic errors.

These acoustic error models have been extended to drive reinforcement learning in a tractable adult model of syllable modification. During “pitch learning,” the experimenter targets acoustic perturbation to a target syllable conditional on that syllable’s rendition-by-rendition pitch. For example, the experimenter may perturb most target syllable renditions, with only the highest pitch 25% of renditions “escap-

ing.” After hours under these example conditions, the experimental bird will shift the distribution of pitches that he sings, producing high pitch renditions that escape the perturbation more often and low pitch renditions that elicit the perturbation less often (Tumer and Brainard (2007), Andalman et al. (2009)). The pitch shifts induced by this paradigm are often subsyllabically localized (Charlesworth et al. (2011)), reminiscent of evidence that subsyllabic regions of juvenile syllables can mature somewhat independently (Ravbar et al. (2012)). Similarly, headphone substitution of pitch-shifted acoustic feedback leads birds to compensate for the pitch manipulation (Sober and Brainard (2009)). Learning from any of these manipulations permits another adult model of syllable modification: after pitch learning drives birds away from baseline behavior, removal of experimental error manipulations causes birds to spontaneously “recover” to their baseline acoustic distributions (Tumer and Brainard (2007), Sober and Brainard (2009)). In these models, birds modify their syllable policies to maximize reward by eliminating actions that produce low reward due to experimental perturbations. However, the exact relationship between this adult behavior and juvenile learning is not known.

A final adult model of syllable modification is deafening-induced song degradation. Song undergoes spectral and temporal degradation in the weeks following deafening (Nordeen and Nordeen (1992), Lombardino and Nottebohm (2000), Tschida and Mooney (2012)). Thus, this manipulation clearly induces vocal motor plasticity, however it provides much less experimental control than pitch learning and can be distinguished from other learning models in that the changes are not adaptive.

1.1.6 Summary

Zebra finch song learning has emerged as a powerful model for studying developmental skill acquisition, especially juvenile imitative learning. This behavior involves multiple interrelated processes. The overarching motivation of my work is to facili-

tate clearer mechanistic and algorithmic understanding of this behavior. I focus on analytic techniques that will facilitate understanding the process of syllable policy modification. Although I hope these tools will be useful in other applications as well, my primary motivation is to render experimentally testable a long-standing neurobiological hypothesis for reinforcement learning of syllable policies. I describe that hypothesis in the following section.

1.2 The zebra finch song system and reinforcement learning

Song production and learning depend on the nuclei of the song system, canonically divided into two interacting forebrain circuits called the Posterior Motor Pathway (PMP) and the Anterior Forebrain Pathway (AFP). These pathways are homologous to mammalian hierarchical cortical motor pathways and cortico-basal ganglia pathways respectively. The PMP implements a stereotyped policy from intrasyllabic times to vocal motor actions. The AFP implements a stochastic policy from intrasyllabic times to perturbations of PMP actions. The AFP policy is well suited to implement a reinforcement learning algorithm. The behavioral analysis developed in this thesis seeks to operationalize predictions of this theory of AFP function.

1.2.1 *The posterior motor pathway*

The PMP consists of the Robust nucleus of the Arcopallium (RA) and HVC (proper name). HVC projects to RA (Nottebohm et al. (1976), Nottebohm et al. (1982)). Both nuclei are specializations of the bird pallium, the embryological tissue that in mammals gives rise to the cortex, as well as the olfactory bulb, claustrum, and pallial amygdala (Striedter (1997)). Although precise homology to the mammalian brain has been controversial, recent transcriptomic evidence suggests that RA and HVC projection neurons are homologs of mammalian ventral pallium (non-cortical pallium), that evolved effector gene expression profiles resembling mammalian cor-

tical layer 5, subcerebral projection neurons. The interneurons of RA and HVC appear to be direct homologs of interneuron types found both in mammalian cortex and mammalian ventral pallium (Colquitt et al. (2021)). Here I present evidence that these nuclei implement a highly stereotyped policy from song temporal contexts represented by HVC activity to vocal actions elicited by RA activity.

RA is topographically organized song motor cortex

The robust nucleus of the arcopallium (RA) exerts proximate control over song by the forebrain song system. RA projects topographically to the brainstem centers that control breathing, and to nXIIIts, which contains myotopically arranged motor neurons that innervate syringeal muscles (Vicario and Nottebohm (1988), Wild (1993)). Unilateral RA lesions degrade song syllable spectral content, although they spare phrase structure in canaries (Nottebohm et al. (1976)). Bilateral RA lesions lead to “silent song” production, in which birds adopt posture and behaviors associated with singing, but do not produce sound (Nottebohm et al. (1976)). Cooling RA degrades syllable acoustics as well (Long and Fee (2008)).

During song motifs, RA neurons burst in temporally precise patterns relative to song time (Yu and Margoliash (1996)). RA burst patterns for spectrally similar syllables are similar (Yu and Margoliash (1996)), and RA neuron bursts can predict spectral features across syllables, in particular pitch (Sober et al. (2008)). Moreover, submillisecond variations in the timing of RA bursts correlate with variations in the timing of acoustic syllable features (Chi and Margoliash (2001)). Together, these results suggest that RA burst patterns specify vocal motor actions producing different sounds.

HVC is a premotor temporal context signal

HVC projects to RA (proper name; Nottebohm et al. (1976)), to form predominantly AMPAR-mediated synapses (Mooney and Konishi (1991)). This projection is critical for normal song production. Bilateral HVC lesions in canaries can lead to “silent singing” (Nottebohm et al. (1976)), though in zebra finches HVC lesions — including lesions of the HVC_{RA} projection specifically — degrade song quality, leading to the production of subsong-like behavior (Aronov et al. (2008)), including a subsong-like relation between breathing and singing, with long expiratory pulses generating multiple syllables (Veit et al. (2011)).

In adults, each HVC_{RA} projection neuron fires an action potential burst at one precise moment in the zebra finch motif. Different HVC_{RA} neurons fire at different times, apparently tiling the duration of each motif (Hahnloser et al. (2002)). It is easy to imagine that if each HVC neuron activates the appropriate effectors in RA for “its” song moment, the stereotyped cascade of activity through HVC would generate a stereotyped sequence of RA bursts that in turn generate the song sequence. In fact, HVC is the dominant influence on RA activity in adult birds (Garst-Orozco et al. (2014)).

1.2.2 Insufficiencies for learning by HVC and RA may be resolved by the AFP

HVC provides temporal context and RA executes context-specific actions under its influence. This input-output map can be reasonably viewed as a policy. However, the neural components described so far are insufficient for reinforcement learning in two respects. First, the influence of HVC on RA, and therefore on song, is extremely stereotyped, even in juveniles (Ölveczky et al. (2005)). It lacks the stochastic quality of policies required for a reinforcement learning. Second, no mechanism to convey reward or reward prediction error directly to HVC or RA has been described, despite direct efforts to find such a signal (Leonardo (2004), Kearney et al. (2019)). These

learning insufficiencies are remedied in a model of the AFP, a second collection of song system nuclei. These nuclei are the Lateral Magnocellular Nucleus of the Anterior Nidopallium (LMAN), the Medial portion of the DorsoLateral nucleus of the anterior thalamus (DLM), and Area X. In fact, lesions that spare the PMP but disrupt the nuclei of the AFP prevent song learning, even though these lesions do not impair production of song that corresponds to a fully learned policy (Bottjer et al. (1984), Sohrabji et al. (1990)).

Overall, the AFP perturbs the stereotyped relationship between HVC temporal context patterns and RA action patterns. As I will review, the AFP receives temporal context information from HVC and generates variable activity along parallel premotor channels that influence different subregions of RA. In this way, the overall system can be viewed as instantiating a policy that stochastically maps HVC contexts to RA perturbations. This map remedies the first learning deficiency apparent in our model of the PMP, allowing stochastic exploration of different perturbations. I will review evidence that the AFP receives dopaminergic reward information for song learning. This dopaminergic afferent provides a plausible mechanism to remedy the second learning deficiency of the PMP model; by providing reward feedback, it allows learning to produce preferences for reward-maximizing AFP perturbations. These observations motivate a model in which juvenile song learning involves reinforcement learning with respect to a reward gradient in the space of AFP timing-to-perturbation policies. This model has support from experiments in adult animals.

In this section, I present the evidence for these claims. At root, the analyses developed in this thesis are motivated by the goal to operationalize the behavioral predictions of this reinforcement learning model of the AFP in juvenile song learning.

AFP output: stochastic perturbation channels

Glutamatergic LMAN projection neurons directly synapse onto RA neurons (Nottebohm et al. (1982), Bottjer et al. (1989), Mooney and Konishi (1991)), where they form predominantly NMDAR-mediated synapses on the same RA neurons that receive direct HVC input (Mooney and Konishi (1991)). This projection is topographically structured with different LMAN subregions projecting to different RA subregions (Iyengar et al. (1999)), suggesting an organization into discrete channels to perturb RA's topographically organized output channels (Vicario and Nottebohm (1988)).

This parallel channel architecture governs connectivity inside the AFP as well. LMAN axons projecting to RA bifurcate and project to Area X (Nottebohm et al. (1976), Nottebohm et al. (1982), Vates and Nottebohm (1995)), providing a perturbation efference copy that is topographically structured: Area X MSNs do not receive inputs from different LMAN output channels (Vates and Nottebohm (1995)). In turn, Area X output neurons project in an almost one-to-one manner onto DLM projection neurons (Luo et al. (2001)). Finally these DLM neurons project back to LMAN (Okuhata and Saito (1987), Vates and Nottebohm (1995)) in a topography-preserving loop (Luo et al. (2001)).

These parallel projection channels stochastically perturb song-related action patterns in RA. LMAN firing rates increase during singing, and LMAN stimulation during singing induces brief acoustic distortions (Kao et al. (2005), Kojima et al. (2018)), indicating the ability of the AFP to modulate ongoing adult song. Although LMAN lesions do not dramatically alter adult singing (Bottjer et al. (1984)), LMAN neurons exhibit firing pattern variability from rendition to rendition during undirected song (Kao et al. (2005), Woolley et al. (2014)) and birds with LMAN lesions produce more stereotyped undirected song (Kao et al. (2005)). Conversely,

directed song is associated with less variable LMAN activity (Kao et al. (2005), Woolley et al. (2014)), reduced LMAN bursting (Woolley et al. (2014)), and less variable behavior (Kao et al. (2005)). In addition to this cross- rendition variation, within-syllable acoustic fluctuations are reduced following LMAN lesions, and during directed singing (Kojima et al. (2018)), indicating that the perturbations have fast temporal dynamics.

LMAN influences the acoustic structure of juvenile syllables as well. Juvenile plastic song syllable exhibit rendition-to-rendition acoustic variability that distinguishes them from adult syllable distributions. LMAN inactivations lead juveniles to produce highly acoustically stereotyped versions of their plastic song syllables (Ölveczky et al. (2005)). Moreover, LMAN inactivations in plastic song lead to prematurely stereotyped song-aligned burst patterns in downstream RA (Ölveczky et al. (2011)).

The AFP is an anatomically distributed map from temporal contexts onto output channels

The prior section indicates that the output of the AFP is partly stochastic across repetitions of an HVC temporal context. However, this output must also be sensitive to temporal context for reinforcement learning to adapt these perturbations if the perturbation-reward contingency is temporal context-sensitive. In this section, I argue that the overall AFP architecture maps HVC patterns onto perturbation channel outputs to accomplish this context sensitivity.

HVC projects to Area X (Nottebohm et al. (1976), Nottebohm et al. (1982)). These HVC_X neurons burst in temporally precise pattern like HVC_{RA} cells, although HVC_X burst patterns are generally denser than those of HVC_{RA} cells; individual HVC_X cells burst up to 4 times per motif (Kozhevnikov and Fee (2007)). By responding to its HVC inputs, Area X mediates the context-sensitivity of the AFP's

perturbation output. The specific transformation constitutes a policy, and the set of all possible transformations is a reinforcement learning policy search space. Suggestively, HVC_X neurons project widely inside Area X (Nottebohm et al. (1982)), so each can influence many perturbation channels.

This gross anatomy is consistent with the ability to find adaptive policies among all possible transformations of song time contexts to RA perturbations. Choice among policies in this neuroanatomical framework constitutes the specification of a pattern of synaptic strengths across the HVC to Area X interface. In the next section I examine how search across policies should be informed by the context-sensitive action-reward outcome contingency.

1.2.3 The AFP may receive dopaminergic prediction errors

In reinforcement learning theory, policy modifications during search are informed by feedback about the reward that is observed following action. This feedback often comes as a reward prediction error signal. In a large body of work outside the songbird field, dopaminergic afferents to the striatum have been observed to convey this kind of information. Next, I motivate the hypothesis that dopaminergic afferents to Area X convey reward prediction error to update perturbation policies. To motivate this hypothesis, I review suggestive homologies between the AFP and non-songbird cortico-basal ganglia-thalamocortical loops for which the dopaminergic pathway is best characterized. In particular, I present evidence for the current understanding that Area X contains the cell types and connectivity patterns of mammalian striatum and pallidum. I also present evidence from recent experiments that directly examine the dopaminergic afferent to Area X in finches.

Area X contains mammalian striatal cell type homologs

Like the medium spiny neurons of mammalian striatum, the most numerous cells in Area X are small, densely spiny neurons with MSN-like intrinsic electrophysiological features. These neurons receive monosynaptic excitatory inputs from cortex-like song nuclei HVC and LMAN (Farries and Perkel (2002), Farries et al. (2005)), express dopamine receptors (Ding and Perkel (2002), Kubikova et al. (2010)), and are innervated by a projection from the dopaminergic midbrain (Bottjer et al. (1989), Lewis et al. (1981)). Moreover, during song these neurons fire sparse, temporally precise bursts (Goldberg and Fee (2010)) similar to mammalian sparse firing MSNs (but see Singh Alvarado for discussion of X MSN variability). Transcriptomic analysis of Area X MSNs reveals that many MSNs express D1 and D5 DA receptor types, and this expression pattern correlates with FoxP2 expression; many other MSNs express D2 DA receptors but not D1/D5 (Xiao et al. (2021)). These profiles are suggestive of the basic distinction between direct and indirect pathway MSNs of mammalian striatum.

Other interneurons of the mammalian striatum appear represented by homologous cell classes in Area X. In particular, Chat+ interneurons in Area X (Zuschratter and Scheich (1990)) have intrinsic electrophysiological properties similar to Chat+ interneurons of mammalian striatum (Farries and Perkel (2002)). Moreover, tonically active neurons with *in vivo* characteristics similar to the cholinergic interneurons of mammalian striatum have been observed during singing (Goldberg and Fee (2010)). Parvalbumin positive Area X interneurons are aspiny and fast spiking (Farries and Perkel (2002)), exhibiting narrow action potentials and brief high-frequency bursts during singing (Goldberg and Fee (2010)), like the fast-spiking interneurons of mammalian striatum. Last, somatostatin positive interneurons in X (Xiao et al. (2021)) likely correspond to the low-threshold spike neurons observed slice (Farries and Perkel

(2002)) and song-selective bursting interneurons recorded during singing (Goldberg and Fee (2010)), consistent with a putative homology to the LTS/PLTS interneurons of mammalian striatum.

Area X contains mammalian pallidal cell type homologs

Area X projects to song motor thalamic nucleus DLM (Okuhata and Saito (1987), Luo and Perkel (1999a), Luo and Perkel (1999b), Luo et al. (2001)). Like mammalian GPi projections to motor thalamus, this projection is GABAergic (Luo and Perkel (1999a), Luo and Perkel (1999b)). The projectors express pallidal marker LANT6 (Reiner et al. (2004)) and are large with aspiny neurites (Farries et al. (2005)). The projectors exhibit high spontaneous firing rates in slice (Farries and Perkel (2002)) and fire rapidly without long pauses or bursting during song (Goldberg et al. (2010)). Although mammalian GPi neurons are inhibited by a long range projection from the striatum by MSNs, Area X projection neurons receive inhibition from local Area X MSNs (Farries et al. (2005)), which likely explains their polysynaptic inhibition by Area X's cortical inputs (Farries et al. (2005)). It remains unknown whether the direct MSN projection onto these GPi-like outputs is disproportionately D1R-expressing, in analogy to the mammalian direct pathway organization. A morphologically similar cell type in Area X was distinguished from the GPi-like projection neurons described above by the fact that cells of this second type synapse onto the GPi-like projection neurons but do not themselves project outside Area X (Farries et al. (2005)). These cells behave similarly to the projection neurons in slice preparations (Farries and Perkel (2002), Farries et al. (2005)), but can be distinguished by their firing pattern during singing. In particular, these fast-spiking aspiny non-projection neurons exhibit long pauses and burst firing during song (Goldberg et al. (2010)). These firing patterns during behavior are similar to the patterns of mammalian GPe neurons of the indirect pathway. Together with their propensity to

project onto the GPi-like cells of Area X, these neurons appear homologous to the pallidal cells of the indirect pathway. However, it is not known whether they receive disproportionate input from the D2-expressing Area X MSNs, a key feature of mammalian indirect pathway organization.

Reward prediction error and the dopamine afferent to Area X

These homologies between Area X and basal ganglia structures are significant because of our more detailed understanding that, in other systems, dopamine afferents to the striatum convey reward prediction error, and influence corticostriatal plasticity in ways that reinforce reward-maximizing policies. This homology suggests that dopaminergic reward prediction errors in Area X could control $HVC_{X_{MSN}}$ in ways that optimize the temporal context-to-perturbation policy. In adult acoustic error models (see 1.1.5), some auditory cortical neurons detect acoustic errors during singing (Keller and Hahnloser (2008)). This “error detection” response pattern is characteristic of auditory cortical neurons projecting to the dopaminergic midbrain (Mandelblat-Cerf et al. (2014)). Downstream, the dopaminergic projection from ventral tegmental area (VTA) to Area X exhibits firing rate modulations to these acoustic errors: after birds acclimate to repeated, stochastic perturbation of a target syllable, some VTA neurons, including VTA_X neurons, exhibit phasic inhibition in response to the perturbation. On the other hand, withholding the perturbation induces target time-locked phasic bursting in these same cells (Gadagkar et al. (2016)). The authors describe these responses as “performance errors,” explicitly analogizing them to classical reward prediction errors believed to support learning in other basal ganglia systems (Schultz et al. (1997)). To be explicit, the authors suggest that the bird predicts some probability of error. True error trials are worse than expected (probability of error resolves to 1) eliciting a negative performance error and dopaminergic cell inhibition. Escape trials are better than expected (error

probability resolves to 0), eliciting a positive performance error and dopaminergic cell bursting.

1.2.4 Policy updates and the AFP

Evidence reviewed in the previous sections suggests that the AFP implements a stochastic context-to-perturbation policy, and may receive rendition-by-rendition performance evaluations of action outcomes. These ingredients are necessary to adapt policies to maximize reward through reinforcement learning. In this section, I present evidence that adaptive AFP policy updates underlie song learning.

Necessity for learning

All the neural components that play a role in the AFP-based reinforcement learning theory are in fact essential to song learning. Juvenile LMAN lesions lead to production as an adult of unusually quiet singing, and simple song composed of a small number of repeating notes (Bottjer et al. (1984), Scharff and Nottebohm (1991)). Juvenile Area X lesions also prevent accurate song copying, however the adult song produced following juvenile Area X lesions is not simple and repeating. Instead, it is composed of “rambling,” variable notes that are often unusually long (Scharff and Nottebohm (1991)). Dysregulation of FoxP2 expression in juvenile Area X also leads to unusually variable adult singing (Haesler et al. (2007)). In adults, Area X is required for pitch learning (Ali et al. (2013)) and deafening-induced song degradation (Kojima et al. (2013)). Similarly Lesions of VTA_X cells or long-term blockade of Area X D1Rs impair juvenile learning outcomes (Hisey et al. (2018)) and impair adult pitch learning (Ali et al. (2013)).

Dopamine-perturbation correlations drive policy updates

In other systems, manipulation experiments reveal action-contingent DA release to cause operant conditioning (Tsai et al. (2009)). The comparison to other dopaminer-

gic prediction error signals invites the question whether Area X dopamine dynamics reinforce song variants. In fact, pitch-contingent syllable-targeted phasic activation of the terminals with optogenetics positively reinforces targeted pitches (Hisey et al. (2018), Xiao et al. (2018)), while similarly targeted phasic optogenetic inactivation of these terminals negatively reinforces targeted pitches (Xiao et al. (2018)).

Policies updated by acoustic error require AFP premotor activity to express

The results above suggest that dopamine action on D1Rs in Area X, under the dynamic control of acoustic feedback, critically underlies syllable modification in pitch learning through negative reinforcement of error trial variants and positive reinforcement of escape variants. No direct tests have explored whether Area X makes vocal “successes” more likely in a pitch learning paradigm. However, after birds have learned to avoid syllable pitches targeted by experimental acoustic perturbations, inactivation of LMAN blocks the expression of this learned bias in subsequent behavior; birds “revert” to singing pitches targeted by the perturbation (Andalman et al. (2009), Warren et al. (2011)). Expression of this adaptive premotor bias depends specifically on the projection from LMAN to RA (Warren et al. (2011)). Moreover, self-driven recovery to baseline from a pitch-shifted repertoire is accomplished by LMAN output biasing behavior back towards its baseline distribution (Warren et al. (2011)).

1.2.5 Consolidation outside the AFP

The previous sections present evidence that the AFP implements a stochastic context-to-perturbation policy that is adaptively updated in response to acoustic song feedback. This attractive model cannot fully explain policy learning in juvenile finches, however. In particular, well learned adult song can be produced by the PMP without AFP perturbations. The AFP learning model requires another mechanism to

incorporate AFP-dependent perturbations into the stereotyped policy implemented by the PMP. An adaptive AFP policy can be expected to induce novel correlations between RA action patterns and HVC_{RA} context patterns. These correlations could in principle drive Hebbian plasticity at HVC_{RA} synapses, eventually enabling HVC to causally induce those RA patterns with which it was previously only correlated. This “two-stage learning” can be accomplished by model neural circuits with appropriately tuned plasticity rules (Teşileanu et al. (2017)).

In fact, the average strength of existing HVC_{RA} synapses rises throughout development (Garst-Orozco et al. (2014)) indicating at least that this interface is dynamic during song learning. The most compelling tests of this consolidation model come from adult pitch learning experiments. In long-term pitch learning experiments, the pitch threshold defining perturbation “hits” and “escapes” is continually updated so that a fixed fraction of song renditions are targeted (say, the lower 75% of pitch variants) despite ongoing adaptive pitch modifications by the bird. These experiments can run for multiple days and induce the bird to sing pitches well outside his baseline range (Andalman et al. (2009)). In these experiments, the learned deviation from baseline acoustics can only be partly explained by AFP bias. In one report of multi-day pitch learning, AFP bias, measured as the difference in mean pitch produced with and without AFP influence, accounted for learning accomplished in the 24 hours prior to LMAN inactivation (Andalman et al. (2009)), suggesting that song changes depend on AFP premotor influence for around a day before those changes are consolidated and become AFP-independent. Another study found that the dependency of learned changes on AFP premotor activity was several days long, but nonetheless found that the learned changes were slowly consolidated outside the AFP. This study also found that the dependence of ‘recovered’ baseline-distributed singing after pitch learning on AFP premotor bias was temporary (Warren et al. (2011)).

1.2.6 Summary and current limitations

Studies in adult learning models have generated exciting mechanistic hypotheses about song system support for song learning, but the applicability of these mechanisms to juvenile song copying remains unclear. There are intrinsic differences between the learning problems. Adult pitch learning is guided by experimentally applied feedback conditioned on variation in the baseline behavior of the animal. Juveniles guide themselves without an obvious external reinforcer, and their target is often outside the range of variation they can produce. Relatedly, juveniles songs and tutor songs exist in a multidimensional acoustic space where many comparisons between the two are plausible. This increases the policy search space and plausibly complicates the feedback signal compared with the adult situation with a univariate meaningful control parameter (pitch) and a binary feedback signal (presence or absence of noise). The policy space may require intelligent search strategies to be efficient. Thus, the hypothesis formulated above strongly motivates testing its component ideas in juveniles. (See Appendix B for current work on this problem.) A major challenge to such testing — the reason pitch learning has been a popular substitute — is the complexity of juvenile syllable policies and the changes they undergo during development. The feasibility of testing these hypotheses as they relate to juvenile learning depends strongly on the ability to analyze and interpret juvenile vocalizations. I next briefly review the strengths and limitations of existing methods for juvenile song analysis. Then I present a novel approach called the variational autoencoder (VAE), which underlies the research in this thesis. I collaborated in the development of the vocal VAE, and in this thesis work I aim to validate its use as a descriptor of song copying, and to develop tools that leverage VAE descriptions to quantify juvenile developmental song changes.

1.3 Quantifying juvenile song behavior

Studying syllable modification requires a description juveniles' changing syllable policies, and a method to infer rendition-by-rendition vocal quality, that is, its reward value. Though pitch learning allows an experimenter to define these attributes, in juvenile learning we must infer them from an animal's spontaneous behavior instead. However, the high-dimensional and complex nature of vocal behavior, and corresponding complex changes during development, make the inference of important behavioral features challenging. Research in this area has produced useful results, but also has conceptual and interpretive limitations. More recently, an unsupervised learning technique called the variational autoencoder has emerged as an alternative that addresses some of these limitations (Goffinet et al. (2021), Kingma and Welling (2013), Rezende et al. (2014)). My thesis research applies this technique to the specific problems of song copying. Here I review influential analysis approaches including the autoencoder, and argue that the autoencoder may address limitations of prior methods.

1.3.1 Feature-based approaches

Early song behavioral research relied on qualitative descriptions of song spectrograms, as well as simple quantitative descriptions, for example syllable duration (e.g., Price (1979)). The development of Sound Analysis Pro (SAP; Tchernichovski et al. (2000)) significantly extended this quantitative approach, facilitating many novel observations. SAP automatically calculates a variety of spectral features from song audio, such as pitch, spectral entropy, frequency modulation, etc. Two key insights motivate the feature-based approach in SAP. First, complex interdependencies between the elements of a spectrogram make it difficult to quantitatively compare two raw spectrograms with linear techniques like correlations. Second, al-

though spectrograms themselves are high-dimensional and unwieldy representations, neural and peripheral production constraints imply that a much smaller number of carefully chosen features should capture the animals' behavioral variability. However, although SAP has generated many insights, some challenges arise from its reliance on experimenter-defined features. In general, it is difficult to assess whether an experimenter-defined feature is capturing significant amounts of variation in the underlying behavior. Including uninformative features decreases the ability to find meaningful behavioral patterns. SAP features are correlated (Goffinet et al. (2021)), so even a feature that captures behavioral variation may be redundant with other measurements, and therefore constitute an uninformative additional dimension. On the other hand, it is difficult to assess whether significant behavioral variation remains unaccounted for in an experimenter-chosen feature set. Overall, it is difficult to assess the relative and combined importance of hand-picked features, and decisions to include or discard features when characterizing a sound can be challenging to justify. Many SAP results are reported as patterns in single features picked out on a *post hoc* basis. In fact, despite SAP's prevalence, other experimenter-defined feature sets have been proposed and justified on account of their utility in distinguishing pupil/tutor animal pairs from arbitrary animal pairs (Mandelblat-Cerf and Fee (2014)). In addition, the SAP summary characterization of syllable renditions collapses temporal organization inside syllables. Syllables are represented as a mean and variance for each acoustic feature, even though the intrasyllabic sequence with which they visit different feature values is copied by pupils from their tutors.

1.3.2 Nearest neighbor age classification

Motivated by the limitations of SAP features, Kollmorgen et al. (2020) published an analysis of song development that does not depend on hand-picked acoustic features. Those authors calculate the median age at production time of every syllable's

nearest neighbors in raw spectrogram space. They use this “neighborhood time” to score the maturity of the query syllable. This approach has similarities to the independently developed predicted age score that will be presented in chapter 4 of this thesis. Both approaches leverage the developmental correlation between age and song acoustics to calculate a sort of maturity score from syllable sound. However, after discarding hand-picked features, the authors only measure neighborhood time, limiting insight into the multidimensional variability of song. To address this limitation, the authors use multidimensional scaling to embed points representing collections of songs at different ages, such that point proximity reflects the overlap in neighborhood time distributions across song collections. However, this embedding is difficult to interpret. It can be affected in similar ways by the locations and the widths of the neighborhood time distributions. We cannot conclude that translations in this space reflect translations in an underlying acoustic space. Additionally, this neighborhood time-based space also has no interpretation for single song renditions, limiting its applicability in experiments that make single-trial brain measurements or manipulations.

1.3.3 Variational autoencoders

Variational autoencoders are a method of unsupervised dimensionality reduction (Kingma and Welling (2013), Rezende et al. (2014)). In the context of song, they leverage the key insight of feature-based analysis approaches: large spectrograms are determined by relatively few parameters. To a first approximation, VAEs discover information-preserving, low-dimensional encodings of spectrograms. The preservation of information in this encoding is enforced by the requirement that the low-dimensional representation suffice to reconstruct the input spectrogram (albeit via a complicated neural network function). In this framework, possibly the most important experimenter decision relates to the evaluating the reconstruction because

this reconstruction loss function guides the machine learning objective. Here I use a simple and generic loss, the sum of squared pixel errors (Goffinet et al. (2021)). Thus, this analysis approach returns a multidimensional but tractable feature space that requires fewer experimenter assumptions than analyses based on hand-picked acoustic parameters. An in-depth presentation of VAEs is given in Appendix A.

1.3.4 *Summary*

Here I have reviewed influential prior methods for analyzing bird song, including the use of low-dimensional latent spaces learned by VAEs. This latter method shares the motivations and insights of SAP-based feature analysis, but in principle avoids the limitations that come from hand-picked features. At the same time, it provides more information and easier interpretation than the approach from Kollmorgen et al. (2020) that also avoids hand-picked features. However, VAEs have not been applied to the analysis of song copying. Moreover, even if autoencoders capture relevant behavioral variation, their feature space lacks a straightforward physical or biological interpretation. Thus we must develop suitable methods to relate autoencoder descriptions of sounds to the problems of song copying we originally identified: describing the syllable policy development, and assessing rendition-specific song quality. Work on these goals constitute the main research advances of this thesis, and is explained in chapters 2, 3, and 4.

1.4 General summary

Juvenile song learning models developmental skill acquisition, especially imitative learning. It is subserved by well defined forebrain circuits. In particular, research suggests a model of reinforcement learning and consolidation by the AFP and HVC_{RA} synapses, respectively. However, this model's predictions depend on abstract parameters of juvenile behavior, like the "reward quality" of a rendition, and changing

syllable policies in vocal motor space. Existing analysis techniques have limitations that have hindered our ability to transform abstract behavioral predictions into readily measurable and testable ones. The variational autoencoder is well positioned to address these limitations, but it has not been applied to song learning problems, so its suitability is experimentally untested. Moreover, realizing this technique's potential as a low-dimensional description of developmental syllable change requires tools to study changing latent distributions over time. Now I turn to my research, which aims to experimentally assess the suitability of this technique to studying song copying, and to develop analyses of autoencoder representations to describe developmental syllable change.

Variational Autoencoder Learns Copied Song Features

2.1 Introduction

Existing acoustic measures of zebra finch song are subject to a variety of limitations, as described previously in Section 1.3. For example, SAP (Tchernichovski et al. (2000)) calculates several features, but some, like pitch, are not applicable to all zebra finch song elements, some of which lack tonality. More generally, song analysis with hand-picked features requires experimenter decisions about which features to calculate and which to exclude. Mandelblat-Cerf and Fee (2014) argue that relevant acoustic features are those that adopt relatively diverse values across arbitrary collections of song sounds but adopt relatively similar values for tutor model sounds and their corresponding pupil copy sounds. This argument motivates evaluating whether pupil and tutor songs are more similar than arbitrary songs under the feature set learned by an autoencoder. I collaborated with Jack Goffinet (at the time, a research technician working with Dr. John Pearson and Dr. Richard Mooney) to make this evaluation. I collected the pupil and tutor song data in this chapter, and analysis

was subdivided such that I performed most of the “shotgun” song analysis and Jack performed most of the segmented “syllable” analysis. The figures are reproductions or modifications of figures appearing in a joint publication (Goffinet et al. (2021)). We found that the syllable and shotgun VAE approaches learn copying-relevant song features.

2.2 Results

2.2.1 Syllable VAE captures pupil/tutor similarity

First, we aimed to evaluate the suitability of VAE latent syllable representations for analyzing the quality of a pupil’s copy of its tutor song. Suitable syllable representations should capture the learned song features that are reliably transmitted through tutoring but also distinguish the song of pupils that were exposed to different tutors. To evaluate VAE latent syllable representations, we recorded song from adult male zebra finches and the tutors they were exposed during juvenile life (n=10 pupil/tutor pairs), and segmented their song motifs into syllables. An example motif from a single tutor and his pupil are depicted in Fig. 2.1A, along with example segmentation boundaries. We trained a single VAE over syllable renditions from all these animals (see methods). After model training, we calculated the 32-dimensional latent posterior mean for each syllable rendition and used this as our latent space representation of the syllable sound.

To visualize the relative locations of our latent vectors across syllable renditions, we embedded our latent descriptions for all syllables into 2 dimensions using the UMAP algorithm (McInnes et al. (2020)). These embedding values are depicted for syllable renditions from the example pupil/tutor pair in Fig. 2.1B, and for the entire population of syllables in Fig. 2.2A. The embedded points suggest that latent syllable representations form clusters, consistent with other work from our group (Goffinet et al. (2021)). These clusters correspond to different syllable types (for ex-

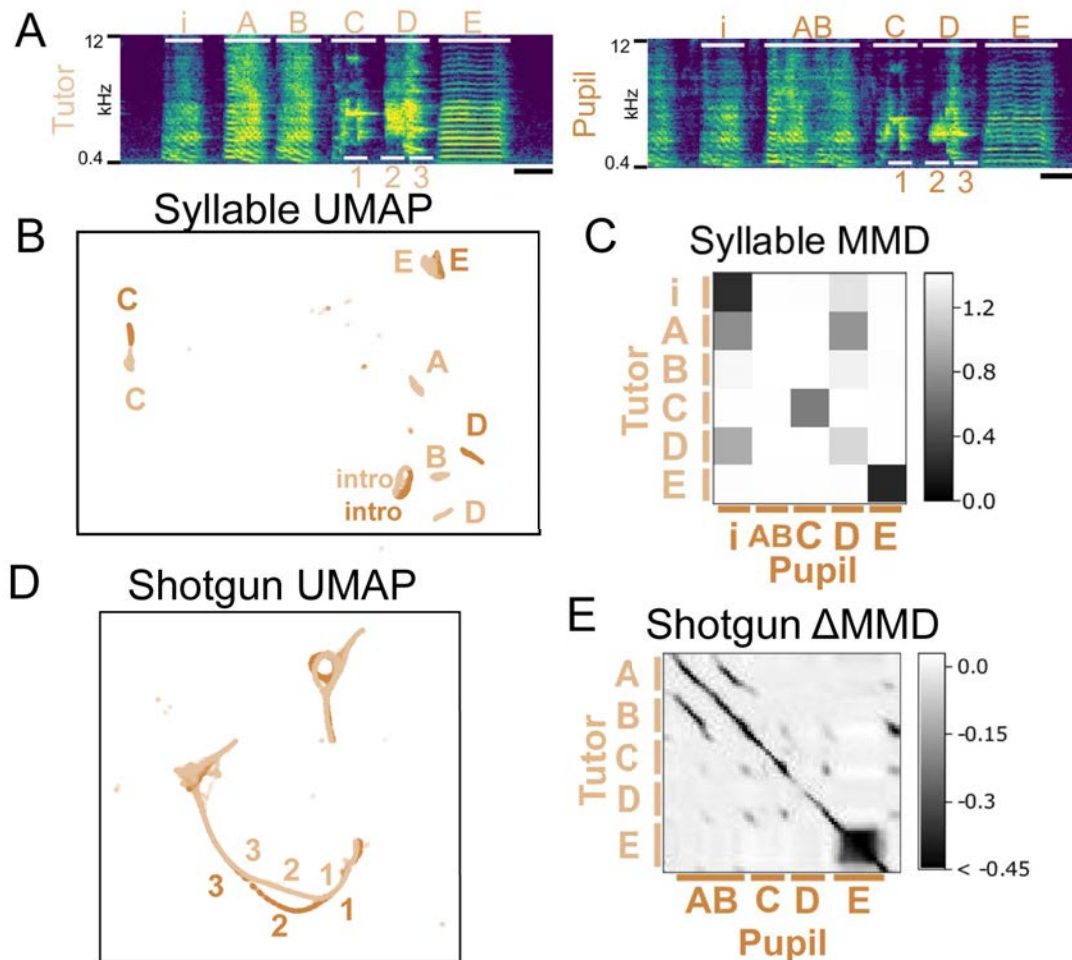


FIGURE 2.1: Similarity analysis of example tutor and pupil. **A** Spectrogram of a tutor motif (top) and adult pupil motif (bottom). Sound amplitude segmentation typically divided occurrences of these motifs into the syllables labeled above the spectrograms. These syllable segmentation decisions underlie the subsequent syllable-based VAE analyses we performed. Shotgun VAE analysis was based on 60ms windows, such as the windows labeled numerically under the spectrograms. Scale bar denotes 100ms. **B** UMAP embedding of the latent description of all syllable renditions from the two animals in A. Note that tutor syllables A and B were sometimes fused, leading to separate tutor clusters A, B, and AB. **C** Pairwise MMD between tutor and pupil syllables for these birds. **D** UMAP embedding of the latent description of all 60ms segments from these animals. Numbers are positioned near points generated by embedding the numbered windows in A. **E** Pairwise MMD between tutor segment collections and pupil segment collections.

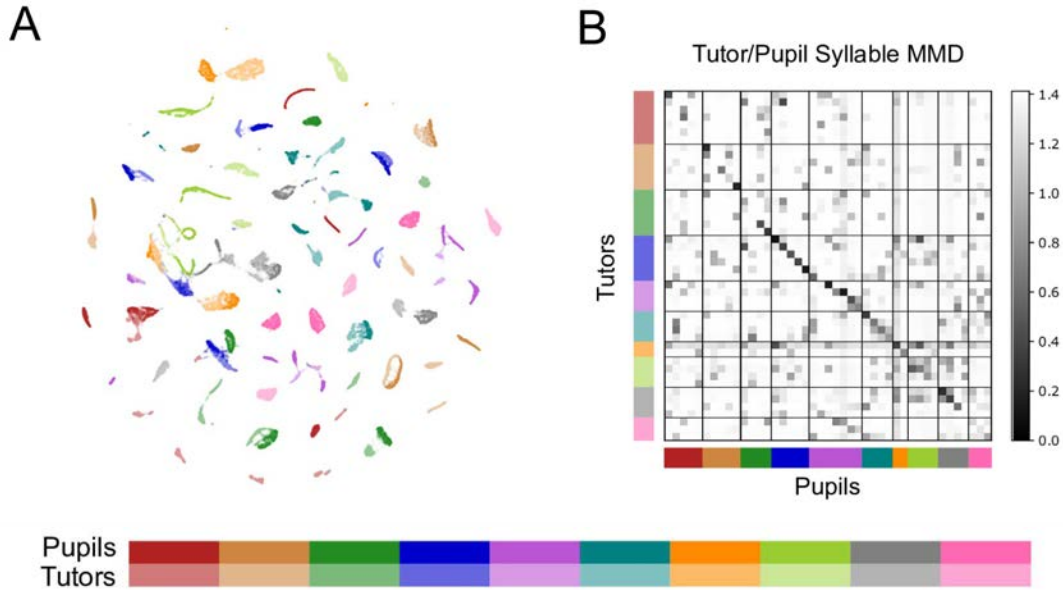


FIGURE 2.2: Syllable-based similarity analysis for all birds. **A** UMAP embedding of the latent representation of all syllables. **B** Pairwise MMD between every tutor syllable and every pupil syllable. Rows and columns are ordered first by bird identity. Next they are ordered by within-motif syllable order.

ample correspondences, see labels in Fig. 2.1A and B). We also observed that pupil syllable clusters often either overlapped or were near clusters from their corresponding tutors. These cluster pairs correspond sensibly to copied syllables, suggesting that the VAE found an underlying acoustic representation in which copied sounds are generally close to one another (for examples, see Fig. 2.1A). These visualizations suggest that in the learned latent feature space, tutor syllable clusters and corresponding copied pupil syllable clusters lie closer together than arbitrary syllable cluster pairs. Although the UMAP visualizations are broadly suggestive, the embedding into two dimensions distorts syllable relationships and does not precisely reflect similarity in latent space. To measure latent space similarity between syllable types quantitatively, we assigned labels to syllable renditions such that that renditions of the same syllable type (A vs B vs C, e.g.) from the same ani-

mal shared a label; otherwise renditions received different labels (see Methods). We then calculated a distance between categorical collections of syllable renditions using Maximum Mean Discrepancy (MMD, see methods; Gretton et al. (2012)). Two collections that mostly overlap in their acoustic distributions have an MMD near 0; two collections with mostly non-overlapping distributions generate larger MMD. In particular, we calculated the MMD between every pair of collections consisting of a tutor syllable collection and a pupil syllable collection. These comparisons are presented for collections from the example pupil/tutor pair in Fig. 2.1C, and for every pupil/tutor pair in Fig. 2.2B. In the example pair matrix (Fig. 2.1C), collection pairs consisting of a pupil syllable collection and tutor syllable collection that are situated in similar intramotif ordinal locations in the pupil song and the tutor song, respectively, lie near the matrix diagonal. These similarly located syllables are most likely to a tutor syllable and its corresponding copy syllable in the pupil song *a priori*. In the larger matrix (Fig. 2.2B), submatrices along the diagonal correspond to syllables from tutors and their respective pupils; these submatrices are arranged like the example matrix with diagonal entries depicting the MMD between syllables with similar intramotif ordinal positions in the tutor and pupil songs. The pattern of low MMD scores along the matrix diagonal indicates that syllables with an *a priori* expected copy relationship have nearby representations in VAE latent space.

2.2.2 Shotgun VAE captures pupil/tutor similarity

The syllable-based analysis above relies on the premise that syllable boundaries in pupil song correspond reliably to syllable boundaries in tutor song. However, pupils sometimes fail to copy their tutors’ syllable boundaries, as demonstrated by a missing syllable boundary in the example pupil compared with his tutor in Fig. 2.1A. Therefore, we sought to develop an analysis of song copying that was independent of syllable boundary definitions, and that could also examine copying at subsyllable

temporal resolution. To this end, we leveraged a “shotgun” VAE method previously developed by our group (Goffinet et al. (2021)). The VAE in this approach is identical to the one used for syllable analysis but training datasets are constructed differently. We identified song motifs, and sampled random onset times within motifs. The shotgun VAE was trained on 60ms duration song spectrogram windows with these random onset times. Thus the shotgun VAE learns to represent arbitrary song segments. To evaluate this approach, we trained a shotgun VAE over 60ms song windows from all birds (see Methods). From a hand-labeled subset of motifs, we generated sliding 60ms spectrogram windows, and calculated a latent feature vector for each such window.

To visualize the relationship between songs encoded in this way, we used a modified 2-dimensional UMAP embedding of the latent encodings. The modification encouraged embeddings that keep temporally adjacent windows from individual motif renditions near one another. The modification was naive to the relationships between different motif renditions and song from different animals (see methods; Fig. 2.4). These embeddings are depicted for the same example pupil/tutor pair described previously in Fig. 2.1D, and for all animals in Fig. 2.3A. The pupil and tutor song embeddings form distinct parallel “strands” that often overlap and occasionally diverge in the UMAP projection. The location of each point along a strand reflects the intramotif position of the sliding window that generated that point. Windows in the pupil motif that correspond through copying to windows in the tutor motif appear as overlapping pupil/tutor strands, as indicated by examining the underlying spectrograms from nearly overlapping points on pupil/tutor strands (Fig. 2.1A and D). By contrast, the embeddings of sounds from arbitrary animal pairs had no consistent relationships. They were generally non-overlapping and often quite far apart in the UMAP projection. Thus, both the modified UMAP and underlying shotgun latent vectors reflect learned acoustic features that relate pupil and tutor songs but

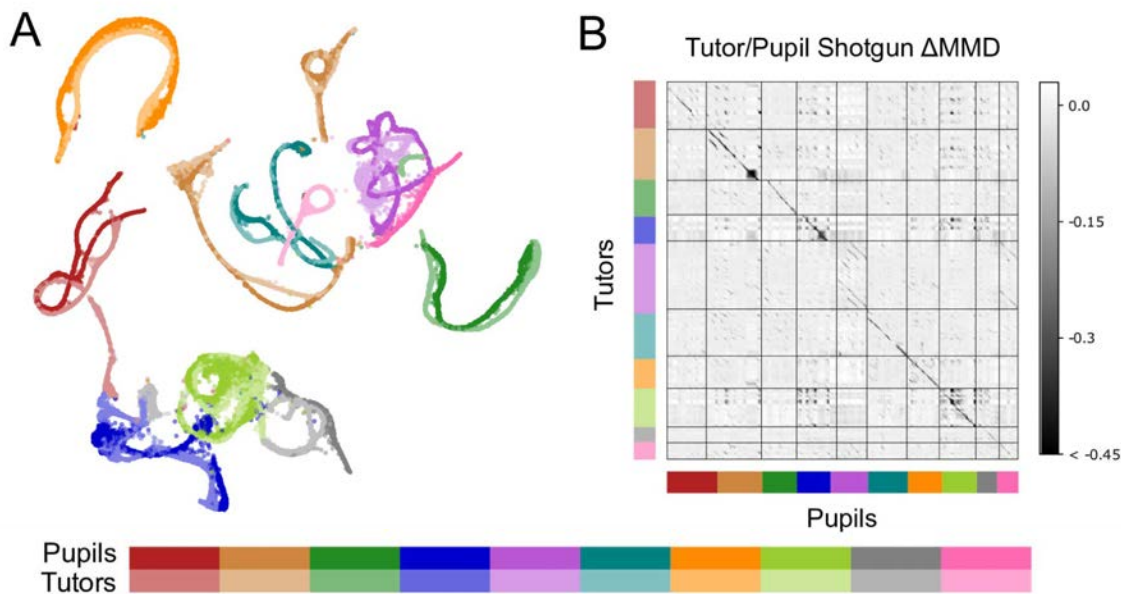


FIGURE 2.3: Shotgun similarity analysis for all birds. **A** UMAP embedding of all segments from all birds. **B** Pairwise MMD between all tutor segment collections and all pupil segment collections. Rows and columns are ordered first by bird identity. Next they are ordered by the within-motif timing of the segment collection.

distinguish arbitrary song pairs.

To quantitatively assess the tutor-pupil relationships in latent space generated using a shotgun VAE, we assigned labels to windows based on the bird that produced them and the window’s intra-motif position. This labeling procedure defines collections of windows: each collection contains sounds produced by a single animal and coming from similar intra-motif times. We calculated the MMD between latent points belonging to every collection pair consisting of a pupil collection and a tutor collection. These comparisons reveal a simple structure unrelated to copying. Some collections are relatively similar to almost all comparison collections; others are relatively different from almost all comparison collections (Fig. 2.5A). To highlight MMD structure that owes to copying, we examined the deviations from a rank-1 approximation of the MMD matrix (see methods and Fig. 2.5). These deviations

are distributed around 0 with a long negative tail corresponding to comparisons of collections that are much more similar than predicted by the simple rank 1 structure. In Fig.2.1E, a subset of MMD deviations from the rank 1 factor are depicted in a matrix with rows and columns sorted by the intramotif time of the window collection. The negative deviations near the matrix diagonal reflect the fact that similar pupil tutor sounds occur in a similar sequence in the two motifs. These deviations also explain patterns in the UMAP visualization. The MMD deviations corresponding to the start of pupil syllable D and tutor syllable D are less negative than elsewhere on the diagonal, suggesting that the shotgun VAE identified the start of syllable D as relatively poorly copied. The divergence between these animals' respective UMAP strands (point 2 in Fig. 2.2A and D) corresponds to the start of syllable D for both animals. In fact, this subsyllabic divergence can explain the relatively poor copying of syllable D quantified and visualized in the syllable-level analysis (Fig. 2.2B and C). Across all pupils and tutors, the on-diagonal, large-magnitude negative MMD deviations reflect the relative similarity of sequential sounds from pupil/tutor pairs, compared with other possible comparisons of sound collections (Fig. 2.3B). This result extends our observation that a syllable-based VAE identifies relevant, learned acoustic syllable features by showing that the shotgun VAE approach does not require syllable segmentation and can be used to quantify copying of subsyllabic structure.

2.3 Conclusions

As summarized in Section 1.3.3 and elaborated in Appendix A, variational autoencoders formalize the insight that variation in vocal behavior reflects a relatively small number of choices impacting acoustics. In many cases those choices are constrained by copying because pupils disproportionately make the same acoustic “choices” as their tutors after successful learning. In fact, previous work has emphasized the value of acoustic features along which pupils and tutors are more similar than arbitrary

animal pairs (Mandelblat-Cerf and Fee (2014)). Since the autoencoder method has not been applied to song learning analysis before, we tested whether training an autoencoder to encode sounds from many animals would find acoustic features with this desirable property. We also tested visualization and quantification methods to explore the autoencoder feature code with respect to cross-bird similarity. We found that, both with and without syllable-level segmentation, the autoencoder readily identified acoustic features according to which pupil and tutor songs were similar, but arbitrary animal song pairs were different. These results advance syllable and shotgun VAE methods as tools that can be used to map the entire process of juvenile song copying at single rendition resolution. Chapters 3 and 4 of this thesis develop and test methods for such a comprehensive and detailed account of sensorimotor learning in the zebra finch.

2.4 Methods

2.4.1 Recordings

We selected 10 adult, normally reared birds from different breeding cages in our colony. Until at least 60 dph, each of these pupil birds had interacted with only one adult male, the tutor from his respective breeding cage. We recorded the adult (>90 dph) vocalizations of these pupil birds for 5–12 days each with Sound Analysis Pro 2011.104 (Tchernichovski et al. (2000)), then recorded their respective tutors under the same conditions for 5–12 days each. Audio was recorded at 44.1 kHz.

2.4.2 Audio Segmenting

For approximately 10 min of song-rich audio per animal, we hand labeled song motif boundaries. These labels were used to train an automated segmentation tool, VAK 0.3.1 (Cohen et al. (2020b)), for each animal. Trained VAK models were used to automatically label motifs in the remaining audio data for each animal. Automatic

segmentation sometimes divided single motifs or joined multiple motifs. To correct for these errors, short (<50 ms) gaps inside motifs were eliminated. After this correction, putative motif segments with durations outside 0.4–1.5 s were discarded. Syllable segments were derived from a subset of the VAK motif segments by aligning the motif amplitude traces and manually determining syllable boundaries, resulting in 75,430 total syllable segments. To generate the shotgun VAE training set for Figure 7, 2000 vak-labeled motifs were selected from each animal. A single 60ms window was drawn from each motif to create a training set of 40,000 total segments across the 20-animal cohort.

2.4.3 Spectrograms

Audio segments were converted to time-frequency spectrograms using the Short Time Fourier Transform with Hann windows of length 512 and 256-sample overlap. These spectrograms were log transformed and interpolated to values at desired frequency and time points. 128 frequency points were mel-spaced between 0.4 and 8kHz. Short syllable spectrograms were scaled in time by $\sqrt{t_{\max}/t}$. Spectrogram values were clipped to hand-tuned maximum and minimum values and the resulting values scaled to lie in the interval $[0,1]$. Syllables shorter than t_{\max} were symmetrically zero-padded so that all spectrograms had 128 time steps and 128 frequency bins.

2.4.4 Model Training

The general variational autoencoder architecture and objective function is described in the introduction. Its specific architecture is given in Fig. 2.6. Separate models were trained for the syllable and shotgun analyses, over spectrogram training sets detailed above. In each case, the model was trained for 10 epochs.

2.4.5 Shotgun Test Data

After training a VAE over the shotgun training spectrograms, the hand-labeled motif segments used to train VAK models (see Audio segmenting) were segmented into overlapping 60 ms windows that spanned each motif with an 8 ms step size between successive windows (52,826 total windows).

2.4.6 Latent Representation of Sounds

We used VAE models trained according to the procedure above to analyze spectrograms from our test datasets. In particular, we calculated the 32-dimensional latent posterior mean for each spectrogram in our test dataset. This mean, called the latent vector, was treated as an acoustic feature vector describing the spectrogram.

2.4.7 UMAP

We used the python implementation of the UMAP algorithm (citation) to embed our test dataset collection of latent means in a two dimensional space while approximately preserving their local relative distances. In general, we used these parameter settings: `n_neighbors=20`, `min_dist=0.1`, `metric='euclidean'`. For the shotgun VAE embedding in this chapter, we modified this UMAP calculation as detailed below. Although the standard UMAP embedding of shotgun VAE latents from single-finch datasets generates points along smoothly varying strands (Goffinet et al. (2021)), UMAP typically broke motifs into multiple strand-like pieces in the 20-animal dataset from Fig. 2.3. To encourage embeddings that preserve the neighbor relationship of successive windows, we modified the distance measure underlying the UMAP. First, we computed the complete pairwise Euclidean distance matrix between all windows in latent space. Then, we artificially decreased the distance between successive windows from the same motif by multiplying corresponding distance matrix entries by 10^{-3} . This precomputed distance matrix was then passed to UMAP as a parameter.

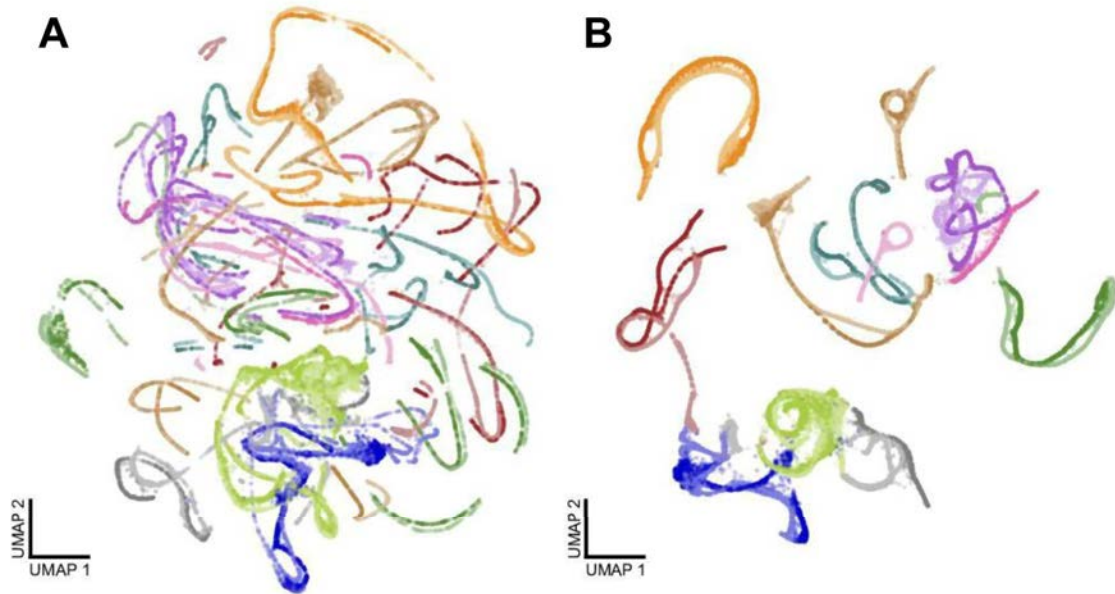


FIGURE 2.4: Effect of UMAP modification on motif “strands.” **A** Standard UMAP embedding of the latent representation of all song segments. Motifs from single animals often appear as multiple strands. **B** UMAP embedding after reducing the pairwise distance between adjacent segment within individual motif renditions. Motifs typically appear as unbroken strands.

See Fig. 2.4 for a comparison of the two UMAP projections.

2.4.8 Syllable Type Clustering

To prepare our data for syllable-level analyses, we assigned a syllable type category to every syllable rendition. To make these categorizations, we plotted the syllable rendition UMAP embeddings, one bird at a time. The clear clustering permitted drawing category boundaries around syllable groups by hand. These groups were then validated and labeled according to syllable order (that is, ABCD etc) by examining the spectrograms underlying several renditions per group in the context of the sound file in which they were produced.

2.4.9 Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) is a flexible measure of the difference between distributions P_x and P_y . In particular, given a flexible function class F , MMD is $\sup_{f \in F} \mathbb{E}_{x \sim P_x} f(x) - \mathbb{E}_{y \sim P_y} f(y)$. That is, MMD is the difference between expected values of a function f under P_x and P_y , for the f that maximizes this difference. Here, f is drawn from the class F of functions on the unit ball in a reproducing kernel Hilbert space with a fixed spherical Gaussian kernel. We chose kernel width equal to 25% of the median pairwise distance between latent points in the multibird test sample. The shotgun MMD matrix between pupil and tutor segments exhibits row and column “bands,” where a segment collection exhibits especially high or low similarity to many comparison collections. The underlying reason for this structure in the encoding is unclear. To focus on the sparse component of the matrix, where MMD entries are large or small compared to other entries in the corresponding row and column we decomposed the MMD matrix into a rank-1 and sparse component. In particular, we modeled the MMD matrix as the sum of a rank-1 matrix and laplace-distributed error matrix by minimizing the L1 error of a rank-1 approximation. This decomposition is depicted in Fig. 2.5.

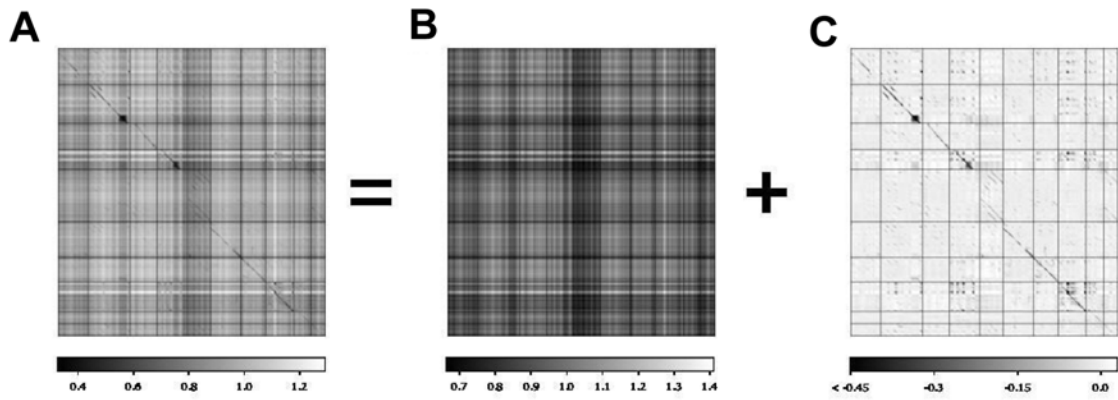


FIGURE 2.5: Shotgun MMD Decomposition into rank-1 and Laplace-distributed matrices. **A** Raw shotgun MMD matrix between tutor and pupil segment collections (rows and columns ordered as in Fig. 2.3B). **B** and **C** are the rank-1 and Laplace-distributed error matrix, respectively. The residuals in **C** isolate the effect of copying on similarity. This matrix is presented in Fig. 2.3B

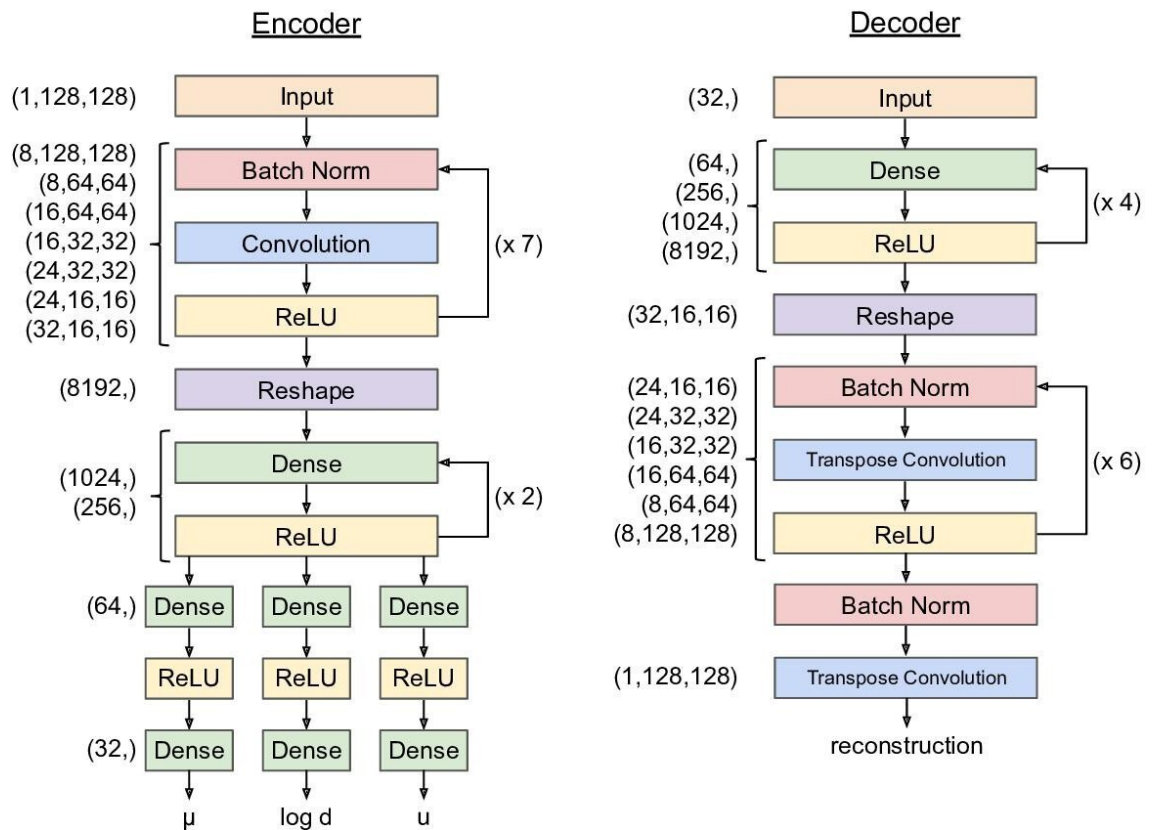


FIGURE 2.6: Autoencoder architecture. This autoencoder architecture was used for all analysis throughout the thesis. Note that arrow loops are shorthand for repeating layers in a feedforward layout, not recurrent connections.

Developmental Forward Models: Learning the dependence of acoustic distributions on age

3.1 Introduction

In the Chapter 2, I demonstrated that autoencoder methods to encode copied song features when trained on datasets that include adult pupils and corresponding tutors. These copied features emerge during juvenile development through practice, a process that has been difficult to study because of behavioral quantification challenges. In this chapter, I apply the autoencoder in the developmental context directly. I confirm that latent representations encode the developmentally changing parameters of song. I develop a flexible approach to model song acoustic development as a time-varying distribution in latent space, accounting for changes at multiple time-scales from hours to weeks. These models quantitatively describe developmentally changing syllable-specific policies. They reveal a circadian component to rendition-to-rendition variability, with variability high in the morning and low in the evening, consistent with a prior study (Ravbar et al. (2012)). I compare models fitted to developmental syllable data in normally tutored and in untutored birds, revealing

that song model isolation reduces this circadian developmental property even though song development in untutored juveniles also depends on auditory feedback (Konishi (1965)) and is thought to involve similar reinforcement learning processes as operate in tutored juveniles.

3.2 Results

3.2.1 Developmental changes in latent space reflect developmental spectrogram changes

By ~60dph, juveniles sing plastic song comprises several (2 to 5) readily distinguishable syllables. These are produced in accordance with policies that change through a process thought to depend on reinforcement learning. I first aimed to assess whether syllable VAE latent representations effectively describe these syllable modifications. Fig. 3.1A depicts the average spectrograms of syllable renditions sampled every 10 days from 60 to 90dph, a period that encompasses much of the syllable modification process. The highlighted spectrogram regions (a and b) are notable because the spectrogram averages reveal clear developmental changes. Region “a” becomes less entropic, with power becoming localized to two frequency bands by the end of syllable modification. Region “b” becomes more temporally complex during the same period. Originally produced as a frequency “downsweep,” by the end of development it begins as an upsweep and ends as a downsweep. Region “b” also develops a more complex harmonic structure across this developmental window. The latent representation of this syllable also undergoes systematic changes in this developmental window, as depicted in Fig. 3.1C. To assess the correspondence between these latent changes and average spectrogram changes, I calculated the average latent location of the syllable renditions whose spectrogram averages appear in Fig. 3.1. Using the trained VAE decoder, I generated spectrograms from these mean latent locations (Fig. 3.1B). As in the averages of real spectrograms, power in the generated spectrograms in region “a” becomes bimodal in frequency space. Similarly, over this

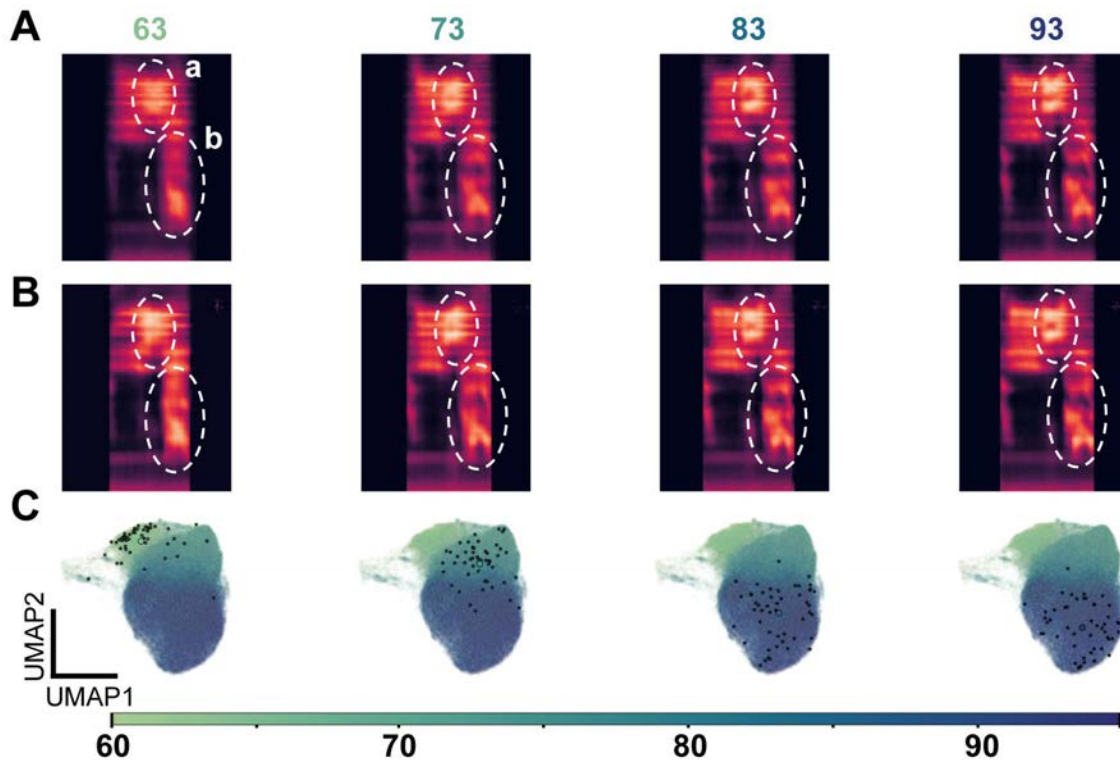


FIGURE 3.1: Example syllable development in spectrogram and latent space. **A** Average of spectrograms produced at 4 different ages in development. **B** VAE decoder reconstruction of the average latent location of the same renditions. **C** UMAP of all rendition latents, color coded by age at production time. The sampled renditions used to create average spectrograms in **A** are depicted as stars. Their average location used to for the reconstruction in **B** is depicted as a large circle.

same time window, region “b” of the generated spectrograms develops temporal and harmonic complexity similar to the average spectrograms. These correspondences show that developmental changes in the low-dimensional latent space can capture complex high-dimensional developmental changes in spectrogram space.

3.2.2 Gaussian models capture within- and between-day changes in latent distributions

The complex developmental changes in spectrogram syllable features are captured by simpler developmental changes in latent space. The simplicity of the latent space representation of development is demonstrated by the fact that 6 to 10 principal

components in latent space were typically sufficient to capture >99% of latent space variation for individual syllable types (see Fig. 3.4B). These results indicate the potential to describe syllable modification as a change in a low-dimensional (6 to 10 dimensional) latent space in place of high-dimensional (>16,000 dimensional) spectrogram space, without greatly reducing the descriptive power of our models. I next developed procedures to model the development of identified syllable types in latent space.

For a given syllable type, a primary goal is to track the ‘typical’ sound produced at different ages, given as a function from age to a location in acoustic (e.g., latent) space. Since the pace and quality of syllable modification is dictated by the pupil rather than the human experimenter, this function cannot be experimentally assigned and must be inferred from data. Moreover, the autoencoder representations do not impose obvious constraints on this function; they are plausibly consistent with complex, nonlinear mappings from age to latent location. (But see the final paragraph of Section 5.4.2 for a discussion of VAE extensions that could constrain this relationship.) In fact, in inspecting my datasets I observed many apparently non-linear developmental ‘trajectories’ through latent space. These considerations motivate the use of very flexible families of functions, like neural networks, to track the typical syllable type sounds produced at different ages.

Even in the best case, a time-varying point estimate of syllable acoustics leaves out critical features of juvenile singing behavior. To facilitate learning, the juvenile policy from intrasyllable times to acoustics should be stochastic (Fiete et al. (2007), Fee and Goldberg (2011)). In fact, rendition-to-rendition variation in the instantiation of this policy is actively maintained by the learning-relevant AFP (Ölveczky et al. (2005), Goldberg and Fee (2011)). Moreover, in adults policy updates in response to feedback may depend on features of this distribution that cannot be captured by a point estimate like a mean (Zhou et al. (2018)). These considerations motivate

modeling the age-dependent syllable policy probabilistically.

To satisfy these considerations, I modeled syllable development as a time-varying multivariate Gaussian distribution in autoencoder latent space. I instantiated this model as a map from production age to Gaussian parameters using a feedforward neural network, whose architecture is given in Fig. 3.2A. The network was trained to produce maximum likelihood estimates of the mean μ and full covariance Σ of a Gaussian distribution conditioned on age (see Gaussian Model Training in Methods).

The first two principal components of the latent representation of the example syllable from the previous section are reproduced in each subpanel of Fig. 3.2B, color coded by age at production time. Each subpanel highlights as stars a different subset of test data that was withheld during training of the Gaussian model network. In particular, data from 63dph, 73dph, 83dph, and 93dph is highlighted in the columns from left to right. Within these days, the top row highlights renditions within a half-hour window in the morning, and the bottom row highlights renditions within a half-hour window in the evening. Consistent with the impression from the age-colored points, the highlighted samples increase in PC1 value across the depicted month of development, but additionally the within-day samples apparently differ as well. To assess the behavior of the Gaussian model network, the center times of the half-hour windows of highlighted data (one center time per subpanel) were input to the trained network and the resulting Gaussian parameters were calculated. The 95% confidence ellipse corresponding to each Gaussian is depicted in its corresponding subpanel in Fig. 3.2B. We see that across days, and even within single days, the Gaussians calculated with the trained model appear to reflect the changing location and spread of test data.

I sought to quantify how age information at different time scales impacts model quality. For the example syllable, I calculated baseline model log likelihood using held out test data (Fig. 3.2C, green point). To quantify the overall contribution of age

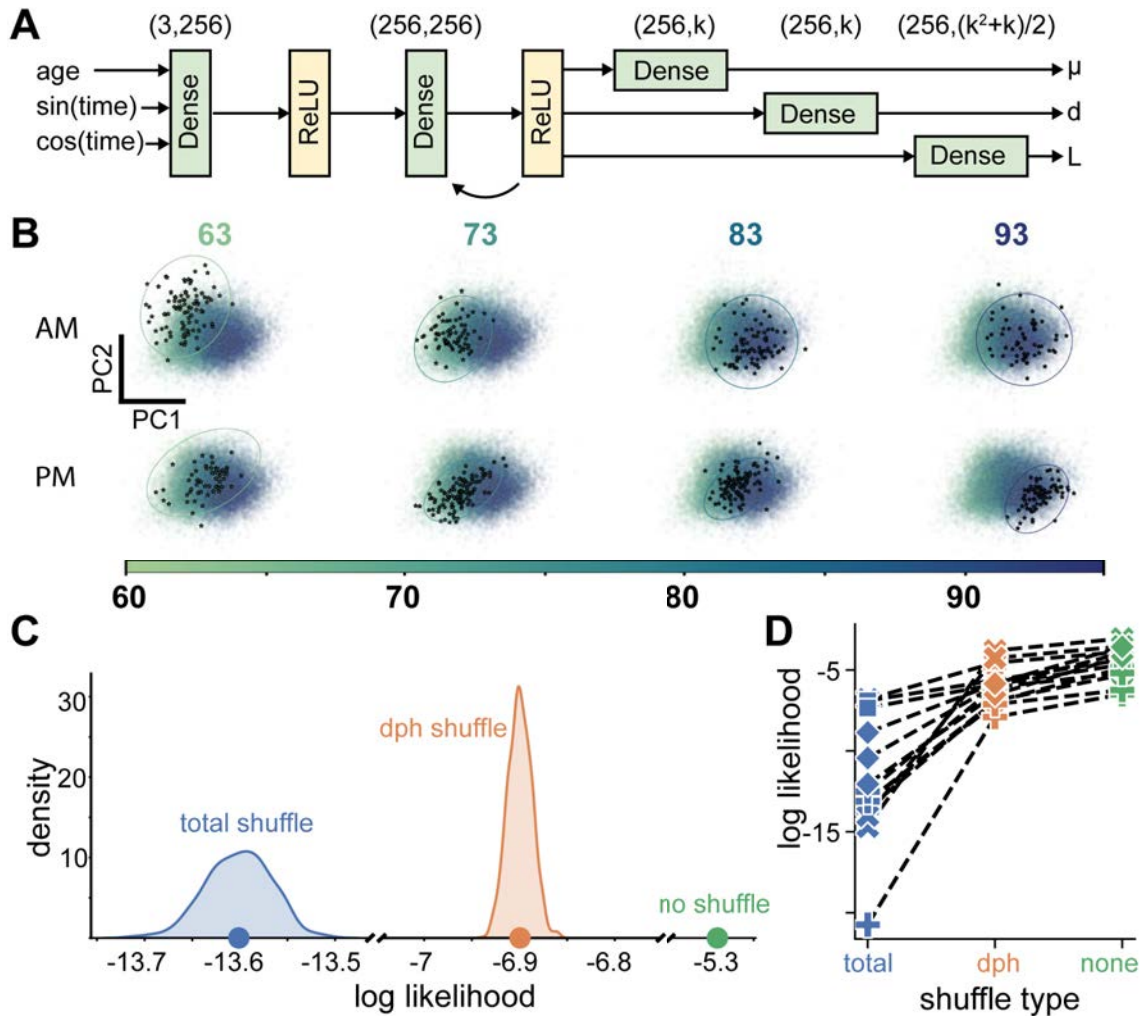


FIGURE 3.2: Design and performance of developmental Gaussian models. **A** Feed-forward network architecture of developmental Gaussian models. For detailed description of inputs and outputs, see methods. Curved arrow indicates repeated feed-forward layers, not recursive connectivity. **B** Example Gaussian model fits. Each panel emphasizes renditions sung within a half-hour window occurring within 63dph morning (top left), 63dph evening (bottom left), etc. The Gaussian predicted at the center time of each half-hour window is depicted as a 95% confidence ellipse in its corresponding panel. Highlighted points from a held-out test set were not used in training. Panel backgrounds contain the principal component projection of all example syllable renditions, colored by age at production time. **C** Log likelihood distribution and mean of panel B model, calculated with respect a test set with totally shuffled production times (1000 permutations), within-day shuffled production times (1000 permutations), and unshuffled production times. **D** Mean log likelihood of shuffling experiments for all syllables. Marker style indicates bird identity.

information in this model, I compared this baseline log likelihood to the log likelihood of models conditioned on shuffled age values (Fig. 3.2C, blue point). Finally, to quantify the specific contribution of variation in age at a within-day timescale, I compared the true log likelihood to a log likelihood of models conditioned on age values shuffled within day but without shuffling ages across days (Fig. 3.2C, orange point). This stratified shuffle preserves long time-scale (multi-day) correlations between acoustics and age, but eliminates short time-scale (within-day) correlations. I repeated this analysis over all syllables in the dataset (Fig. 3.2D). Shuffles that disregard all production time information profoundly disrupt model performance, indicating that the forward model captures the large differences in sound distributions between syllable renditions produced at very different ages. Within-day shuffling also impairs model performance, albeit more subtly, indicating that within-day model changes capture real, within-day changes in the distributions of syllable renditions.

3.2.3 Gaussian model entropy is high in the morning and low in the evening

Exploratory behavioral variability across repeated occurrences of a context plays a critical role in reinforcement learning generally. Moreover, juvenile song variation depends on specific, learning-related regions in the AFP, suggesting this variation is centrally controlled in order to facilitate song learning. It is recognized that juvenile produce more stereotyped behavior as they mature over multi-week time scales, and neural mechanisms have been proposed to facilitate this change (Ölveczky et al. (2011), Garst-Orozco et al. (2014)). But birds can rapidly alter the variability of their behavior in response to social cues (Kao et al. (2005), Kojima and Doupe (2011)), so it is possible that the variability of syllable policy is controlled at relatively rapid (<1 day) timescales during learning. In fact, prior studies report within-day changes in song variability but draw different conclusions about whether it rises or falls within the day (Miller et al. (2010) vs Ravbar et al. (2012)). Moreover, prior work has

focused on learning-related changes at the timescale of single days (Derégnaucourt et al. (2013), Kollmorgen et al. (2020)) and improved description of the behavior at that timescale could help to resolve conflicts between these results.

The Gaussian models I develop above are the first attempt to characterize the overall, age-dependent covariance structure of juvenile syllable policies. I sought to leverage this innovation to determine how the overall amount of rendition-to-rendition variability changes across time, particularly at the timescale of individual days. I queried the entropy of the fitted models at regular intervals, and discarded query times that were poorly covered by training data (see Methods). These entropies are presented over a range of ages for an example syllable in Fig. 3.3A. I observed a daily pattern in the entropy of syllable distributions, with syllable distributions exhibiting relatively high entropy early in the day and relatively low entropy at the end of the day. To quantitatively assess this pattern across multiple syllables and animals, I constructed a model of entropy as a linear function of time of day, and included random effects components in the intercept and slope to account for the possibility that individual birds and syllables exhibit correlated deviations from the group average behavior (see Methods). In Fig. 3.3B, I depict this model’s predictions of entropy by time of day, incorporating the random bird and syllable effects. While successfully accounting for differences between syllables, the model identifies a circadian entropy decrease in every case. Not surprisingly then, the fixed effect indicated the same circadian trend (entropy decreases by $0.073275/\text{hr}$, $p < 10^{-8}$). These individual fits and the partial dependence of entropy on time of day after marginalizing out random effects of bird and syllable are depicted in Fig. 3.3C.

3.2.4 Gaussian models of isolate song

Even juveniles raised in social isolation from a tutor use auditory feedback to learn some species-appropriate song characteristics during development (Konishi (1965)).

Indeed, such social isolates are speculated to engage many of the same behavioral and neural mechanisms to “learn” their isolate songs. Nonetheless, the acoustic development of isolate song remains largely uncharacterized. Consequently, we do not know if the circadian patterns I just described that are exhibited by normally tutored animals depend specifically on tutoring, or instead reflect other aspects of sensorimotor experience that accompany juvenile singing, such as vocal motor activity and associated auditory feedback. Therefore, I sought to characterize the entropy of isolate syllable distributions from juvenile birds raised in isolation from a tutor, and compare these measurements to those I made in normally tutored animals. I recorded isolate song development and processed these audio data according to the procedure developed for tutored animals (see methods).

As with tutored syllables, I modeled syllable-level distributions over time as multivariate Gaussians and then calculated these models’ entropy over time. The entropy dependence on time of day was compared for isolates and tutored animals by adding a tutoring main effect term and a tutoring-by-time of day interaction term to the fixed effects design of the linear mixed effects model described earlier (see Methods). In general, the models of isolate syllables had greater entropy than models of tutored syllables (tutoring coefficient = -2.3153, $p < 0.001$). This increased entropy may owe to the generally increased dimensionality of isolate syllable variation (see Fig. 3.4B; Methods). In addition, the entropy of isolate syllable distributions depended less on time of day (difference in slopes = -0.0683/hr, $p < 0.05$). Therefore, song development in juveniles raised without tutor experience differs from that in tutored animals in two important ways: isolate syllable distributions are more entropic and also show more modest circadian fluctuations in entropy. These findings are consistent with the idea that tutoring exerts identifiable effects on the sensorimotor learning process that are difficult to explain exclusively through tutoring’s widely recognized effect on parameterizing the internal template.

3.3 Conclusions

Having previously established the ability of autoencoder-based methods to identify copied syllable features, in this chapter I developed and tested quantitative models of the age-sensitive distributions of syllables in an autoencoder-based latent space. In particular, I trained a neural network map from production age to underlying Gaussians in latent space that are approximate generative models of syllable renditions. These models leverage both multi-day and within-day developmental trends to predict the acoustic distributions of syllables.

Trained models of tutored syllable distributions reliably exhibit high entropy in the morning and low entropy in the evening. This pattern suggests that syllable rendition-to-rendition variability is high in the morning and low in the evening. Such a pattern superficially conflicts with the conclusions of a prior study (Miller et al. (2010)) that does not distinguish within- and between-rendition acoustic variability, but is consistent with a prior investigation isolating rendition-to-rendition variability (Ravbar et al. (2012)). More broadly, this result underscores a theoretical gap in models of song development. Some existing learning models emphasize how mean acoustic output can be rapidly adapted (Andalman et al. (2009), Fee and Goldberg (2011)), while others emphasize slow changes in variability at long time-scales (Garst-Orozco et al. (2014)). Rapid adaptation of rendition-to-rendition variability has received less attention in a learning context, even though learning-related forebrain structures in the AFP rapidly adapt vocal variability to varying social context (Kao et al. (2005)). Future work can establish whether rapid regulation of variability directly relates to the learning algorithms that are actually employed by juvenile songbirds during syllable modification. Plausibly, within-day variability changes directly reflect an error-correcting learning process, like the elimination of undesirable repertoire variants.

Isolate song undergoes developmental changes that, like those in tutored song development, depend on auditory feedback (Konishi (1965)). Nonetheless we do not know if isolates exhibit similar behavioral patterns as a consequence of this shared developmental mechanism. To address this gap, I modeled isolate song sound distributions using the procedure developed to model the distribution of normally tutored juvenile syllables. Overall, isolate sound distributions had higher entropy than distributions of developing tutored syllables, which may owe in part to the increased number of linearly independent dimensions of isolate sound variation. Plausibly, a specific tutor model may constrain pupil output more than innate targets governing isolate development. In addition to its main effect on entropy, isolation reduces the dependence of entropy on time of day. On its face, this result implies that singing-auditory feedback-dependent juvenile song development is not a sufficient condition for the entropy dynamics observed in tutored birds. This observation can constrain future theoretical accounts of the relation between juvenile learning and the dynamic regulation of rendition-to-rendition variability, accounts that are currently lacking in the field as noted above. However, I also cannot rule out the possibility that the Gaussian modeling procedure is less sensitive to rapid changes in isolate distribution entropy. This possibility could arise if the isolate syllable collections I labeled for modeling are less Gaussian – and therefore more poorly modeled using the procedures in this chapter – than the syllable collections labeled in the tutored case. Future work with more flexible models, such as density mixture models, can relax the Gaussian assumptions here to resolve this issue.

The models developed in this chapter provide quantitative descriptions of changing syllable policies, though they are agnostic to the reinforcement learning concepts that are theorized to underlie these changes. Reinforcement learning requires actions to be evaluated, ultimately, with respect their effect on a one-dimensional ‘reward’ quantity, but we have not developed a method to evaluate the quality of syllable

renditions. In the Chapter 4, I develop a method to quantify individual syllable rendition quality.

3.4 Methods

3.4.1 Recordings

3 of the tutored birds in this dataset were raised in their home cages with their parents and clutch mates until approximately 50dph, at which time they were singly housed in boxes equipped with microphones and recorded at 41kHz using SAP (Tchernichovski et al. (2000)). 2 additional tutored birds in the dataset were recorded from 72dph to adulthood after rearing in their home cages. These were recorded at 32kHz using EvTaf. Isolate birds were transferred with their clutchmates and mother from the primary aviary to sound-attenuating boxes that blocked outside adult male sounds before 10dph. They were raised in this environment until 30 to 40dph, at which time males were housed singly in boxes equipped with microphones. They were then recorded until adulthood at 41kHz using SAP.

3.4.2 VAE latent scoring

VAE processing was performed separately for each bird in this dataset. I generated syllable spectrograms from the audio recordings by manually tuning sound amplitude thresholding parameters to segment audio into syllables, and then manually tuning spectrogram floor and ceiling values to a range that captured variation in vocal sound intensity but excluded quiet background noise. These clipped syllable spectrograms were rescaled so all values fell in the interval $[0,1]$. These spectrograms were used to train bird-specific syllable VAEs according to the procedure outlined in Chapter 2, “Model Training.” Finally, the resulting VAE was used to calculate a latent representation of every syllable in the dataset by calculating the mean of the latent posterior given by the trained VAE encoder.

3.4.3 *Preparing syllable type datasets*

After calculating a latent representation for every sound in the dataset, the latents were embedded in a 2-dimensional UMAP to visualize clusters. By investigating the underlying spectrograms of renditions in each cluster, I was able to assign meaningful category labels to different clusters. Some categories (like cage noise and call types) were discarded. I retained for further analysis only clear clusters corresponding to syllable types represented in the animal’s crystallized endpoint song. In some cases of <60dph song, I observed *in situ* differentiation of a unimodal syllable distribution into two distinct syllable categories. For all analysis in this chapter, I focused on song produced after 60dph after which time syllable number was stable. I also limited analysis to song produced before 95dph. In the case of isolate data, in addition to well-defined clusters, I often observed large, complex groups of sounds that included song elements and calls. Sounds in these complex groups were discarded unless the song elements were localized to protruding ‘peninsulas’ in the UMAP. In these cases, a coarse categorization was made, followed by manual categorization of sounds near the category boundaries. One of 5 isolates exhibited no apparent song sound clusters at all, and was dropped from further analysis. After assigning syllable labels in this way, I performed within-syllable PCA to find the primary axes of variation exhibited by syllables over the course of development. Finally, the data were further subdivided. 80% of renditions of each syllable type were randomly assigned to a training dataset, and the remaining 20% were assigned to a test dataset. The training dataset was used to train Gaussian model networks as explained in the next section.

3.4.4 *Gaussian model training*

Training datasets were used to train Gaussian model networks for each syllable type. Observations consisted of pairs: production time and the vector of k principal components, with k chosen to cumulatively explain >99% of the variation in latent data.

Note that isolate syllable data typically required greater k than tutored syllable data (see Fig. 3.4B). The training data was further subdivided into a train, test, and validation sets according to a 70/20/10 split. These were used for weight updating, training stopping, and evaluation respectively. The architecture for these Gaussian model networks is given in Fig. 3.2A. The model input “age” was generated from observed data by z-scoring all production ages. The $\sin(\text{time})$ and $\cos(\text{time})$ inputs were generated from production time of day by calculating the sine and cosine, respectively, of production time of day in radians ($24 \text{ hrs} = 2\pi$). Thus, the quantity $[\sin(\text{time}), \cos(\text{time})]$ took unique values at every time of day, and identical values for renditions at the same time of day on different days. The output values of the network were used to construct the parameters of a multivariate Gaussian distribution with full covariance matrix Σ of dimension k . Output l is reshaped into a lower triangular matrix L reflecting the Cholesky decomposition of Σ : $\Sigma = LL^T + \text{diag}(e^d + \epsilon)$. μ is directly interpreted as the multivariate Gaussian mean. Given training examples of paired production times and latent locations, the network learned to minimize the negative log likelihood given latent observations: $\sum_i -\log(P(\text{latent}_i))$, $P \sim N(\mu, \Sigma)$. To minimize this loss with respect to the network parameters, I used the Adam optimizer with learning rate = 0.001. The model was trained until it experienced 10 consecutive epochs without improvement on the test set, at which point the model producing the best test set performance was saved and subsequently used in analysis.

3.4.5 Evaluation of Gaussian models

I scored the performance of the trained Gaussian models using held out syllable rendition data. First, I calculated the log likelihood of the trained model by evaluating $\frac{1}{n} \sum_i -\log(P(\text{latent}_i | \text{age}_i))$. To evaluate the contribution of age information to the performance of each model, I generated 1000 permutations of the age parameter and computed for each permutation the “total shuffle” log likelihood

$\frac{1}{n} \sum_i^n -\log(P(\text{latent}_i | \text{age}_{\text{total shuffle index}}))$. I summarized this experiment for each syllable with the mean of the total shuffle log likelihood. To separate the contribution of within- and between-day age information to the performance of each model, I generated 1000 within-day permutations of the age parameter. That is, permutations that associated latent observations with wrong-day production times were prohibited. For each permutation I computed the “dph shuffle” log likelihood $\frac{1}{n} \sum_i^n -\log(P(\text{latent}_i | \text{age}_{\text{dph shuffle index}}))$. I summarized this experiment for each syllable with the mean of the dph shuffle log likelihood.

3.4.6 Entropy datasets

The trained Gaussian models map production ages to multivariate Gaussian distributions. The volume of these distributions is summarized by their entropy: $\frac{1}{2} \ln |\Sigma| + \frac{k}{2}(1 + \ln(2\pi))$, where k is the number of dimensions. In order to evaluate the behavior of the fitted models I generated ‘query times’ at 5 minute intervals during daytime hours of birds’ light cycles. Because the model output is not reliable at times without nearby training data, I then discarded query times at which fewer than 30 training renditions of the modeled syllable were produced in the half hour window centered at the query time. At the remaining query times, I generated predicted covariance matrices and calculated entropy from these.

3.4.7 Linear mixed effects models

The entropy models corresponding to the syllables of different birds, or to a bird’s different syllables, may differ from one another in idiosyncratic ways. One important feature that differs by syllable is the number of dimensions, k , used to represent syllable rendition acoustics. In general, entropy scales with distribution dimension, so the differences in k between syllables almost certainly leads to entropy variation unrelated to a dependence on production time. To account for idiosyncratic features

exhibited by syllables and by birds, I constructed a linear mixed effects model for the dependence of entropy on time of day: $entropy = X\beta + Z_{bird}b_{bird} + Z_{syll}b_{syll} + \epsilon$, where

- X is a fixed-effects design matrix containing a column of ones for the intercept, and a column of time-of-day values for every query time,
- β is the fixed-effect intercept and time-of-day slope,
- Z_{bird} is a random-effects design matrix containing an indicator column for each bird ID, and a time of day * indicator column for each bird,
- b_{bird} is a vector of bird-specific random intercepts time-of-day slopes,
- Z_{syll} is a random-effects design matrix containing an indicator column for each syllable ID, and a time of day * indicator column for each syllable,
- b_{syll} is a vector of syllable-specific random intercepts and time-of-day slopes,
- ϵ is a vector of random errors.

The bird-level random effects vector b_{bird} is populated with intercept-slope pairs $(b_{0,i}, b_{1,i})$ for each bird i distributed according to the 2D Gaussian $\mathcal{N}(0, \Theta)$ estimated from the data. Similarly, the syllable-level random effects vector b_{syll} is populated with intercept-slope pairs $(b_{0,j}, b_{1,j})$ for each syllable j according to $\mathcal{N}(0, \Phi)$ estimated from the data. The entire model is fit using Matlab's `fitlme` function and providing the Wilkinson notation formula 'entropy ~ 1 + timeOfDay + (1 + timeOfDay | bird) + (1 + timeOfDay | syllable).'

To compare isolate and tutored syllable distributions, I fit a similar model that included 2 additional fixed effects in β : a binary regressor set to 1 for models fit over tutored bird syllable and 0 otherwise; and an interaction between this tutoring variable and the timeOfDay regressor.

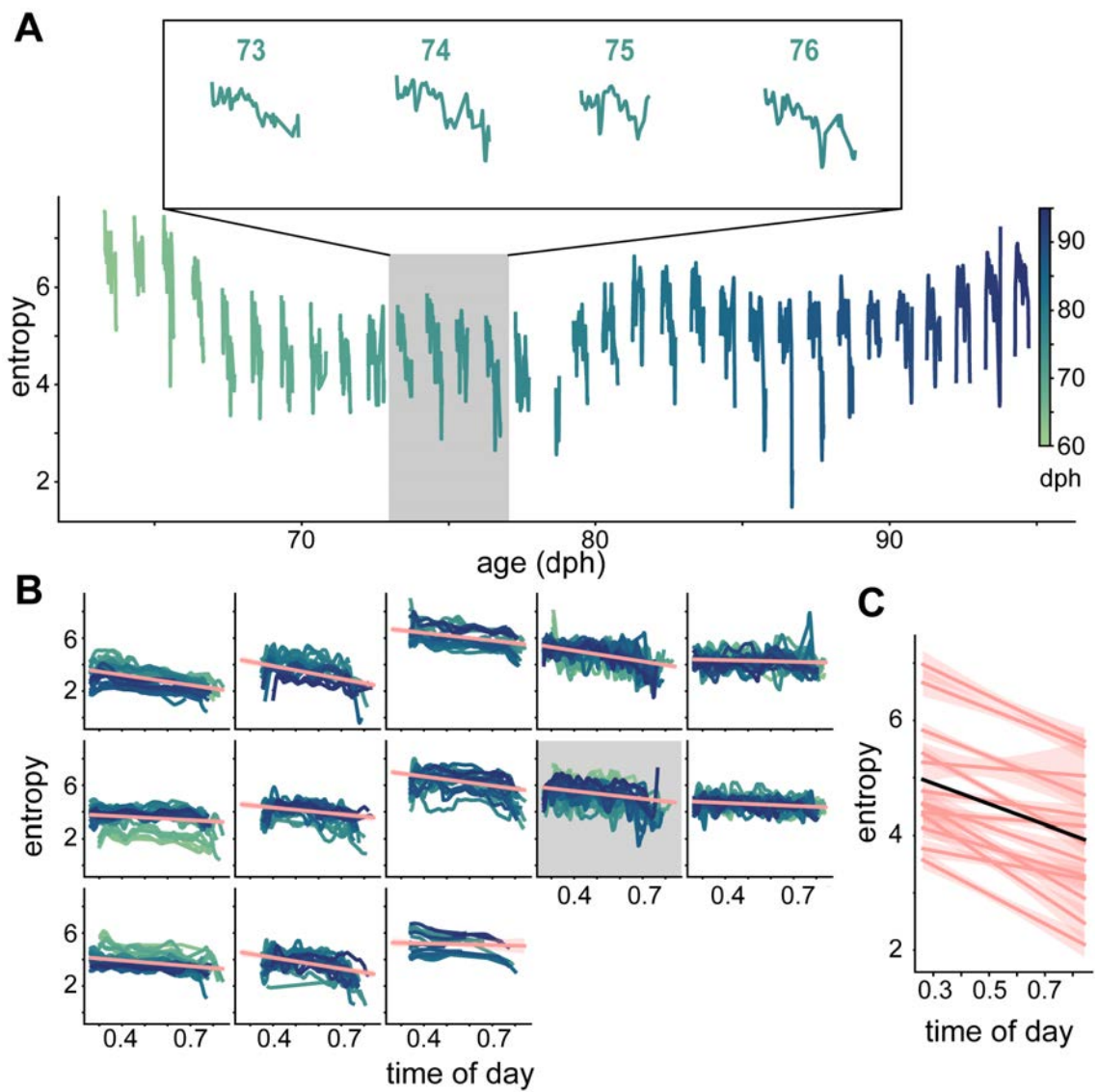


FIGURE 3.3: Daily entropy variation. **A** Gaussian model entropy versus age for the example syllable. Inset depicts example days at higher temporal resolution. **B** Multiple days

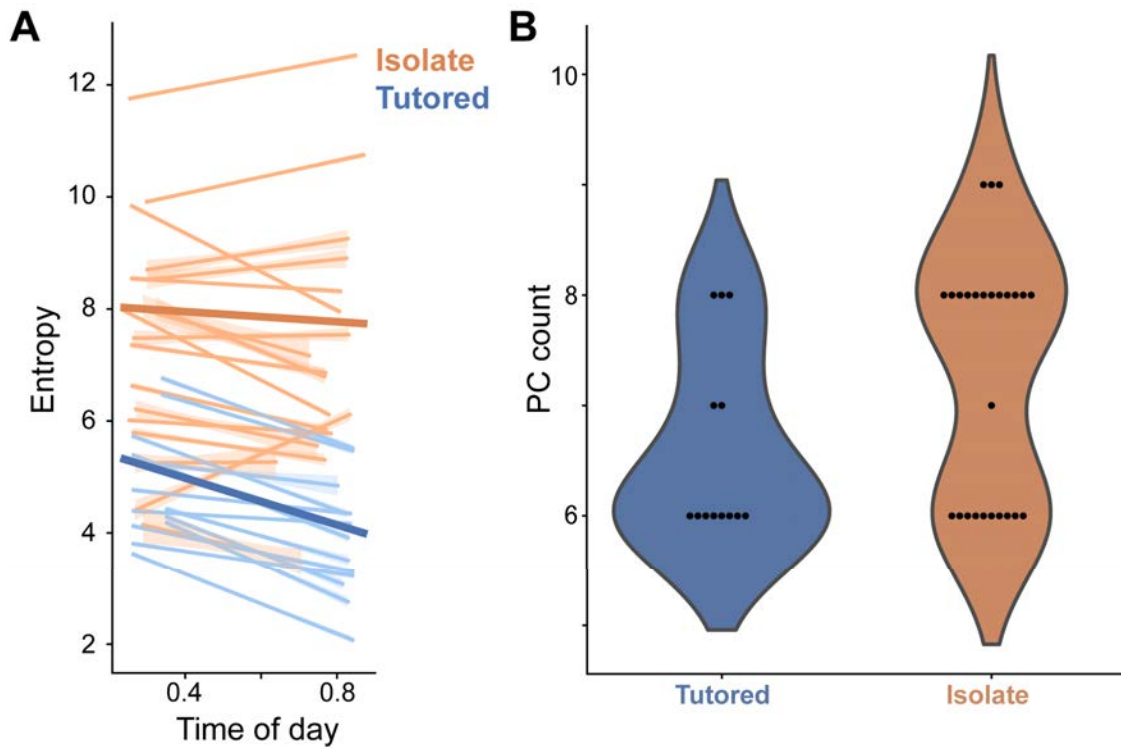


FIGURE 3.4: Isolate and tutored circadian entropy. **A** Entropy was modeled as a linear function of time of day, with random slope and intercept components added at a bird and syllable level (see methods). Syllable-level predictions and 95% confidence intervals are thin lines; tutoring group-level partial dependence on time of day is presented as thick lines. **B** More PCs were required to explain 99% of variance of isolate syllables than in tutored syllables ($p < 0.05$, t-test with unequal variance).

Reverse Models: Scoring individual rendition maturity

4.1 Introduction

In the last chapter, I developed methods to model the age-dependent acoustics of plastic song syllables. These acoustic changes are thought to depend in part on an evaluation of individual song renditions that subsequently facilitates reinforcement learning: activity producing renditions evaluated to be ‘good’ is positively reinforced and vice versa (eg, Fiete et al. (2007), Fee and Goldberg (2011)). Song quality in this model arises in the brain and cannot be determined definitively on the basis of behavior alone. However, if song quality controls developmental change through reinforcement learning, then we can make inferences about song quality from the statistics of the behavioral timecourse. In particular, to the extent that song changes arise from song evaluation and downstream reinforcement, we can infer that sounds that are produced early and become infrequent are relatively ‘bad,’ generating negative reinforcement. Conversely, we can infer that sounds that become more frequent with practice are relatively ‘good,’ generating positive reinforcement. In this chapter, I

develop this logic to create an explicit quantitative metric of the quality of individual syllable renditions by estimating the age at which each was most likely produced. I leverage data collected by another graduate student, Jiaxuan Qi, to show that a mechanism underlying putative reinforcement learning in juveniles is required for the daily improvements in song quality captured by this metric. I relate this metric to previously developed measures of syllable maturity, showing that it exhibits circadian patterns comparable to patterns seen with other measures. Finally, I show that isolation from a tutor model alters these circadian patterns.

4.2 Results

4.2.1 *Scoring syllable rendition maturity by predicting age*

In the last section, I modeled the changes in probability of producing sounds with different acoustics as birds practice. This developmental trajectory induces a complex joint distribution between the acoustic features of syllable renditions and production ages. That joint distribution induces a conditional distribution of the form $P(\text{age} \mid \text{latent})$. The mean of this conditional distribution is a function over latent space that minimizes the square deviation from actual syllable production age given syllable acoustics. I trained a neural network to estimate this function. The architecture of the network I used is given in Fig. 4.1A. (Training details are given in methods.) In the following, I refer to this network’s output as a syllable’s predicted age, given its acoustics. To assess the performance of this network, I compared actual syllable production ages to predicted ages on renditions held out from model training. The left panel of Fig. 4.1B shows the UMAP embedding of these held out renditions of an example syllable, color-coded by the age at which each was produced. The spatial organization of production ages conceptually reflects the function that the neural network is instructed to learn. The right panel of Fig. 4.1B shows the UMAP embedding of the same held-out dataset, color-coded instead by the predicted age

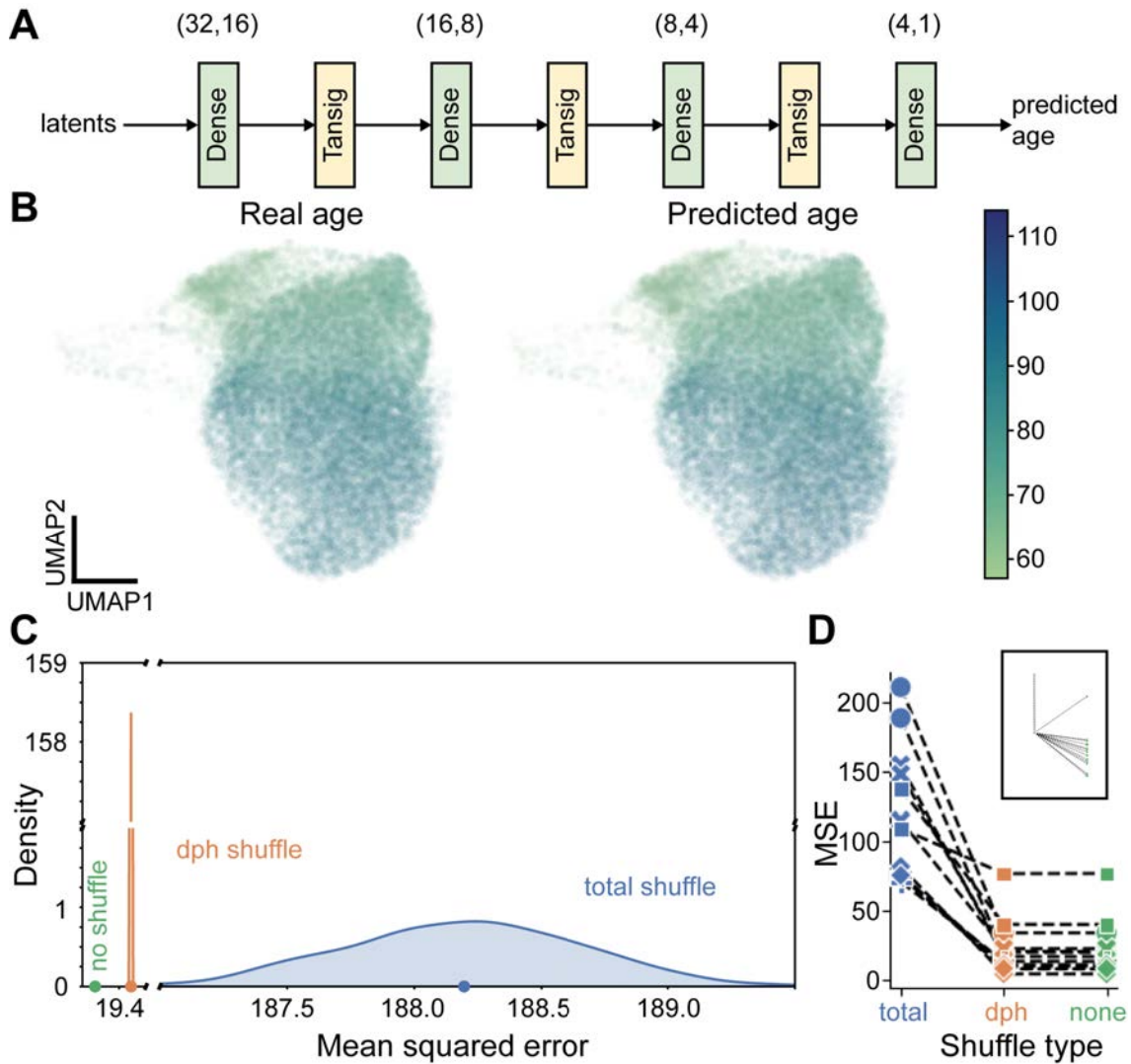


FIGURE 4.1: Predicting syllable age. **A** Neural network design for predicting age. **B** UMAP embedding of VAE latents for held out example syllable renditions, color-coded by age at production time (left), or predicted age (right). **C** Mean squared error of age predictions (green). Distribution and mean of MSE over 1000 random shuffles of age (blue). Distribution and mean of MSE over 1000 within-day shuffles of age (orange). All values were based on permutations of the example syllable data in **B**. **D** Means of MSE for 1000 total (blue) or within-day (orange) permutation tests for every syllable, along with true MSE on held out data (green). Repeated measures ANOVA (with syllable type as level) indicates significant effect of permutation, and *post hoc* tests show all comparisons are significant ($p < 0.05$). Inset shows ‘dph shuffle’ and ‘none’ values with ‘dph shuffle’ value subtracted to emphasize that for all syllables but one within-day shuffling increases MSE.

output of the trained network. The apparent similarity between these plots suggests the success of the network in learning the intended function. I quantified network performance over these held out data using the mean squared error (MSE) of the network’s age predictions. This average error is presented as the ‘no shuffle’ value in Fig. 4.1C. Network performance exceeds the expected performance under two null models, assessed via bootstrap comparisons (see Predicted age model evaluation in Methods). In particular, performance greatly exceeds what is possible in the absence of a meaningful relationship between production age and acoustics (‘total shuffle’ condition). Performance also exceeds what is possible in the absence of meaningful within-day improvements in predicted age (‘dph shuffle’ condition), although these performance gains are relatively small in magnitude. The evaluations of this example syllable are consistent with evaluations across all syllables (Fig. 4.1D; syllable-level repeated-measures ANOVA, effect of shuffle type $p < 10^{-9}$; ‘none shuffle’ MSE less than ‘total shuffle’ MSE by 95.09 ± 12.395 , $p < 10^{-4}$ in Tukey-Kramer post hoc test; ‘none shuffle’ MSE less than ‘dph shuffle’ MSE by 0.059 ± 0.018 , $p < 0.05$ in Tukey-Kramer post hoc test).

4.2.2 Daily predicted age increases depend on D1R signalling in Area X

Changes in song acoustics occurring over hours in juveniles have been attributed to learning (Tchernichovski et al. (2001), Derégnaucourt et al. (2005), Ravbar et al. (2012), Kollmorgen et al. (2020)). However, prior work has not established a dependence of these rapid changes on the neurobiological mechanisms known to support learning in adult birds. For example, within-day changes in pitch in response to pitch-contingent noise playback require activation of D1 dopamine receptors (D1Rs) in Area X. Additionally, these receptors are implicated in juvenile song learning at long timescales, since their long-term blockade prevents the acquisition of tutor-similar song structure (Hisey et al. (2018)). The predicted age measure developed

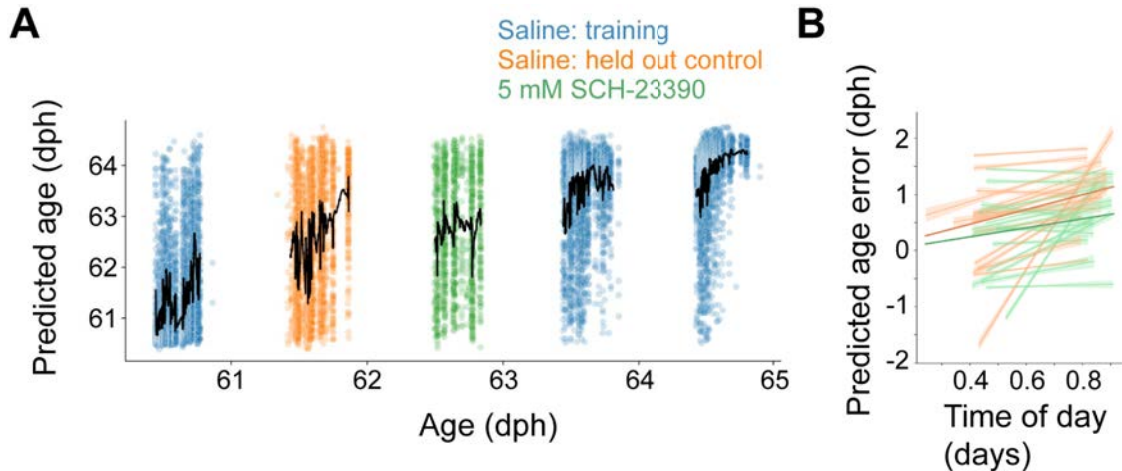


FIGURE 4.2: D1R signaling underlies within-day changes in predicted age. **A** Predicted vs real production age for renditions of an example syllable, color-coded by condition. **B** Errors in predicted age by time of day. Thin lines and shaded regions are means and 95% confidence intervals for every syllable in the experiment, with linear relationships fitted as described in the methods. Darker thick lines show the predicted age error’s partial dependence on time of day and treatment condition. Interaction between time of day and treatment is significant ($p < 10^{-10}$) in mixed-effects ANOVA (see methods).

here exhibits within-day changes, as established by the comparison of ‘none shuffle’ to ‘dph shuffle’ in the last section. I collaborated with another graduate student, Ji-axuan Qi, to determine whether within-day changes in juvenile predicted age depend on D1R signaling.

Ji-axuan used reverse microdialysis to deliver Ringer’s solution or the D1R antagonist SCH-23390 to Area X in juvenile birds (61 to 72dph). Drug was administered for the duration of a single waking day, surrounded by at least two days of vehicle control that permitted unmanipulated singing and learning. Ji-axuan represented syllable renditions as VAE latent vectors and assigned syllable type labels to different clusters of renditions. To test whether D1R activation supports the daily song changes reflected in changing predicted ages, I trained predicted age networks separately for each syllable in the dataset. I withheld from training all renditions

performed on the day of drug treatment as well as all renditions on the preceding control day. I used these networks to calculate predicted age for all syllables, including the held out renditions (for example, see Fig. 4.2A). To test whether D1R activation supports within-day increases in predicted age, I modeled predicted age network errors as a linear function of time of day, drug treatment condition, and an interaction between these factors. Modeling the network errors removes absolute (intercept) differences between predicted age values on different days while introducing a constant offset in the dependence on time of day (slope): the slope of predicted age vs. time of day is $1 +$ the slope of network errors vs. time of day. In the linear model, I included random effects of bird and syllable on the intercept and time of day slope (see methods).

Fig. 4.2B depicts the model predictions and confidence intervals for individual syllables (including their modeled random effects), as well as the overall partial dependence of network error on treatment and time of day. Within-day predicted age increases were significantly larger on control days with intact D1R signalling than on days with D1R blockade (interaction value is 0.511 ± 0.077 predicted age days/day, $p < 10^{-10}$). No other model terms reached significance at an alpha threshold of 0.05, but a main effect of model error on time of day exhibited a trend (1.59 ± 0.916 predicted age days/day, $p < 0.1$). (Note that this trending dependence of errors suggests a significant dependence of predicted age on time of day.) The steeper dependence of network error on time of day in the control condition indicates that within-day changes in predicted age at least partly reflect learning through a D1R-dependent mechanism in Area X, consistent with the observation that long-term blockade of these receptors interferes with the acquisition of tutor-similar song structure (Hisey et al. (2018)), and consistent with the mechanisms of pitch learning in adult birds.

4.2.3 *Overnight consolidation is quantile-dependent*

Previous work has examined other measures as proxies for rendition maturity or quality. Derégnaucourt et al. (2005) used within-syllable entropy variance as a measure of acoustic complexity. They report that syllables that become increasingly acoustically complex over development gain complexity in a circadian, non-monotonic pattern. In particular, they report that syllable renditions rapidly increase in complexity during the morning to a plateau value that is maintained during afternoon singing. Overnight, most of the day’s gains are lost; birds begin the next morning singing less complex syllables than the prior day’s plateau value. In this way, high syllable complexity is achieved through a “two steps forward, one step back” circadian pattern. Although only some syllables exhibit this pattern with respect to entropy variance, this result has been widely accepted as evidence of a general pattern of syllable acoustic development. Independent from my research, Kollmorgen et al. (2020) developed a general approach to scoring syllable maturity with similarities to the predicted age metric I developed. Their work apparently qualifies the conclusions that were drawn by Derégnaucourt et al. (2005). In particular, they observe that birds are able to produce syllables that range widely in maturity during plastic song. Overnight ‘reversion’ is asymmetric in this distribution. Birds produce highly immature songs in the morning, including less mature variants than they produced the previous evening. On the other hand, mature morning songs are *not* ‘regressive’ compared with mature evening songs.

Although the role of sleep in juvenile song learning is poorly understood, these behavioral signatures of overnight consolidation (or deconsolidation) have influenced the field. I sought to determine whether the predicted age metric that I developed exhibits overnight patterns similar to or different from those seen by Derégnaucourt et al. (2005) or Tchernichovski et al. (2001). As mentioned above, the findings

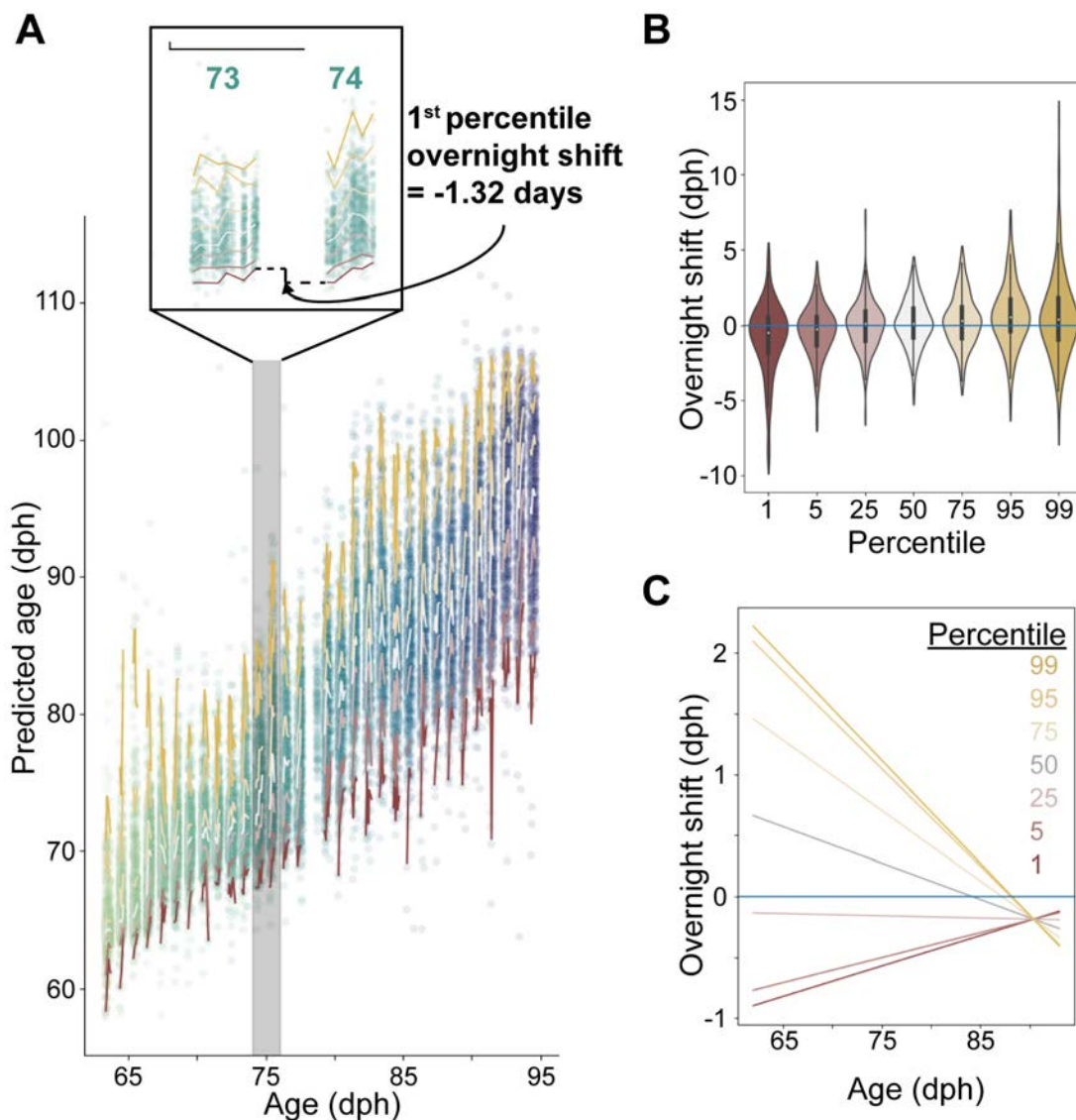


FIGURE 4.3: Overnight consolidation. **A** Predicted vs actual production age for held out renditions of an example syllable. Overlaid traces plot percentiles of the predicted age measure by production age for time-binned data. Inset shows two consecutive days, and the calculation of a percentile-wise overnight shift. **B** Histograms of percentile-wise overnight shifts for all overnight shifts in the dataset. **C** Partial dependence of overnight shift value on age and percentile after marginalizing dependence on random effects. The main effect of percentile, main effect of age, and interaction between age and percentile are significant, as reported in the text.

of Kollmorgen et al. (2020) suggest that overnight shifts in quality may vary as a function of predicted age quantile. As illustrated in an example syllable in Fig. 4.3A, I binned data by age at production time into 0.1 day bins and calculated predicted age at each of several quantile levels in each bin. By comparing predicted age at each quantile level in adjacent evenings and mornings, I calculated a quantile-level overnight shift, as illustrated by the calculation of a 1st percentile overnight shift in the inset in Fig. 4.3A. I present the distributions of quantile-level shifts for all nights across all syllables in the dataset in Fig. 4.3B. Consistent with Kollmorgen et al. (2020), only immature data quantiles tended to exhibit overnight ‘reversion,’ which results in a tendency towards negative shift values at the 1st and 5th percentiles. In fact, I observed the opposite pattern in mature data quantiles, which tended to exhibit positive shift values.

To quantify the dependence of overnight shift on quantile level, and to examine these data for age-dependent patterns in overnight consolidation, I modeled overnight shift magnitudes as a linear function of quantile level and age, as well as the interaction between quantile and age. I also included random effects of bird and syllable on shift intercepts and quantile slopes. This modeling indicates a significant negative shift intercept (intercept value is -2.481 ± 1.064 , $p < 0.05$), but this tendency towards overnight reversion is significantly quantile-dependent (main effect of quantile (predictor values in range $[0,1]$) is 10.11 ± 1.641 , $p < 10^{-8}$) and slightly moderated with age (main effect of age (dph) is 0.0259 ± 0.0127 , $p < 0.05$). Finally, the quantile slope decreases with age (age by quantile interaction is -0.112 ± 0.0202 , $p < 10^{-7}$). The joint impact of these model factors is depicted in Fig. 4.3C, which shows the linear fits of overnight shift by age separately for different quantile levels.

4.2.4 *Developing isolate song exhibits consistent overnight reversion*

In the previous section, I demonstrated that the predicted age metric I developed reproduces prior reports of a quantile-dependent overnight shift in syllable maturity (see Kollmorgen et al. (2020)). This pattern is regarded as an important signature of song copying (Derégnaucourt et al. (2005), Kollmorgen et al. (2020)) although the reasons for the pattern remain unclear. I next sought to determine whether this pattern reliably occurs during auditory feedback-guided juvenile song development even in the absence of tutoring. I fit predicted age networks to training partitions of isolate syllable clusters. As a prerequisite to examining overnight shifts, I confirmed that isolate renditions exhibit multiday and within-day advancement in predicted age. As with tutored animals, MSE on held out isolate data is reliably lower than MSE on total and dph shuffled datasets (see Fig. 4.4A; syllable-level repeated-measures ANOVA, effect of shuffle type $p < 10^{-11}$; ‘none shuffle’ MSE less than ‘total shuffle’ MSE by 116.38 ± 14.63 , $p < 10^{-5}$ in Tukey-Kramer post hoc test; ‘none shuffle’ MSE less than ‘dph shuffle’ MSE by 0.108 ± 0.01 , $p < 10^{-7}$ in Tukey-Kramer post hoc test).

Having established that predicted age exhibits reliable within-day trends in isolate singing, I sought specifically to compare overnight shift patterns in tutored and isolate birds. I generated quantile-level predicted age scores for isolate syllables using the same procedures I used for tutored birds. I then used these quantile scores to calculate overnight shifts in isolate song as in tutored song. I modeled tutored and isolate overnight shifts as a linear function of quantile, age, and tutoring condition, as well as all two-way interactions and the three-way interaction between these predictors. I also allowed terms for random intercepts and quantile slopes at the bird and syllable level. The predictions of the fixed-effects model are given in Fig. 4.4B. Notably, the model indicates that the dependence of shift on quantile is significantly abridged in isolate birds (as given by two-way interaction term, quantile

slope is 10.797 ± 3.048 greater in tutored than isolate birds, $p < 10^{-2}$). In particular, isolate birds appear to undergo overnight ‘reversion’ at all quantile levels, whereas in tutored birds only immature quantiles exhibit this pattern.

4.3 Conclusions

In influential models of song learning, the performance of ‘bad’ renditions generates a negative reinforcement signal that discourages similar sounds from being produced in the future. Conversely, the performance of ‘good’ renditions leads to a positive reinforcement signal that makes the production of similar sounds more likely. In these models, changes in the acoustic distribution of vocal output relate directly to the evaluation of different sounds. The primary aim of the research in this chapter is to develop from this theoretical framework a procedure to infer the quality of different sounds. Since the theory suggests that low-value sounds produced early in development will become rare late in development, and vice versa for high-value sounds, I reasoned that high- and low-value sounds can be distinguished by the ages at which they are likely to be produced. In that case, I aimed to quantify the value of a sound as the age at which its production is most likely. This inference is accomplished using a feedforward neural network that takes location in VAE latent space and predicts age. This method generalizes to unseen data and explains long-term trends as well as within-day acoustic trends. Note that some acoustic changes in song may arise during development for reasons unrelated to song evaluation and downstream reinforcement. The approach here does not currently distinguish between learned and unlearned sources of behavioral change. Even so, predicted age remains likely to correlate with song quality and reinforcement, even if its incorporation of extraneous, unlearned features adds noise to the correlation.

Leveraging data from a pharmacological manipulation performed by Jiaxuan Qi, I demonstrated that within-day changes in predicted age require D1R activation.

D1R-dependent plasticity mechanisms play a role in adapting song output to auditory feedback in adult finches and are more broadly known to support song copying in juveniles (Hisey et al. (2018)). Thus, the result here suggest that within-day trends in predicted age are generated by learning mechanisms in Area X, at least in part. This result mitigates the concern raised above about unlearned features influencing predicted age. It motivates the use of this metric for future mechanistic studies of juvenile song learning. These avenues for research are developed more fully in the next chapter.

Having validated the relevance of this metric to Area X-dependent learning, I sought to determine whether it exhibits the circadian patterns that have been reported for other measures of syllable maturity. Specifically I sought to determine whether predicted age regresses overnight (Derégnaucourt et al. (2005)), and whether this pattern is uniform or differs by quantile (Kollmorgen et al. (2020)). Overnight predicted age patterns were broadly consistent with patterns described previously. Only immature quantiles regress overnight. I observed that mature quantiles actually advance overnight – a stronger quantile dependence than previously reported (Kollmorgen et al. (2020)). The map from acoustics to predicted age (or ‘neighborhood time’ in the case of the maturity score developed by Kollmorgen et al. (2020)) is complex. Moreover, my research in Chapter 2 suggests, in agreement with Ravbar et al. (2012), that acoustic development not only exhibits trends in the mean behavior but also circadian patterns in variability. The relationship between the circadian variability pattern and the circadian predicted age pattern remains unclear, and it is possible that the former partly explains the latter. In particular, variable morning singing may more readily produce especially good *and* especially bad variants than less variable evening singing. Kollmorgen et al. (2020) do not explicitly consider the role of rendition-by-rendition variability in the generation of patterns they describe. Without explicit forward models in an acoustic space, like those developed in the

previous chapter, it is challenging to explain how patterns in a more more abstract measure like predicted age arise. These considerations point to the utility of the framework developed in this thesis, where syllable development can be explicitly scored with a metric like predicted age, but also described more agnostically as a distribution in an informative, tractable acoustic space. For example, in future work the dependence of predicted age patterns on circadian variability fluctuations can be assessed by simulations of development in acoustic space with different variability patterns. For example, future work can simulate developing syllable renditions as draws from forward models with fixed within-day variability to determine whether syllable renditions produced in that way still exhibit quantile-dependent overnight shift effects.

After confirming these basic similarities between predicted age and other measures of developmental maturity, I sought to determine whether these canonical overnight patterns require tutoring, or if they also occur during auditory feedback-dependent learning by isolate juveniles. As a prerequisite, I established that isolate song exhibits within-day and multiday predicted age trends. I determined that the dependence of overnight shifts on quantile is much greater in tutored than untutored birds. As mentioned above, the underlying acoustic patterns responsible for overnight shifts in maturity remain unclear and may relate to circadian patterns in variability. In light of this point, the finding in the last chapter that circadian patterns of variability appear attenuated in isolate singing may partly explain differences between the overnight shifts in maturity exhibited by isolate and tutored juvenile song.

4.4 Methods

4.4.1 Developmental dataset preparation

The normally developing and isolate syllable level datasets from the last chapter were used for the analyses of normal and isolate song development in this chapter. Detailed methods on data collection, representation of component sounds with VAE latents, syllable-type labeling, and partitioning into training and test sets are presented in the last chapter.

4.4.2 Pharmacology datasets

For reverse microdialysis experiments, 7 juvenile (50 to 60 dph) birds were implanted with microdialysis probes in Area X using stereotaxic coordinates. After implantation, probes were flushed with a saline solution daily to minimize clogging. After post-operative singing returned to baseline rates (3 to 7 days), bird vocalizations were continuously monitored and recorded with SAP (Tchernichovski et al. (2000), see recording methods in prior chapter). On experiment days, probes were flushed with saline or 5mM SCH-23390, a D1 dopamine receptor antagonist, an hour before the bird’s lights came on. D1R antagonist was flushed from probes with saline at the end of the days on which it was applied. These experimental procedures were executed by Jiaxuan Qi.

4.4.3 Predicted age model training

The structure of the feedforward neural network used to make production age predictions is given in Fig. 4.1A, where “tansig” layers perform a hyperbolic tangent sigmoid transformation. Network input is the 32-dimensional latent vector describing observed spectrograms. Network training occurred separately for different syllable classes. During training, the network minimized an error consisting of squared prediction error and a regularization term consisting of the sum of squares of the network

weights. These terms were combined with weights given by Bayesian regularization (Dan Foresee and Hagan (1997)) by training the network in Matlab with the training function parameter set to ‘trainbr.’ I included regularization in order to “smooth” the predicted age function in acoustic space, to improve generalizability when regions of data were systematically withheld from training sets (as in pharmacology experiments). The training data input to this procedure was automatically partitioned by Matlab into a 70/15/15 split corresponding to training, test, and validation sets respectively. Training iterated until performance on the test set failed to improve for three consecutive epochs, or until 30 minutes had elapsed. The validation subdivision was unused, because I used a separate held-out dataset (see Developmental dataset preparation) for subsequent analysis.

For the D1R pharmacology dataset, all renditions on the drug treatment and preceding saline control day were withheld from predicted age network training. The remaining data was split 70/15/15 into training, test, and validation datasets for model training and for an stopping criterion.

4.4.4 Predicted age model evaluation

I quantified the performance of the predicted age networks relative to null models as mean squared error (MSE) on held out test sets. Unlike the training data, test sets were limited to renditions produced from 60 to 95dph. I compared to network performance against two benchmarks. In the ‘total shuffle’ benchmark, I shuffled production ages relative to latent descriptions for all syllables in the test set. Then I computed the MSE of model predictions based on shuffled ages. For each syllable, I performed this permutation test 1000 times and recorded the average MSE from these experiments as the ‘total shuffle’ benchmark performance score. In the ‘dph shuffle’ benchmark, I again shuffled ages relative to latent vectors, but I prevented shuffles that swapped the ages of syllables produced on different days. As with

the other benchmark, I performed 1000 ‘dph shuffle’ experiments and recorded the average MSE from these experiments as the ‘dph shuffle’ benchmark performance score. The same analysis was performed to quantify performance of predicted age networks on isolate song.

4.4.5 Analysis of D1R treatment

Predicted age networks, trained as described in “Predicted age model training,” were used to calculate predicted age for all renditions on each drug treatment day and on the saline control day immediately preceding each drug treatment day. I calculated prediction error by subtracting actual production times from these predictions. This subtraction zero-centers predictions at different ages in different animals, but has no effect on a potential difference between the drug treatment conditions with respect to daily changes in predicted age. I fit a linear mixed effects model of these residual errors. The fixed effects model included an intercept, a main effect of drug condition, a main effect of time of day, and an interaction between treatment and time of day. The model included a bird-level random effects vector b_{bird} , populated with intercept-time of day slope pairs $(b_{0,i}, b_{1,i})$ for each bird i distributed according to the 2D Gaussian $\mathcal{N}(0, \Theta)$ estimated from the data. Similarly, the model included a syllable-level random effects vector b_{syll} , populated with intercept-slope pairs $(b_{0,j}, b_{1,j})$ for each syllable j according to $\mathcal{N}(0, \Phi)$ estimated from the data. The model was fit using Matlab’s `fitmle` function and providing the Wilkinson notation formula: ‘predicted age error $\sim 1 + \text{timeOfDay} * \text{treatment} + (1 + \text{timeOfDay} | \text{bird}) + (1 + \text{timeOfDay} | \text{syllable})$.’ More details about the general approach of linear mixed effects modeling is provided in the methods of Chapter 3.

4.4.6 Analysis of overnight shifts

For every syllable, renditions were binned by age at production time, with a bin width of 0.1 days. I calculated the 1st, 5th, 25th, 50th, 75th, 95th, and 99th percentile predicted age in each data bin, and discarded bins containing fewer than 50 syllable renditions. I subsequently calculated percentile-wise overnight shifts in predicted age by subtracting the i th predicted age percentile in the last time bin of day j from the i th predicted age percentile in the first time bin of day $j+1$. In the event that the last time bin on day j and the first time bin on day $j+1$ were separated by more than 0.6 days (which occurred in infrequent cases where morning or evening data was lost due to acquisition equipment errors), the overnight comparison was discarded.

I created linear mixed effects models of percentile-wise overnight shifts. For normally developing animals, I included fixed effects of intercept, percentile, age (given as integer-valued age on the evening of day j), and the interaction between percentile and age. The model included a bird-level random effects vector b_{bird} , populated with intercept-percentile slope pairs $(b_{0,i}, b_{1,i})$ for each bird i distributed according to the 2D Gaussian $\mathcal{N}(0, \Theta)$ estimated from the data. Similarly, the model included a syllable-level random effects vector b_{syll} , populated with intercept-percentile slope pairs $(b_{0,j}, b_{1,j})$ for each syllable j according to $\mathcal{N}(0, \Phi)$ estimated from the data. The model was fit using Matlab's `fitmle` function and providing the Wilkinson notation formula: 'overnight shift \sim 1 + age*percentile + (1 + percentile | bird) + (1 + percentile | syllable).'

I compared overnight shifts in isolates to overnight shifts in normally developing birds by modifying the model presented above. In particular, the model also included a fixed main effect of tutoring condition, the 2-way interactions of tutoring with percentile and tutoring with age, and the three way interaction of tutoring, percentile, and age. The model was fit using Matlab's `fitmle` function and providing

the Wilkinson notation formula: 'overnight shift ~ 1 + age*percentile*tutoring + (1 + percentile | bird) + (1 + percentile | syllable).'

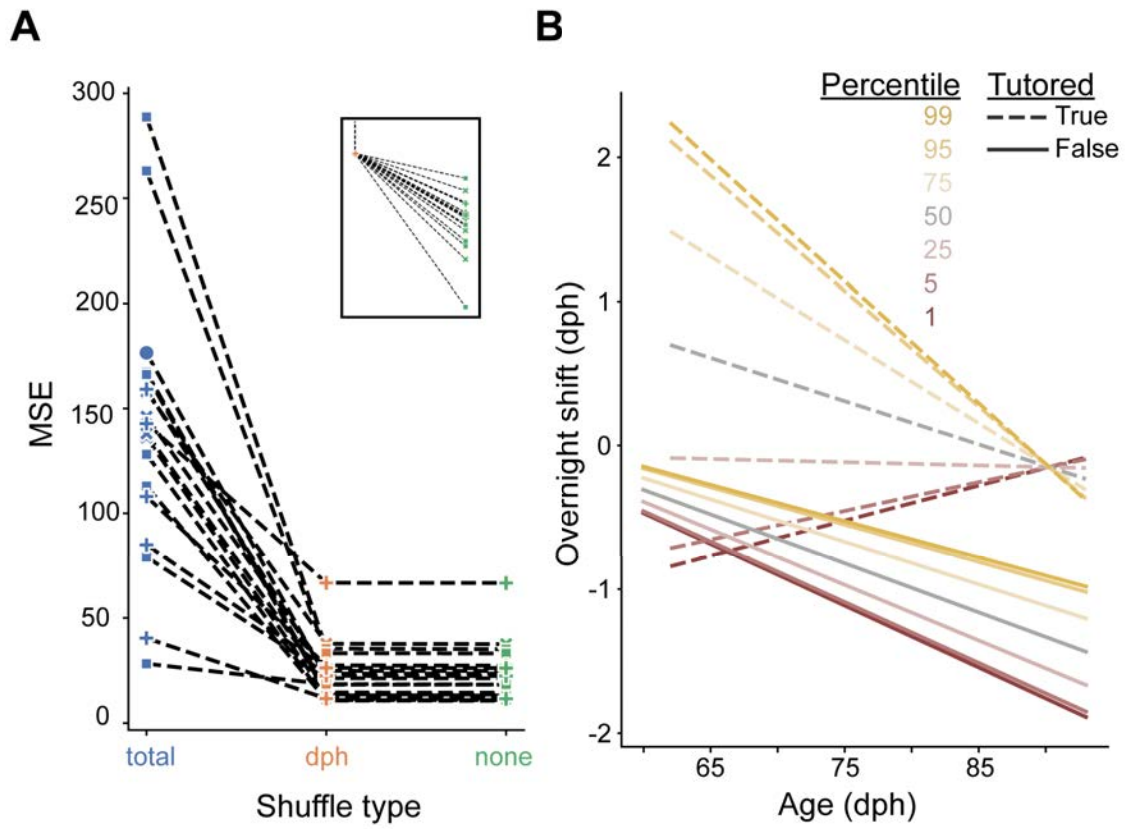


FIGURE 4.4: Overnight consolidation in isolated and tutored birds. **A** Means of MSE for 1000 total (blue) or within-day (orange) permutation tests for every isolate cluster, along with true MSE on held out data (green). Inset shows ‘dph shuffle’ and ‘none’ values with ‘dph shuffle’ value subtracted to emphasize that for all syllables within-day shuffling increases MSE. **B** Dependence of overnight shift value on age, quantile level, and tutoring status as given by linear mixed effects model. The main effect of age and tutoring, the interactions between tutoring and percentile, between tutoring and age, and between age, tutoring, and quantile are significant, as reported in the text.

Conclusions and future directions

5.1 Summary

The zebra finch model offers an unparalleled opportunity to understand how learning shapes juvenile behavior into a target adult behavior, particularly in cases of imitation, where the target is set by observing the expert performance of an adult. In the lab, we can readily record all or most juvenile ‘practice’ song renditions, as well as the model song target. In addition, previous work clearly indicates the importance of well-delineated forebrain structures — the “song system” — to song learning and production. The scientific promise of studying this natural learning process is coupled with challenges. In many other learning paradigms, experimenters control the task structure and in this way define the significance of an animal’s actions. The significance of juvenile vocal acoustics are instead defined by their utility to the animal in a learning process we do not fully understand.

First, the apparent high dimensionality raw sound measurements of juvenile singing presents a challenge. Recently a collaboration between the Mooney and Pearson labs developed a principled method to extract highly informative, low-dimensional descriptions of song renditions by modeling the constraints on song re-

vealed by correlations in the raw data. This approach, based on a variational autoencoder (VAE), has fewer experimenter biases than approaches leveraging experiment-defined features, but similarly reduces the complexity of the downstream inference problems about behavioral significance and meaning. First, I aimed to assess this approach as a description of song copying. Then I aimed to develop tools to explicitly describe the course of learning and the significance of rendition-by-rendition acoustic variation, leveraging VAE-based low-dimensional descriptions of song renditions.

In collaboration with Jack Goffinet, I assessed VAE performance extracting copying-relevant features using a sensible benchmark suggested by prior literature. Mandelblat-Cerf and Fee (2014) suggest that meaningful song features will readily differentiate songs produced by arbitrary pairs of animals, but represent as relatively similar the songs produced by pupils and their corresponding tutors. In the second chapter of this thesis, I present evidence that unsupervised feature extraction by the VAE meets this criterion, validating the application of this technique to study song copying broadly.

In the next thesis chapter, I applied the VAE to study plastic song. I developed a tool for tracking the acoustic development of plastic song syllables, represented in VAE feature space. The neural network-based tool implements a relatively unconstrained map from age to Gaussian distribution parameters. The approach is innovative in the song development field for explicitly modeling multidimensional behavioral variation. I demonstrate that the model not only learns slowly changing features of song, but also fast (within-day) changes in song distributions. In particular, the model exhibits a within-day pattern of rendition-by-rendition variability reduction that is consistent across syllables produced by normally tutored animals. I presented evidence that the magnitude of this circadian pattern requires tutor exposure.

Finally, I developed an analytic tool to quantify plastic song syllable rendition

quality, working from the assumption that changes in the probability of different syllable variants depend on reinforcement based on rendition quality. In particular, I used a neural network to estimate the mean of the conditional distribution $P(\text{age} \mid \text{latent})$. I demonstrate that this approach successfully estimates this quantity, enabling it to predict a rendition's production age from its latent description. I show that, in addition to the large changes occurring over weeks, within-day changes in acoustics contribute to the network's performance. In collaboration with another student, Jiaxuan Qi, I show that D1R signalling in Area X, known to support copying outcomes (Hisey et al. (2018)), supports within-day improvements in predicted age. This result suggests that changes in predicted age are supported by basal ganglia-based learning mechanisms, and motivates exploring these mechanisms with this tool. I demonstrated that the predicted age measure only 'reverts' overnight at immature quantiles, consistent with a recent report on syllable maturation (Kollmorgen et al. (2020)). In fact I observe that mature quantiles are advanced in the morning compared with the preceding evening. Somewhat unexpectedly, I determined that this quantile dependence is reduced in the song of developing isolates, even though isolate juveniles also use auditory feedback to learn species-typical song features (Konishi (1965)). This result suggests that auditory feedback-based learning from an internal template is not sufficient to drive the typical pattern of behavior. The pattern instead results from processes that are specifically dependent on tutoring.

Taken together, this thesis work provides tools for rigorous, unbiased investigation of the learning mechanisms underlying plastic song. It raises a number of questions and avenues for development that are elaborated below. These topics relate to (1) further investigation of the results presented in this thesis; (2) application of the tools developed in the thesis to long-standing neurobiological questions; (3) generalizations and extensions of the tools developed here.

5.2 Interpretation

5.2.1 *What features of acoustic change underlie the quantile dependence of overnight predicted age shifts?*

Interest in overnight reversion in juvenile song maturation stems originally from the observation that (for ultimately complex syllables) syllable complexity increases each day and decreases overnight (Derégnaucourt et al. (2005)). These results suggest that sleep promotes learning ‘reversion,’ a phenomenon that would distinguish sleep’s role in juvenile song learning from the effect of sleep in several forms of human skill learning, where it leads to improved motor performance (Walker et al. (2003), Fischer et al. (2002)) and decreased susceptibility of recent motor learning to interference by subsequent experience (Korman et al. (2007)). I observed that ‘overnight reversion’ of predicted age is quantile-dependent, occurring only for the most immature quantiles of the behavior. Multiple, non-exclusive patterns of behavioral change in acoustic space could contribute to this effect. On the one hand, an entropy-preserving transformation (like a rotation) of the underlying distribution in acoustic space could lead to different-direction overnight shifts of predicted age in different parts of the predicted age distribution. On the other hand, higher entropy in the morning than the evening (as observed) can also generate this pattern of overnight shifts. Lastly, these overnight transformations of the shape of the underlying acoustic distribution can combine with overnight changes in the mean of the distribution to influence measured overnight shifts in predicted age.

The quantile-dependent ‘overnight reversion’ of predicted age is broadly consistent with the results of Kollmorgen et al. (2020). Those authors suggest that overnight reversion occurs principally with respect to features orthogonal to learning direction; reversion in the learning direction is small and limited to regressive quantiles. This conclusion apparently conflicts with the work of Derégnaucourt et al.

(2005), where syllable complexity in ultimately complex syllables was chosen for study *because* increases in complexity were a long-term, slowly acquired feature of the syllables under study. The inferences by Kollmorgen et al. (2020) about a learning-orthogonal axis of daily progress and overnight reversion depend on those authors' choice to model large collections of renditions at different times of day as single points in an underlying acoustic space. The distributional differences between such collections are reflected as distances between points in their framework. However, because these abstract distances between points approximately reflect distributional overlap between collections of renditions, these distances can be influenced by changes in the mean and variance of the underlying collections of renditions. Because variance (entropy) changes systematically each day (see 3.2.3), it plausibly contributes to the behavioral trajectory identified by these authors. Although the basis for the quantile-dependence of overnight reversion remains unresolved, the tools in this thesis can offer insight, as I explain below.

The forward models developed in Chapter 3 provide a basis for testing the relative contribution of the orientation, shape, and scale of variation as well as changes in the mean to patterns of maturity scores like predicted age. For example, we can compare the flexible models in Chapter 3 to matched models with fixed entropy within each day. We can simulate behavioral datasets by sampling from these models. If draws from the fully flexible models but not the fixed-entropy models recapitulate a pattern in predicted age like the quantile-dependence of overnight consolidation, we can conclude that the pattern depends on systematic fluctuations in variability. On the other hand, if fixed-entropy models can generate data with quantile-dependent overnight reversion, we can conclude that changes in the syllable mean and distribution shape account for the pattern. If even the more flexible models cannot produce a predicted age pattern, we could consider increasing the flexibility (and by extension the complexity) of the forward modeling procedure. I present attractive next steps

to extend the forward modeling approach in Section 5.4.3.

5.2.2 Mechanisms of variability patterns

The mechanisms underlying the circadian changes in distribution entropy are unclear. Two broad, non-exclusive causes are possible. First, the negative reinforcement of ‘bad’ variants may directly cause a reduction of entropy each day. If birds learn to avoid inappropriate song variants present in their morning repertoire faster than the rate at which they acquire new ‘good’ variants, this increased selectivity of production would result in reduced entropy on its own. Importantly, this possible mechanism only explains within-day decreases in entropy; an additional mechanism must increase entropy at the start of each day, for example, by reintroducing rejected variants. Such a mechanism plausibly contributes to the entropy effect I observed and is consistent with the pattern of overnight reversion at immature predicted age quantiles described in 4.2.3. In particular, it is consistent with within-day improvement of immature quantiles by elimination of the worst variants, followed by reversion of immature quantiles overnight by reintroducing this variation. Of particular interest for future work will be whether the contractions of variability are symmetric in acoustic space or more limited to dimensions that impact song quality, in which case variability mechanisms offer a strategy to reduce the production of especially poor renditions (while also reducing the production of especially good renditions).

On the other hand, influences on behavior besides reinforcement may also impact variability. In adults (Kao et al. (2005)) and late plastic song juveniles (Kojima and Doupe (2011)), the introduction of a female immediately decreases the rendition-to-rendition variability of song, indicating that variability can be controlled independently of learning. This social context-dependent control relies on processes in the AFP (Kao et al. (2005)), including Area X (Woolley et al. (2014), Singh Alvarado et al. (2021)). Plausibly, circadian processes in the AFP that do not depend on

reinforcement could also drive circadian variability changes. In mammalian striatum, dopaminergic tone undergoes circadian oscillations through regulation of the dopamine transporter driving reuptake (Castañeda et al. (2004), Ferris et al. (2014)). It is plausible that dopamine levels fluctuate with a circadian pattern in Area X. Previous research indicates that D1R signalling in Area X can influence adult song variability (Leblois et al. (2010)), although noradrenaline has also been implicated (Singh Alvarado et al. (2021), Castelino and Ball (2005)). In addition, Area X FoxP2 expression falls in juveniles' first waking hours each day (Teramitsu et al. (2010)). FoxP2 dysregulation in Area X increases the variability of syllable sequencing in adults (Xiao et al. (2021)) and disrupts the regulation of acoustic variability by social context (Murugan et al. (2013)). The involvement of dopamine tone or FoxP2 expression in directly regulating circadian variability because these factors are independently implicated in juvenile learning outcomes (Hisey et al. (2018), Haesler et al. (2007), Heston and White (2015)). In this way, the within-day regulation of variability may be mechanistically linked to learning-related processes. To this point, I note finally that reinforcement learning depends crucially on the actual actions an animal takes and is influenced by the variability in its behavior. The circadian regulation of song variability may play a direct role in a juvenile learning algorithm, although I do not know of theoretical work exploring this possibility in formal models (Doya and Sejnowski (1996), Fiete et al. (2007), Fee and Goldberg (2011)).

5.2.3 The role of tutoring in song development

Influential early research determined that isolate birds develop species-typical song characteristics that are absent in the adult songs of birds deafened as juveniles (Konishi (1965)). This result has widely been interpreted as evidence for an auditory error-driven learning process in isolates that converges on a generic 'innate template.' In this model, tutoring may serve only to increase the specificity or add features to an

internal template (Marler (1970)). This simple model predicts similar intrinsic behavioral patterns in tutored and isolate birds as they differ only in their template or target, not in the learning algorithm they are implementing. By contrast, tutoring may impact juveniles in many other ways. The first tutoring experience changes the morphology and physiology of cells in juvenile HVC within 24 hours (Roberts et al. (2010)). It also leads rapidly to increased acoustic feature diversity in juvenile singing (Tchernichovski et al. (2001)). These rapid effects on a song production nucleus and on behavioral output suggest the possibility that tutoring impacts behavioral patterning directly, that is, independent of its effect on an evaluative reward function that impacts song through reinforcement learning. These effects do not rule out the possibility that tutoring affects behavior exclusively through its impact on the evaluative function, provided reinforcement learning under updated evaluative function rapidly alters behavior.

I sought to determine whether the plastic song of tutored and isolate birds differed with respect to intrinsic patterns – that is patterns of behavior relative to itself that do not depend on the specific acoustics of the template. In particular, I evaluated whether isolate and tutored plastic song syllables have similar circadian dynamics in their distribution entropy, and similar quantile-dependence of overnight predicted age shifts. I found significant differences between tutored and isolate song with respect to both of these patterns. These results raise the question whether tutoring affects juvenile song production exclusively through its influence on the template guiding learning, or whether the tutoring experience impacts vocal production in independent, direct ways as well. Future work is required to explain these circadian patterns in tutored animals and subsequently to explain the dependence of those mechanisms on tutoring. I have already discussed the possible involvement of the AFP in generating circadian variability patterns. In fact, song learning requires NMDAR currents in LMAN during tutoring (Aamodt et al. (1996)). Tutoring affects

the timecourse with which LMAN neurons develop mature physiology (Livingston and Mooney (2001)) and leads to phosphorylation of CaMKii in Area X spiny neurons (Singh et al. (2005)). Thus, several avenues exist for the influence of tutoring on juvenile song variability via a direct influence on the AFP.

5.3 Testing a model for basal ganglia-based reinforcement learning

Perhaps the most important contribution of this thesis is the operationalization of a number of concepts in a long-standing neurobiological theory of song learning. I review this theory and explain its interaction with the methods developed in this thesis below.

5.3.1 *Review of an influential model*

The AFP has been theorized to play critical roles in reinforcement learning-based song copying for many years (Doya and Sejnowski (1996), Fiete et al. (2007), Fee and Goldberg (2011)). An influential current model (Fee and Goldberg (2011)) is reviewed more completely in the introduction to my thesis, and briefly summarized here. The model posits that vocal motor variability is introduced to a song production pathway that otherwise produces stereotyped but immature song. In particular, variable LMAN activity induces exploratory variability in the premotor firing of RA, which is otherwise controlled by stereotyped inputs from HVC. At the same time, these variable LMAN patterns are transmitted to Area X via axon collaterals (Vates and Nottebohm (1995)). In adults responding to experimenter feedback, when a variant produced this way is ‘better’ than expected, a phasic activity burst in dopaminergic afferents to Area X (Gadagkar et al. (2016)), controlled by auditory evaluative circuits (Keller and Hahnloser (2008), Kearney et al. (2019)), communicates this success to Area X. Conversely, ‘worse’ than expected vocal motor outcomes are communicated by phasic suppression of baseline tonic firing in the same

dopaminergic projection (Gadagkar et al. (2016)). Plasticity in Area X, organized by rendition-to-rendition information about LMAN-induced motor variation (Kearney et al. (2019)) and its dopamine-reported quality (Hisey et al. (2018), Xiao et al. (2018)), generates song-locked Area X activity patterns that causally induce those LMAN premotor patterns that lead to good outcomes (Andalman et al. (2009), Warren et al. (2011)). Finally, because this mechanism induces the preferential adoption of RA premotor activity patterns that generate ‘good’ variants, it induces activity-dependent plasticity in RA that makes these activity patterns intrinsically preferred (i.e., they eventually occur even without a biasing signal from LMAN). Except for the observation that LMAN induces rendition-to-rendition song variability in juveniles (Ölveczky et al. (2005), Goldberg and Fee (2011)), and the observation that manipulation of AFP nuclei interfere with learning outcomes (Bottjer et al. (1984), Scharff and Nottebohm (1991)), almost no features of this attractive model have been explicitly tested during juvenile song learning. This lack of experimental support owes largely to the abstract predictions of the model with respect to juvenile behavior: generally the predictions involve concepts like ‘good’ and ‘bad’ song renditions, and stochastic policies in ‘acoustic space,’ concepts have been hard to operationalize in practice. The work in my thesis operationalizes much of this model in a plausible way. In this section, I will leverage the tools of my thesis to reframe this theory using quantities we can now calculate from experimental data.

5.3.2 Dopaminergic reinforcement signals

The idea that dopaminergic afferents to striatal structures reinforce behaviors has recently been experimentally supported in adult zebra finches as well (Hisey et al. (2018), Xiao et al. (2018), Gadagkar et al. (2016)). In the case of juvenile bird song, it is technically challenging but feasible to collect exhaustive recordings of juvenile vocalizations and, on a representative subset of vocalizations, paired recordings of

dopaminergic cells. The analysis procedures developed in this thesis could be applied to such a dataset to test the hypothesis of vocal variant-driven reinforcement in the following way. The first test is analogous to the first experiment type described above. In particular, *post hoc* calculations of song rendition predicted age can serve as a correlate of variant quality, and therefore as a correlate of the hypothesized reinforcement value of the rendition as an acoustic stimulus. Thus the theory of juvenile song reinforcement learning, combined with the tools developed here, make a concrete, testable prediction: rendition predicted age should positively correlate with subsequent phasic firing in dopaminergic afferents to Area X. My modeling work also makes possible a more complete account of the relationship between song learning and dopamine dynamics. Analogous to experiments that link dopamine firing to changes in behavior, it is possible to use the forward models from chapter 3 to explicitly represent the changes in behavior during learning and relate those changes to a neural signal like dopaminergic cell firing rate. In particular, we could calculate explicitly a dynamic representation of the behavior from the forward models as follows. Given the distributional model $P(\text{latent} \mid \text{age})$ developed in Chapter 3 and a small time increment δ , we can approximate time-varying dynamics of the behavior as function of time t : $P(\text{latent} \mid t + \delta) - P(\text{latent} \mid t - \delta)$. This difference is positive at latent locations that are becoming more likely and negative at latent locations that are becoming less likely. We can then ask if these local dynamics could be explained by dopamine-based reinforcement by testing if renditions that precede phasic dopamine release have a latent representation that is becoming more likely. Although this design does not involve experimenter control of dopamine dynamics it offers the ability to look in detail at whether dopamine dynamics could explain changes in behavior.

5.3.3 Premotor contributions of the Anterior Forebrain Pathway

The reinforcement learning theory outlined above predicts abstract premotor functions of Anterior Forebrain Pathway nuclei that can be operationalized concretely with the tools developed in this thesis. In particular, the theory predicts that song is initially updated in response to auditory feedback through reinforcement-induced changes to AFP activity, which influences premotor activity in downstream RA. In this theory, activity at many successive nodes of the AFP carries this adaptive bias signal: Dopamine-dependent plasticity in Area X initially alters song-locked medium spiny neuron activity; these alterations ramify downstream as altered pallidal neuron activity, DLM neuron activity, and LMAN neuron activity. The model does not predict specific differences between the premotor function of these nodes.

If song changes that have been acquired recently through reinforcement learning depend on a premotor signal in the AFP, song produced without the influence of biased AFP activity will lack those recently acquired changes. Since those changes presumably acted to maximize reinforcement based on song quality, the song produced without them should be ‘worse’ than it would otherwise be. Thus the reinforcement learning theory predicts the effect of transiently silencing nodes in this network in terms of recently acquired behavioral changes and improvements and regressions of song quality. The tools developed in this thesis allow us express these abstract predictions using quantities we can calculate from behavioral data. In particular, the prediction that song produced without AFP contribution should be ‘worse’ can be operationalized concretely as a prediction that predicted age networks trained on unmanipulated song renditions will assign lower predicted age values to renditions produced during AFP suppression than to held out unmanipulated renditions. Moreover, the prediction that recently acquired changes will fail to express during AFP suppression can be framed as a prediction that a manipulated rendition with

acoustics L at time t will have higher probability under $P(L | t - \delta)$ for some short lag δ than under $P(L | t)$, given a trained forward Gaussian model $P(\text{latent} | \text{age})$. More generally, the forward models can be adapted to take AFP manipulation state as an input, allowing them to model the difference between manipulated and unmanipulated song very freely. Ongoing work to developing these ideas is presented in Appendix B.

Beyond this role in adaptively modifying song, a more securely established premotor function of the Anterior Forebrain Pathway is its role increasing the rendition-to-rendition variability of song. With respect to this function, manipulations at different nodes of the pathway may yield different results. Some reports indicate that normal juvenile song variability requires activity in LMAN (Ölveczky et al. (2005) and DLM (Goldberg and Fee (2011)), but not Area X (Goldberg and Fee (2011)). On the other hand, recent results indicate that Area X may generate the residual variability found in adult song that varies with social context (Singh Alvarado et al. (2021)). As such, the developmental timeline of Area X's role in generating variability remains unclear. The forward modeling tools in chapter 3 provide a foundation to characterize behavioral variability in detail. They model the overall amount of variability and its orientation along different directions in latent space. Thus an experimental design like the one proposed above can be combined with forward modeling to address the possibly age-dependent role of Area X in variability generation. As mentioned above, we can adapt the forward models to include a manipulation input. This input enables to the models to generate different distributions in latent space for different experimental conditions (with and without Area X participation, for example) for the same age. By examining the covariance matrix of models fit with and without AFP participation, future work can explore in greater detail the contribution of different AFP nodes to the shape of song variation.

5.3.4 *Consolidation outside the Anterior Forebrain Pathway*

The reinforcement learning model outlined above predicts that reinforcement adaptively organizes activity in Area X, leading to that structure's production of activity that biases behavior towards 'good' song variants. However, adult song that fully incorporates the adaptations learned in development can be produced without Anterior Forebrain Pathway participation (Bottjer et al. (1984)). This observation indicates that if adaptive biases are learned and implemented in Area X initially, they must consolidate elsewhere in the song system. Theoretical work describes how repetition of adaptive RA firing under LMAN 'guidance' could induce plasticity in RA making those patterns intrinsically preferred and LMAN-independent (Teşileanu et al. (2017)). As with other questions in this section, this model has little experimental support in juveniles. Even if it is broadly correct, the timecourse of consolidation is not yet determined. Ölveczky et al. (2005) analyzed songs produced during LMAN inactivation with traditional acoustic features and reported that these did not fall outside the normal acoustic range of unmanipulated song, indicating that LMAN-independent song cannot lag normal song drastically. In adult birds undergoing pitch learning, song produced without AFP participation changed pitch at a lag behind unmanipulated song on the order of one (Andalman et al. (2009)) or many (Warren et al. (2011)) days. The experimental design indicated above, combined with the tools developed in this thesis, can permit a more thorough characterization of the 'progress' of AFP-independent song. In particular, the forward models with an input for treatment will permit representing the evolution of AFP-independent song. Application of the tools developed in this thesis to AFP-independent song can reveal whether the already observed circadian dynamics require an AFP contribution or if they present in the consolidated (AFP-independent) singing behavior.

5.4 Limitations and extensions

5.4.1 Autoencoder assumptions and alternative assumptions

The analyses developed in this thesis leverage the acoustic representation of song by a variational autoencoder (developed by Goffinet et al. (2021)). This method is motivated as providing an unbiased low-dimensional representation of song acoustics. However, the architecture and training regime for this tool rely on assumptions without direct biological motivation. Concerns about these limitations are mitigated by the experiments in Chapter 2 demonstrating that the tool captures copied acoustic features. Nonetheless, here I discuss potentially significant assumptions and future directions that would relax these assumptions or at least assess their impact further.

Reconstruction loss function

The variational autoencoder loss function includes two terms. One stems from the prior distribution over the latent space. The other term is a reconstruction loss that reflects the ‘distance’ between the decoder output and the input spectrogram. Ideally, the distance metric might reflect a relevant aspect of zebra finch biology, like the perceptual distance between the input and output sounds. Despite extensive work to characterize zebra finch auditory responses (reviewed, e.g., in Woolley (2012)), we cannot yet calculate such a metric. Absent such a metric, we use a generic distance metric. I trained the autoencoder in these analyses to minimize the squared error, summed across pixels, of the spectrogram reconstruction. This squared error loss is widespread in a variety of applications, but other general loss functions are possible. For example, minimizing the absolute error, summed across pixels, would encourage maps that had many pixel errors equal to zero even if occasionally large errors at a pixel were made. By contrast, our error term prefers eliminating all large errors, at the cost of making small value but non-zero errors throughout. It is not clear that

such a training regimen would substantially alter important latent space patterns. However, future work could test whether any of the analysis results presented in this thesis depend on a choice among reasonable alternative reconstruction loss functions.

Encoding of time and dynamics

Perhaps the most important assumptions underlying the analyses in this thesis stem from decisions about how to represent the temporal patterns in song. Song production is organized at multiple timescales from multisecond bouts to subsyllabic elements lasting only tens of milliseconds. With the exception of shotgun VAE comparisons between pupils and tutors in Chapter 2, the analyses presented in this thesis rely on characterizing song as a sequence of discrete syllables. Several lines of evidence motivate this organizational framework. Adults expire to produce syllables and inspire briefly in the gaps between syllables (Franz and Goller (2002)). The emergence of an overrepresented syllable duration is a developmental milestone in early juvenile singing (Aronov et al. (2008)). During early juvenile singing, HVC neurons exhibit an increased probability of burst firing at syllable onsets (Okubo et al. (2015)). However, these motivating observations are consistent with the possibility that juvenile syllable boundaries are a graded rather than a binary phenomenon. Although some juvenile syllable boundaries are marked by inspiration like adult syllable boundaries, juveniles also produce sequences of syllables separated by sound amplitude ‘gaps’ — sometimes partially voiced — during long expiratory pulses (Veit et al. (2011)). Technically, these ‘fuzzy’ boundaries limit the consistency of segmentation decisions using simple tools like amplitude thresholding (Mackevicius et al. (2019)); conceptually they raise the possibility that the underlying neural motor representation at these boundaries is intermediate to the representation of boundaries between intrasyllabic sounds and minibreath boundaries between adult syllables. Thus my decision to base my analysis on syllables limits its applicability to later stages of song

learning for technical and also possibly conceptual reasons.

Beyond the decision to represent song as a sequence of syllables, the spectrogram representation of syllables and corresponding convolutional network architecture of the autoencoder correspond to further decisions about how to treat temporal structure. In particular, all the convolutional filters are square and there is no *a priori* difference between spectrally extended patterns and temporally extended patterns. However in reality the frequency and time axes of a syllable spectrogram have completely different meanings corresponding respectively to the vocal motor action and temporal context components of a behavioral policy. These representational decisions may not impair the information encoding by the latent embeddings. However, these choices may make it harder to decompose spectral and temporal structure in the latent space. Given that control over timing and spectral features may depend on different neural structures, this decomposition may be desirable in some applications. In particular, the blended representation of these features in the latent space I have used here may add challenges in future work relating song data and neural data.

In light of these considerations, in the next section I sketch some methods to extend this work to incorporate the complexity of song’s temporal organization.

5.4.2 *Time-based extensions*

In Chapter 2, I used an approach called the “shotgun VAE” to represent pupil and tutor sounds without using syllable segmentation boundaries. In this approach, we use VAE compression of short song segments with arbitrary onsets. With sufficiently short segments, this approach corresponds to a compression of spectral structure from moment to moment. In this framework, song is represented as a continuous, temporally extended trajectory through a low-dimensional acoustic space. The VAE itself is used only to compress spectral patterns, while temporal patterns can be

represented separately by some expression of song’s dynamics in latent space. This approach also provides flexibility in representing syllable segments. In particular, it does not require the identification of syllable boundaries as a preprocessing step. Instead syllable structure is a property of the dynamics, as trajectories from relatively low-amplitude regions of acoustic space, through vocalized, high-amplitude regions, and back to low-amplitude regions. The increased flexibility of the shotgun approach has a cost, however. Dynamical models must be incorporated to analyze temporal structure, whereas intrasyllabic temporal structure was previously captured ‘for free’ by syllable-based VAEs. Thus the use of shotgun VAE approaches is perhaps best motivated for use in analyzing very early song where the cost of identifying syllable boundaries during preprocessing is highest. Otherwise, the shotgun VAE may be useful when separating spectral and temporal structure is especially important in its own right, for example when trying to relate song to neural time-series data.

Last in this discussion of dynamics, I will highlight a developing idea in machine learning with applicability to dynamic song representations. A recent category of approaches involves training variational autoencoders on paired sequential samples of high-dimensional time series data. As with the approach in my thesis, the autoencoder trains to optimize a map from the high-dimensional data to a low-dimensional latent space. Simultaneously, the objective function optimizes a latent space dynamics $f(z_t) = z_{t+1}$. The value of this approach is that the loss can incorporate priors on the dynamics that make the system easier to analyze. This training procedure will produce a latent space that facilitates that representation of the dynamics. That is, this approach enables the experimenter to choose among low-dimensional embeddings of raw data those that yield simple dynamics (e.g., Champion et al. (2019)).

5.4.3 Modeling non-Gaussian song distributions in latent space

In Chapter 3, I sought to describe age-dependent changes in latent space song distributions. In particular, I found functions of age that returned the parameters of multivariate Gaussian distributions. This approach is motivated as a trade-off between flexibility and simplicity. It makes some important assumptions; future work can extend the methods I used in order to relax these assumptions.

Perhaps the most limiting feature of Gaussian models is that they are unimodal. The distribution of song sounds at any age is typically multimodal. Thus, song data must first be clustered into its prominent modes (syllable types) in order to apply the modeling approach I developed. During the last several weeks of typical song development, syllable clustering is usually straightforward. At earlier ages, syllable clustering is more challenging. First, multiple syllables in an adult song may arise through differentiation *in situ* from a single prior syllable type (Liu et al. (2004), Tchernichovski et al. (2001)). In this case, syllable clustering may be possible both before and after differentiation, but challenging in the days during which differentiation occurs. At early ages, clustering may be difficult and somewhat arbitrary. Even at older ages, manipulations can lead to syllable clustering challenges. In this thesis, I clustered isolate song where possible, but left some isolate sounds unlabeled and unanalyzed because their categorization seemed arbitrary.

Even if clustered syllable data is approximately unimodal, it may also exhibit non-Gaussian characteristics. Most notably, it may be skewed or heavy tailed. In fact, some research suggests that adult syllables' heavy-tailed pitch distributions influence aspects of pitch learning (Zhou et al. (2018)). Multiple possibilities exist to relax the assumptions of the Gaussian models. I briefly summarize some of these avenues and their pros and cons relative to the techniques developed here.

A straightforward and general extension of the forward models presented in this

thesis would be to model age-dependent sound distributions as mixtures of Gaussians. Any smooth probability density can be approximated with arbitrary precision by a mixture of a sufficient number of Gaussian distributions, so this approach is very general in principle. This architecture requires the *a priori* specification of the number of Gaussian components, although a large number of components and a sparsity prior on the mixture weights could be used in combination to methodically explore tradeoffs between model simplicity and model accuracy.

Some sound distributions, like those corresponding to the ridge-like distributions of shotgun VAE embeddings, are sufficiently non-Gaussian that they may be more readily represented with a completely different technique. In fact, I worked with Jack Goffinet to use model latent space distributions using ‘normalizing flows,’ an extremely flexible approach to modeling arbitrary probability distributions (De Cao et al. (2020)). This technique identifies a one-to-one map (flow) from latent space onto latent space that can be used to warp a Gaussian distribution into an arbitrary shape that maximizes the probability of observed points. In initial efforts, this approach was able to identify slowly changing dependencies of acoustic distributions on age, but failed to identify rapid (within-day) changes in acoustic distributions. As a result, this method did not initially capture the circadian effects I observed with simpler models. However, including a periodic circadian parameter to the flow model could provide a helpful inductive bias to facilitate discovery of circadian patterns.

Finally, individual syllables could be modeled using more flexible parametric models than Gaussian distributions. For example, similar neural network procedures can be used to identify parameters of multivariate skew normal distributions, which include a vector of skew parameters controlling asymmetry in different directions (Azzalini and Valle (1996)). These distributions have more recently been generalized as multivariate skew t-distributions which include a single additional parameter controlling the relative weight of the distribution tails (Azzalini (2005)). Although

less flexible than Gaussian mixtures and normalizing flows, these distributions have more readily interpreted parameters. They are only suited to unimodal distributions however, so they require syllable classification like the Gaussian models I used in Chapter 3.

5.5 Conclusion

In this thesis, I validated the use of variational autoencoders to produce feature spaces that capture song copying. Then, I developed methods to use VAE latent space representations to quantitatively model juvenile song development. In particular, I developed methods to model time-varying syllable acoustic distributions, and to quantify the maturity of individual syllable renditions. These techniques demonstrate that behavior exhibits circadian patterns with respect to rendition-to-rendition variation and quantile-dependent maturation. The specific patterns exhibited by normally developing juveniles are altered in isolate juveniles, suggesting that tutoring not only affects the acoustic parameters of the target behavior, but also changes the intrinsic organization of practice behavior. I am extremely excited to see these tools used to answer long-standing neurobiological questions about the control of practice song, especially by the Anterior Forebrain Pathway. This work is ongoing. Moreover, the tools that I developed are readily extended so that they can be applied to other phases of song learning that violate assumptions about segmentation reliability and syllable clustering. More broadly, I believe the development of tools like those presented here will improve the tractability of studying natural behavior in the lab. In many cases, brain structures are likely optimized for very specific behaviors in an animal's natural repertoire, and understanding brain organization will require an understanding of these natural behaviors. I anticipate that improved ability to quantify natural behaviors, through methods like those developed here, will be essential to upcoming developments in neurobiology.

Appendix A

VAE Mechanics

Doersch (2016) provided a lot of insight in formulating the argument here, although the variational autoencoder technique is developed elsewhere (Kingma and Welling (2013), Rezende et al. (2014)). The vocalization-specific autoencoder presented here is from Goffinet et al. (2021).

A.1 The decoder formalizes generative constraints

Spectrogram representations of vocal sounds are comprehensive and have an intuitive visual representation. Nonetheless, they are typically high-dimensional. For example, the spectrograms used to represent individual syllables in the analyses in this thesis contain values at every pair of one in 128 time bins and one in 128 frequency bins, for a total of 16,384 observed values in the unit interval per vocalization. In what follows, I will refer to the pixel values of a spectrogram as a 16384-length vector \mathbf{x} . This high dimensionality significantly reduces statistical power when analyzing vocalizations, makes visualizing relationships between large numbers of vocalizations challenging, and even makes it difficult to store large numbers of vocalization obser-

vations in PC computer memory. The fundamental insight of feature-based analyses like SAP is that the constraints of the finch brain and vocal apparatus enable a much smaller number of appropriately chosen features to represent the variation in a group of finch vocalizations. These constraints manifest as complex statistical dependencies between pixel values in the spectrogram image. More formally, singing generates observations of the random variable \mathbf{X} distributed non-uniformly according to $P_{\mathbf{X}}$. As a result, although there are 16,384 values in the spectrograms I am describing, realistic spectrograms do not correspond to separate 16,384 choices. The degree of freedom of choice in specifying spectrograms is in fact much smaller.

The idea that a relatively small number of freely made choices underlie each high-dimensional spectrogram is formalized by the generative “decoder” of variational autoencoders. In particular, the small number of free choices required to parameterize a spectrogram is formalized as a random draw \mathbf{z} of the low-dimensional latent random variable \mathbf{Z} , distributed according to latent prior distribution $P_{\mathbf{Z}}$. (In the models in this thesis, $\mathbf{z} \in \mathbb{R}^{32}$.) At this abstract level, the particular shape of $P_{\mathbf{Z}}$ does not matter much. Assuming we have a lot of flexibility in mapping a numeric latent choice \mathbf{z} to a spectrogram \mathbf{x} , we could formalize our choice as a draw from a variety of prior distributions. Because it will make calculations detailed below tractable, a multivariate standard normal distribution is typically used, and is used in the analyses in this thesis. In other words, in what follows $P_{\mathbf{Z}} \sim \mathcal{N}(0, I)$, where I is the identity matrix.

It is easy to imagine that every song rendition corresponds to a low-dimensional free choice \mathbf{z} that is deterministically mapped to a spectrogram \mathbf{x} by the function $f : \mathbb{R}^{32} \rightarrow \mathbb{R}^{16384}$. In practice, to facilitate learning algorithms I will soon describe, we replace f with a probabilistic map that depends on f , namely: $P_f(\mathbf{x} \mid \mathbf{z}) \sim \mathcal{N}(f(\mathbf{z}), \sigma^2 * I)$. Our model defines an approximation of $P_{\mathbf{X}}$ albeit in abstract terms: $P_{\mathbf{X}}(\mathbf{x}) \approx P_f(\mathbf{x}) = \int P_f(\mathbf{x} \mid \mathbf{z})P_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$. If P_f is a good approximation of $P_{\mathbf{X}}$, real

spectrogram datasets generated by $P_{\mathbf{X}}$ will have high probability under P_f .

A.2 The decoder motivates a computational and inferential problem

In principle, we can assess the quality of any generative decoder model by looking at the model’s likelihood using the equation above. We could imagine searching a parameter space of f for high-quality models and stopping when we find one. A computational problem and an inference problem quickly surface. Assessing model likelihood involves evaluating $\int P_f(\mathbf{x} | \mathbf{z})P_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$. Although this integral is determined following selection of a function f and latent prior $P_{\mathbf{Z}}$, we lack a method to calculate its numerical value. The integral is a weighted average, the expectation of $P_f(\mathbf{x} | \mathbf{z})$ under $P_{\mathbf{Z}} : \mathbb{E}_{\mathbf{z} \sim P_{\mathbf{Z}}}[P_f(\mathbf{x} | \mathbf{z})]$. It is conceptually easy to imagine approximating this integral by drawing samples \mathbf{z}_1 to \mathbf{z}_n from $P_{\mathbf{Z}}$ and calculating for every \mathbf{x} in our dataset: $P_f(\mathbf{x}) \approx 1/n \sum_{i=1}^n P_f(\mathbf{x} | \mathbf{z}_i)$. However, under this naive sampling scheme, most of the time $f(\mathbf{z}_i)$ will be sufficiently far from our sample \mathbf{x} (relative to σ) as to provide almost no information about the quality of our model. As a result, we would need to sample a prohibitively large number of \mathbf{z} s to accurately estimate the likelihood of our model.

In addition to this computational problem, an inference problem is especially relevant to our interest in analyzing vocalizations. We modeled the generation of sounds as a random sample from a latent distribution, and an operation to transform that sample into a spectrogram. But we are not primarily interested in generating spectrograms. We are interested in the set of choices underlying each spectrogram that we observe in our dataset, formalized as the \mathbf{z} that gave rise to each observed spectrogram. In other words, if we have an effective generative model P_f , that model implicitly defines for an observed spectrogram \mathbf{x} the latent posterior distribution $P_{\mathbf{Z}}(\mathbf{z} | \mathbf{x})$. But it is not yet clear how to calculate anything about this distribution of interest.

A.3 The encoder

These computational and inferential problems are directly related to one another. The uninformative sampling challenge to estimating $\int P_f(\mathbf{x} | \mathbf{z})P_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$ does not arise in estimating the related quantity $\int P_f(\mathbf{x} | \mathbf{z})P_{\mathbf{Z}}(\mathbf{z} | \mathbf{x}) d\mathbf{z}$, provided we could infer our distribution of interest $P_{\mathbf{Z}}(\mathbf{z} | \mathbf{x})$. Samples from $P_{\mathbf{Z}}(\mathbf{z} | \mathbf{x})$ are distributed precisely according to their probability of generating the observation \mathbf{x} ; they are unlikely to be wholly uninformative samples corresponding to regions of \mathbb{R}^{16384} far from \mathbf{x} . Thus, we can more efficiently estimate $\mathbb{E}[P_f(\mathbf{x} | \mathbf{z})]$ over \mathbf{z} distributed according to its posterior than over \mathbf{z} distributed according to its prior (provided we can calculate the posterior in the first place). However, these expectations are not the same quantity — how are they related? To answer this question, we can consider a more general relationship between the expectation $\mathbb{E}[P_f(\mathbf{x} | \mathbf{z})]$ when \mathbf{z} is distributed according to its prior (in which case the expectation equals $P_f(\mathbf{x})$), and the expectation of the same quantity when \mathbf{z} is distributed according to an arbitrary distribution Q over \mathbb{R}^{32} . In that case,

$$\log P_f(\mathbf{x}) - D_{KL}[Q||P_{\mathbf{Z}}(\mathbf{z} | \mathbf{x})] = E_{\mathbf{z} \sim Q}[P_f(\mathbf{x} | \mathbf{z})] - D_{KL}[Q||P_{\mathbf{Z}}]$$

where $D_{KL}[A||B]$ is the Kullback-Leibler divergence between distributions A and B.

Since this relationship is true for any distribution Q , it is true for $Q(\mathbf{z} | \mathbf{x}) \sim \mathcal{N}(\boldsymbol{\theta}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}))$. In that case, we can write:

$$\log P_f(\mathbf{x}) - D_{KL}[Q(\mathbf{z} | \mathbf{x})||P_{\mathbf{Z}}(\mathbf{z} | \mathbf{x})] = E_{\mathbf{z} \sim Q(\mathbf{z}|\mathbf{x})}[\log P_f(\mathbf{x} | \mathbf{z})] - D_{KL}[Q(\mathbf{z} | \mathbf{x})||P_{\mathbf{Z}}]$$

The first term on the left hand side is the log likelihood of our decoder model, as previously discussed. In addition, if $Q(\mathbf{z} | \mathbf{x})$ is distributed similarly to the posterior

latent distribution of interest $P_{\mathbf{z}}(\mathbf{z} \mid \mathbf{x})$, the second term on the left hand side of this equation will be small. Thus the combination of a good generative model and good approximation $Q(\mathbf{z} \mid \mathbf{x})$ to the posterior latent distribution of interest jointly maximize the left hand side. Even though we do not have a method to estimate either of the terms on the left hand side, the right hand side consists of a term we can estimate efficiently and a term we can calculate directly. In particular, we have discussed that the expectation on the right hand side can be approximated with only limited sampling. In addition, since we chose to formalize our uncertainty as a latent draw from a standard normal distribution $P_{\mathbf{z}}$, and since we have also restricted ourselves to $Q(\mathbf{z} \mid \mathbf{x})$ from the multivariate normals, the divergence term on the right hand side has an explicit formula.

Given a dataset D of spectrogram observations $\mathbf{x}_1 \dots \mathbf{x}_n$, we want to find $Q(\mathbf{z} \mid \mathbf{x})$ and $P_f(\mathbf{x} \mid \mathbf{z})$ that minimize:

$$-\sum_{i=1}^n [E_{\mathbf{z}_i \sim Q(\mathbf{z} \mid \mathbf{x}_i)}[\log P_f(\mathbf{x}_i \mid \mathbf{z}_i)] - D_{KL}[Q(\mathbf{z} \mid \mathbf{x}_i) \parallel P_{\mathbf{z}}]]$$

We can calculate the gradient for the component terms for each sample and use gradient descent to arrive at a solution.

A.4 Implementation in neural networks

We have spoken only abstractly about gradient descent to optimize functions. In practice, we will limit our search for functions to a class of flexible functions with parameters that can be easily updated with respect to an objective function gradient. In particular, both $Q(\mathbf{z} \mid \mathbf{x})$ and $P_f(\mathbf{x} \mid \mathbf{z})$ will be instantiated as neural networks that take \mathbf{x} or \mathbf{z} as inputs, respectively, and return distributions in \mathbb{R}^{32} or \mathbb{R}^{16384} , respectively. In this thesis, Q is based on a convolutional neural network over spectrograms, which takes advantage of the relationships between neighboring regions of

the spectrogram to efficiently infer patterns. The family of posterior latent distributions is multivariate gaussians with diagonal plus rank 1 covariance structure. The output layer of this encoder network is 3 vectors in \mathbb{R}^{32} : one expresses the mean location of the latent posterior, \mathbf{m} ; the second, \mathbf{u} , and third, \mathbf{d} , parameterize the covariance matrix: $\Sigma = \mathbf{u}\mathbf{u}^T + \text{diag}(e^{\mathbf{d}})$. In this way the encoder network output is interpreted as the distribution $Q(\mathbf{z} | \mathbf{x}) \sim \mathcal{N}(\mathbf{m}, \Sigma)$. The decoder function takes a 32-length vector \mathbf{z} , which can be drawn from the distribution just mentioned, and expands it using transpose convolutional layers to $\mathbf{x} \in \mathbb{R}^{16384}$. We interpret this output as the distribution $P_f(\mathbf{x} | \mathbf{z}) \sim \mathcal{N}(\mathbf{x}, 0.1 * I)$.

A final point remains in connection to this implementation. As it is currently formulated, a “forward pass” through the encoder and decoder involves sampling from the posterior latent distribution, which is defined in terms of encoder parameters we want to optimize. Backpropating gradients to the encoder requires that the autoencoder is composed of deterministic functions of its input. This structural issue is resolved using the “reparameterization trick.” We introduce another autoencoder input, s , generated randomly from the standard normal in \mathbb{R}^{32} on every autoencoder forward pass. Given this random sample \mathbf{s} , and the Cholesky decomposition of $\Sigma = \mathbf{u}\mathbf{u}^T + \text{diag}(e^{\mathbf{d}})$ into lower triangular matrix \mathbf{L} and \mathbf{L}^T , the entire forward pass of the autocoder consists of backpropagation-friendly operations, with decoder input given deterministically by $\mathbf{L}\mathbf{s} + \mathbf{m}$. The complete architecture is given in Fig. 2.6.

Appendix B

AFP experiments

B.1 Introduction

A primary aim of the analysis work developed in this thesis was to render the predictions of a basal ganglia-mediated reinforcement learning theory of song copying concrete and testable. This theory is reviewed in detail elsewhere (Section 1.2; Fee and Goldberg (2011)). It predicts that auditory evaluations of rendition-to-rendition acoustic variability guide plasticity in Area X; this plasticity in turn shapes the AFP policy to prefer temporal context-sensitive perturbations associated with positive evaluations. This model predicts that recently acquired adaptive changes to song acoustics result from the adapted AFP policy. Consequently, manipulations that transiently remove AFP perturbations will block the expression of recently learned song changes. This prediction has been confirmed in adult birds learning to adapt their song pitch in response to pitch-contingent experimental punishment: recently acquired pitch shifts are eliminated when the influence of LMAN on RA is blocked pharmacologically (Andalman et al. (2009), Warren et al. (2011)).

In conjunction with developing these analyses, I conducted multiple experiments

aimed to test the juvenile premotor function of the AFP. These experiments do not support the reinforcement learning theory outlined above, but I do not regard these results as conclusive owing to experimental limitations I will describe. Finally, I present avenues to address or circumvent these limitations in future work.

B.2 Results

B.2.1 Area X optogenetics

As explained above, bias in AFP policy has been explored in adults through manipulations of LMAN (Andalman et al. (2009), Warren et al. (2011)). The more complex behavior of juveniles motivated the use of optogenetics instead of pharmacology because optogenetic tools could permit better coverage of developmental age through interleaved sampling of manipulated and unmanipulated song. However, I could not reliably express optogenetic tools in LMAN using AAV vectors. (The resistance of LMAN to infection with AAVs and Lentivirus was confirmed by Carlos Lois in personal communication.) Because AFP bias is thought to depend on Area X activity under the control of HVC_X activity, I sought to manipulate Area X activity instead.

Recently, my lab used ArchT to inactivate Area X neurons in adult birds. This manipulation reduced rendition-by-rendition acoustic variability (Singh Alvarado et al. (2021)). I sought to express ArchT in Area X neurons generally (using AAV2.9-CAG-ArchT.GFP injection, $n=3$) or spiny neurons specifically (using AAV2.9-CaMKii-ArchT.GFP injection, $n=2$). In either case, 3 weeks after injection I recorded multiunit spiking responses while delivering ArchT-activating light (532nm wavelength) to Area X. In most recordings, I observed firing rate increases during laser presentation, consistent with light-induced changes in circuit dynamics. In a minority of cases, I observed acute inactivation consistent with direct recording from an ArchT-expressing neuron. Despite interpretative challenges associated with light-evoked firing rate increases in ArchT-expressing tissue, birds with bilateral light-evoked re-

sponses of either type were implanted with optic fiber stubs in Area X (n=5).

Following this implantation surgery between 50 and 55dph, I recorded birds' vocalizations continuously. After post-operative song rates returned to normal levels, I recorded vocal audio from several days of unmanipulated singing. Using this baseline data, I created a highly permissive song detection template (using software from Tumer and Brainard (2007)) to trigger frequently during juvenile song. A random minority (15%) of files were preselected as 'laser trials.' When writing these files, song detections elicited 2 seconds of continuous laser (15 mW power at input to implant, following Singh Alvarado et al. (2021)) delivery to Area X, followed by a forced 3 seconds of laser inactivity regardless of song detection events. When writing the remaining majority (85%) of files, laser remained off regardless of song detection events. This regime was followed every 2 or 3 days, with intervening days consisting entirely of laser-free singing, until late plastic song (~80dph). I continued to collect song data until adulthood (>95dph)

Following behavioral data collection, I trained syllable VAEs and classified syllable types without reference to manipulation status, using methods described earlier. Subsequently, syllable renditions were assigned to one of three manipulation categories. Assuming a 30ms lag between activity in Area X and downstream effects on song acoustics (Kao et al. (2005)), I determined whether an Area X premotor window was unmanipulated for the duration of syllable production ("Laser Off" rendition), whether laser light was applied for the duration of that premotor window ("Laser On" rendition), or whether the laser onset or offset occurred during the premotor window for a syllable ("Partial" rendition). Partial renditions were discarded from subsequent analysis. 80% of Laser Off renditions were used for predicted age model training.

I used trained predicted age networks to calculate predicted ages for held-out Laser Off renditions, as well as all Laser On renditions. The quantify the impact of

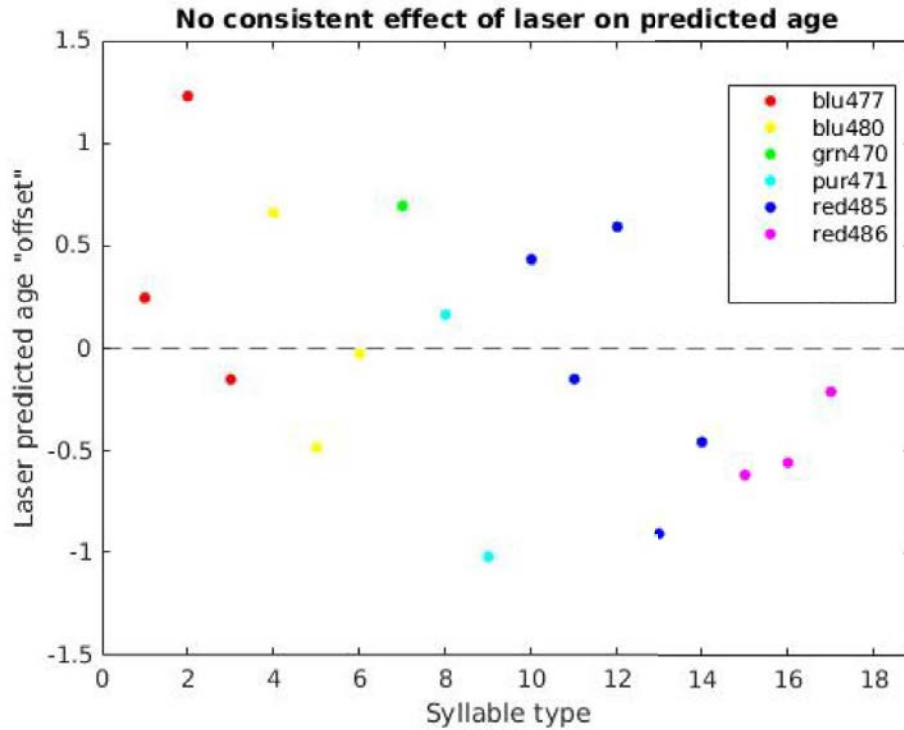


FIGURE B.1: No consistent effect of laser on predicted age **A** Laser “offset” term from linear models depicted for every syllable in the dataset ($n=5$ birds). Fits are distributed about equally in positive and negative directions.

the manipulation on predicted age, I modeled the predicted age of these renditions for each syllable in the dataset (18 syllables from 5 animals). In particular, I found the least square error linear model of predicted age, z , in terms of real production age, x , and laser status, y :

$$z = \beta_0 + \beta_1 x + \beta_2 y, \text{ where } y = \begin{cases} 0 & \text{if Laser Off,} \\ 1 & \text{if Laser On} \end{cases}$$

In this model, β_2 represents contribution of manipulation status to measured syllable maturity. Negative β_2 values are consistent with the hypothesis that inac-

tivating Area X reduces syllable maturity by eliminating the learned modulation of AFP perturbation by timing context. Figure B.1 depicts the β_2 value calculated for each syllable in the dataset, with marker color indicating bird identity. The observed β_2 values do not indicate any tendency to exhibit positive or negative sign.

B.3 Discussion

B.3.1 Interpreting the null result

I did not find a consistent effect of my Area X manipulation on predicted age in the analysis described. Two primary alternatives are possible. First, the results are consistent with the possibility that the reinforcement learning theory of the AFP is incorrect. Second, the theory could be true, with the experiment failing for technical reasons. Unfortunately it is difficult to assess the first possibility directly; we can only methodically assess the the plausibility of the second possibility. If the AFP hypothesis is true, there are two general, non-exclusive types of experimental failure that could prevent detecting an effect. First, despite my best efforts to develop relevant analytic tools in my thesis work, the analysis framework I used could be insensitive to a real behavioral effect. Second, the null result could owe to failures of the biological manipulation approach. I will discuss these possibilities in turn.

There are a handful of modeling components to consider. If the VAE dimensionality reduction fails to encode acoustic features under the control of AFP perturbation and learning, that failure would preclude all downstream analysis from detecting an effect of Area X inactivation. Notwithstanding limitations of the VAE noted in Section 5.4.1, I consider this possibility implausible. First, in Chapter 2 I demonstrate the utility of the autoencoder method to encode copied acoustic features in a cohort of pupil/tutor pairs. In virtue of being learned, these features must be controllable by the AFP in the reinforcement learning framework under consideration. In downstream analysis, relevant variation in latent space should be encoded by predicted

age to the extent that changes in song over time are learned by AFP-dependent reinforcement. Although this consideration indicates that predicted age will encode relevant acoustic variation, irrelevant acoustic variation may also contribute to predicted age, decreasing the power of our analyses. This point is discussed in Section 4.3. As noted there, however, the ability to detect reduced within-day increases in predicted age on days with Area X D1R antagonist microdialysis motivates the conclusion that predicted age is reasonably sensitive to song changes that require AFP mechanisms. Finally, the quantification of the manipulation effect as β_2 is only meaningful to the extent the linear model of predicted age is reasonable. While the linear model does not capture all the dynamics of predicted age — including circadian dynamics described in 4.2.3 — it has the virtue of being relatively simple. It seems like a straightforward way to operationalize the simple theoretical prediction that predicted age will be numerically reduced by Area X inactivation.

In addition to analysis considerations, failures of the biological manipulation could lead to null results even if the AFP-based learning hypothesis is true. In general AAV vector-driven expression of transgenic tools is less reliable and less routine in finches than in mouse, with less established protocols. The injection and implant strategy I used was developed in Singh Alvarado et al. (2021), and executed in consultation with that study’s lead author, Jonna Singh Alvarado. However, that study focused on adult animals, whereas I injected birds at ~ 20 dph. This difference may be significant because Area X adds neurons throughout juvenile development (Nordeen and Nordeen (1988)). Thus only a fraction of the Area X neurons relevant to behavior at the manipulation ages (>60 dph) are available to infect at the injection age, even if the virus infects available cells at high rates. The added neurons seem principally to be spiny neurons (Sohrabji et al. (1993), Rochefort et al. (2007)), so this explanation requires an experimental dependence on high spiny neuron infection rates. That requirement certainly obtains in the case of CaMKii-driven expression

(n=2 animals), since in Area X that promoter drives spiny neuron-specific expression (Hein et al. (2007), Singh Alvarado et al. (2021)) and that cell type must mediate putative optogenetic manipulations. It is less clear how the migration of new neurons into Area X would affect experiments using pan-neuronal expression vectors (CAG promoter, n=3 animals) because the full complement of Area X output neurons is present at the injection age and successful inactivation of these plausibly “overrides” challenges of expression in their Area X afferents; variation in upstream circuit activity cannot impact behavior when Area X efferents are silenced.

To assay the physiological impact of ArchT expression in my preparation, I recorded extracellular spiking in Area X in each hemisphere before chronically implanting fiber stubs. The results of this assay were complex. I recorded laser-evoked activity patterns in all hemispheres, but in only a minority of cases did I observe rapid inhibition. More often, I observed an increased rate of action potentials. This pattern could be accomplished through ArchT-mediated disinhibition of recorded cells — in fact the vast majority of Area X neurons are locally projecting inhibitory cells. Nonetheless I observed this result in every bird injected with CAG-driven Arch plasmid, conflicting with the baseline prediction that this manipulation would reduce firing pan-neuronally. Certainly this result speaks to complex laser-induced dynamics that complicate interpretation of the manipulation. Moreover, my results are qualitatively consistent with the electrophysiological verifications in Singh Alvarado et al. (2021), where laser induced inhibition in some recordings and activation and others; however, laser drove direct inhibition more often in that work, with activation occurring more rarely (personal communication with Jonna Singh Alvarado).

Ideally, the biological manipulation would be verified in singing birds using a measure independent from the hypothesized effect on predicted age. In principle, this verification could involve simultaneous electrophysiology from Area X, but in practice this added recording channel increases the technical challenges of the ex-

periment considerably. A plausible alternative approach involves measuring song variability during laser presentation. Certainly some AFP manipulations — LMAN inactivations for example — are expected to reduce song variability, providing an independent behavioral measure against which to assess manipulation success. Unfortunately, the effect of Area X inhibition on juvenile song variability is uncertain. Recent experiments in adults indicate that variable Area X activity plays a role in generating rendition-to-rendition acoustic variability in song (Singh Alvarado et al. (2021)). However, Area X lesions in juveniles suggest that the nucleus is not required for plastic song variability or the generation of high-variability subsong (Goldberg and Fee (2011)). Nonetheless, to permit assessments of optogenetic manipulation on syllable acoustic distributions, I have extended the Gaussian distribution networks developed in Chapter 3. In particular, I developed networks that take an additional input reflecting manipulation status (Laser On vs Laser Off). These networks find the most likely Gaussian distribution parameters for syllable distributions as a function of age and conditioned on laser status. Reduced song variability with Area X inactivation should manifest as reduced entropy for distributions generated with the manipulation input set to Laser On, and greater entropy for distributions generated with the manipulation input set to Laser Off. Testing this prediction is a key next analysis step with the potential to aid interpretation of these data. Reduced variability would be a strong indicator of the success of the manipulation; no effect on variability is hard to interpret in light of conflicting reports of the role of Area X in generating variability.

B.3.2 Alternative approaches for future work

The challenges laid out above prevent me from ruling out failures of the biological manipulation. In future work, several alternative approaches to block the hypothesized AFP policy remain viable. Recently a tool for improved optogenetic inhibition

of axon terminals was developed by Copits et al. (2021). This raises the prospect of preventing the expression of the AFP perturbation policy by removing the influence of temporal context, communicated by HVC_X axons, on AFP activity. This approach is attractive because HVC has proved more amenable to AAV-driven expression than many other song system nuclei. Like Area X, HVC adds neurons during development (Nordeen and Nordeen (1988)) but these may be principally HVC_{RA} projection neurons and GABAergic interneurons (Scott and Lois (2007)), indicating the HVC_X projection neuron class is available for infection early. Alternatively, neuron excitation with channelrhodopsin has been used more widely in the finch than optogenetic inhibition (Hisey et al. (2018), Kearney et al. (2019), Zhao et al. (2019)) and may work more reliably. Because Area X pallidal spikes completely eliminate spiking in their target DLM neurons for several milliseconds (Goldberg et al. (2012)), it is conceivable to imagine eliminating DLM firing with tetanic optogenetic terminal stimulation of pallidal terminals in DLM. Finally, LMAN inhibition remains the most interpretable experimental manipulation. The challenge in its infection may arise from its dense myelination (Carlos Lois and Mingshan Xue, personal communication). However, this pattern of myelination arises at around 20dph and is significantly reduced at even 15dph (Champoux et al. (2021)), so early injections might enable greater infection.

Bibliography

- Aamodt, S. M., Nordeen, E. J., and Nordeen, K. W. (1996), “Blockade of NMDA receptors during song model exposure impairs song development in juvenile zebra finches.” *Neurobiology of learning and memory*, 65, 91–98.
- Ali, F., Otchy, T. M., Pehlevan, C., Fantana, A. L., Burak, Y., and Ölveczky, B. P. (2013), “The basal ganglia is necessary for learning spectral, but not temporal, features of birdsong.” *Neuron*, 80, 494–506.
- Andalman, A. S., Fee, M. S., and Nottebohm, F. (2009), “A Basal Ganglia-Forebrain Circuit in the Songbird Biases Motor Output to Avoid Vocal Errors,” *Proceedings of the National Academy of Sciences of the United States of America*, 106, 12518–12523.
- Aronov, D., Andalman, A. S., and Fee, M. S. (2008), “A Specialized Forebrain Circuit for Vocal Babbling in the Juvenile Songbird,” *Science*, 320, 630–634.
- Aronov, D., Veit, L., Goldberg, J. H., and Fee, M. S. (2011), “Two distinct modes of forebrain circuit dynamics underlie temporal patterning in the vocalizations of young songbirds.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31, 16353–16368.
- Azzalini, A. (2005), “The Skew-Normal Distribution and Related Multivariate Families,” *Scandinavian Journal of Statistics*, 32, 159–188.
- Azzalini, A. and Valle, A. D. (1996), “The Multivariate Skew-Normal Distribution,” *Biometrika*, 83, 715–726.
- Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983), “Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems,” *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13, 834–846.
- Bottjer, S. W., Miesner, E. A., and ARNOLD, A. P. (1984), “Forebrain Lesions Disrupt Development but Not Maintenance of Song in Passerine Birds,” *Science*, 224, 901–903.

- Bottjer, S. W., Halsema, K. A., Brown, S. A., and Miesner, E. A. (1989), “Axonal connections of a forebrain nucleus involved with vocal learning in zebra finches.” *The Journal of comparative neurology*, 279, 312–326.
- Castañeda, T. R., de Prado, B. M., Prieto, D., and Mora, F. (2004), “Circadian rhythms of dopamine, glutamate and GABA in the striatum and nucleus accumbens of the awake rat: modulation by light.” *Journal of pineal research*, 36, 177–185.
- Castelino, C. B. and Ball, G. F. (2005), “A role for norepinephrine in the regulation of context-dependent ZENK expression in male zebra finches (*Taeniopygia guttata*).” *The European journal of neuroscience*, 21, 1962–1972.
- Champion, K., Lusch, B., Kutz, N., and Brunton, S. (2019), “Autoencoders for discovering coordinates and dynamics from data,” in *APS Division of Fluid Dynamics Meeting Abstracts*, p. P10.006.
- Champoux, K. L., Miller, K. E., and Perkel, D. J. (2021), “Differential development of myelin in zebra finch song nuclei.” *Journal of Comparative Neurology*, 529, 1255–1265.
- Charlesworth, J. D., Tumer, E. C., Warren, T. L., and Brainard, M. S. (2011), “Learning the microstructure of successful behavior,” *Nature Neuroscience*, 14, 373–380.
- Chi, Z. and Margoliash, D. (2001), “Temporal precision and temporal drift in brain and behavior of zebra finch song.” *Neuron*, 32, 899–910.
- Clayton, N. S. (1989), “The Effects of Cross-Fostering on Selective Song Learning in Estrildid Finches,” *Behaviour*, 109, 163–175.
- Cohen, Y., Shen, J., Semu, D., Leman, D. P., Liberti, W. A., Perkins, L. N., Liberti, D. C., Kotton, D. N., and Gardner, T. J. (2020a), “Hidden neural states underlie canary song syntax.” *Nature*, 582, 539–544.
- Cohen, Y., Nicholson, D., Sanchioni, A., Mallaber, E. K., Skidanova, V., and Gardner, T. J. (2020b), “TweetyNet: A neural network that enables high-throughput, automated annotation of birdsong,” *bioRxiv*, p. 2020.08.28.272088.
- Colquitt, B. M., Merullo, D. P., Konopka, G., Roberts, T. F., and Brainard, M. S. (2021), “Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits,” *Science*, 371.
- Copits, B. A., Gowrishankar, R., O’Neill, P. R., Li, J.-N., Girven, K. S., Yoo, J. J., Meshik, X., Parker, K. E., Spangler, S. M., Elerding, A. J., Brown, B. J., Shirley, S. E., Ma, K. K. L., Vasquez, A. M., Stander, M. C., Kalyanaraman, V., Vogt,

- S. K., Samineni, V. K., Patriarchi, T., Tian, L., Gautam, N., Sunahara, R. K., Gereau, R. W., and Bruchas, M. R. (2021), “A photoswitchable GPCR-based opsin for presynaptic inhibition.” *Neuron*, 109, 1791–1809.e11.
- Curtiss, S., Fromkin, V., Krashen, S., Rigler, D., and Rigler, M. (1974), “The Linguistic Development of Genie,” *Language*, 50, 528–554.
- Dan Foresee, F. and Hagan, M. T. (1997), “Gauss-Newton approximation to bayesian learning,” in *IEEE International Conference on Neural Networks - Conference Proceedings*, pp. 1930–1935, Nokia Corporation, Espoo, Finland.
- De Cao, N., Aziz, W., and Titov, I. (2020), “Block Neural Autoregressive Flow,” in *Uncertainty in Artificial Intelligence*, pp. 1263–1273, PMLR.
- Derégnaucourt, S., Mitra, P. P., Feher, O., Pytte, C., and Tchernichovski, O. (2005), “How sleep affects the developmental learning of bird song,” *Nature*, 433, 710–716.
- Derégnaucourt, S., Poirier, C., Kant, A. V. d., Linden, A. V. d., and Gahr, M. (2013), “Comparisons of different methods to train a young zebra finch (*Taeniopygia guttata*) to learn a song.” *Journal of physiology, Paris*, 107, 210–218.
- Ding, L. and Perkel, D. J. (2002), “Dopamine modulates excitability of spiny neurons in the avian basal ganglia.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22, 5210–5218.
- Doersch, C. (2016), “Tutorial on Variational Autoencoders,” *arxiv*.
- Doya, K. and Sejnowski, T. (1996), “A Novel Reinforcement Model of Birdsong Vocalization Learning,” *Advances in neural information processing systems*, 7.
- Farries, M. A. and Perkel, D. J. (2002), “A Telencephalic Nucleus Essential for Song Learning Contains Neurons with Physiological Characteristics of Both Striatum and Globus Pallidus,” *Journal of Neuroscience*, 22, 3776–3787.
- Farries, M. A., Ding, L., and Perkel, D. J. (2005), “Evidence for “direct” and “indirect” pathways through the song system basal ganglia,” *Journal of Comparative Neurology*, 484, 93–104.
- Fee, M. S. and Goldberg, J. H. (2011), “A hypothesis for basal ganglia-dependent reinforcement learning in the songbird,” *Neuroscience*.
- Ferris, M. J., España, R. A., Locke, J. L., Konstantopoulos, J. K., Rose, J. H., Chen, R., and Jones, S. R. (2014), “Dopamine transporters govern diurnal variation in extracellular dopamine tone,” *Proceedings of the National Academy of Sciences of the United States of America*, 111, E2751–E2759.

- Fiete, I. R., Fee, M. S., and Seung, H. S. (2007), “Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances.” *Journal of neurophysiology*, 98, 2038–2057.
- Fischer, S., Hallschmid, M., Elsner, A. L., and Born, J. (2002), “Sleep Forms Memory for Finger Skills,” *Proceedings of the National Academy of Sciences of the United States of America*, 99, 11987–11991.
- Franz, M. and Goller, F. (2002), “Respiratory units of motor production and song imitation in the zebra finch,” *Journal of neurobiology*, 51, 129–141.
- Gadagkar, V., Puzerey, P. A., Chen, R., Baird-Daniel, E., Farhang, A. R., and Goldberg, J. H. (2016), “Dopamine neurons encode performance error in singing birds.” *Science*, 354, 1278–1282.
- Garst-Orozco, J., Babadi, B., and Ölveczky, B. P. (2014), “A neural circuit mechanism for regulating vocal variability during song learning in zebra finches.” *Elife*, 3, e03697.
- Goffinet, J., Brudner, S., Mooney, R., and Pearson, J. (2021), “Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires.” *Elife*, 10.
- Goldberg, J. H. and Fee, M. S. (2010), “Singing-Related Neural Activity Distinguishes Four Classes of Putative Striatal Neurons in the Songbird Basal Ganglia,” *Journal of neurophysiology*, 103, 2002–2014.
- Goldberg, J. H. and Fee, M. S. (2011), “Vocal babbling in songbirds requires the basal ganglia-recipient motor thalamus but not the basal ganglia.” *Journal of neurophysiology*, 105, 2729–2739.
- Goldberg, J. H., Adler, A., Bergman, H., and Fee, M. S. (2010), “Singing-Related Neural Activity Distinguishes Two Putative Pallidal Cell Types in the Songbird Basal Ganglia: Comparison to the Primate Internal and External Pallidal Segments,” *Journal of Neuroscience*, 30, 7088–7098.
- Goldberg, J. H., Farries, M. A., and Fee, M. S. (2012), “Integration of cortical and pallidal inputs in the basal ganglia-recipient thalamus of singing birds.” *Journal of neurophysiology*, 108, 1403–1429.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schoelkopf, B., and Smola, A. (2012), “A Kernel Two-Sample Test,” *Journal of Machine Learning Research*, 13, 723–773.
- Haesler, S., Rochefort, C., Georgi, B., Licznarski, P., Osten, P., and Scharff, C. (2007), “Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X.” *PLoS Biol*, 5, e321–2897.

- Hahnloser, R. H. R., Kozhevnikov, A. A., and Fee, M. S. (2002), “An ultra-sparse code underlies the generation of neural sequences in a songbird.” *Nature*, 419, 65–70.
- Hein, A. M., Sridharan, A., Nordeen, K. W., and Nordeen, E. J. (2007), “Characterization of CaMKII-expressing neurons within a striatal region implicated in avian vocal learning.” *Brain research*, 1155, 125–133.
- Heston, J. B. and White, S. A. (2015), “Behavior-linked FoxP2 regulation enables zebra finch vocal learning.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35, 2885–2894.
- Hisey, E., Kearney, M. G., and Mooney, R. (2018), “A common neural circuit mechanism for internally guided and externally reinforced forms of motor learning,” *Nature Neuroscience*, 21, 589–597.
- Immelmann, K. (1969), “Song development in the zebra finch and other estrildid finches,” *Bird Vocalizations*, pp. 61–77.
- Iyengar, S., Viswanathan, S. S., and Bottjer, S. W. (1999), “Development of topography within song control circuitry of zebra finches during the sensitive period for song learning.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 19, 6037–6057.
- Kao, M. H., Doupe, A. J., and Brainard, M. S. (2005), “Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song.” *Nature*, 433, 638–643.
- Kearney, M. G., Warren, T. L., Hisey, E., Qi, J., and Mooney, R. (2019), “Discrete Evaluative and Premotor Circuits Enable Vocal Learning in Songbirds.” *Neuron*, 104, 559–575.e6.
- Keller, G. B. and Hahnloser, R. H. R. (2008), “Neural processing of auditory feedback during vocal practice in a songbird,” *Nature*, 457, 187–191.
- Kingma, D. P. and Welling, M. (2013), “Auto-Encoding Variational Bayes,” *arxiv*.
- Kojima, S. and Doupe, A. J. (2011), “Social performance reveals unexpected vocal competency in young songbirds.” *Proceedings of the National Academy of Sciences of the United States of America*, 108, 1687–1692.
- Kojima, S., Kao, M. H., and Doupe, A. J. (2013), “Task-related ”cortical” bursting depends critically on basal ganglia input and is linked to vocal plasticity.” *Proceedings of the National Academy of Sciences of the United States of America*, 110, 4756–4761.

- Kojima, S., Kao, M. H., Doupe, A. J., and Brainard, M. S. (2018), “The Avian Basal Ganglia Are a Source of Rapid Behavioral Variation That Enables Vocal Motor Exploration.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 38, 9635–9647.
- Kollmorgen, S., Hahnloser, R. H. R., and Mante, V. (2020), “Nearest neighbours reveal fast and slow components of motor learning.” *Nature*, 577, 526–530.
- Konishi, M. (1965), “The role of auditory feedback in the control of vocalization in the white-crowned sparrow.” *Zeitschrift fur Tierpsychologie*, 22, 770–783.
- Korman, M., Doyon, J., Doljansky, J., Carrier, J., Dagan, Y., and Karni, A. (2007), “Daytime sleep condenses the time course of motor memory consolidation,” *Nature Neuroscience*, 10, 1206–1213.
- Kozhevnikov, A. A. and Fee, M. S. (2007), “Singing-Related Activity of Identified HVC Neurons in the Zebra Finch,” *Journal of neurophysiology*, 97, 4271–4283.
- Kubikova, L., Wada, K., and Jarvis, E. D. (2010), “Dopamine receptors in a songbird brain,” *Journal of Comparative Neurology*, 518, 741–769.
- Leblois, A., Wendel, B. J., and Perkel, D. J. (2010), “Striatal dopamine modulates basal ganglia output and regulates social context-dependent behavioral variability through D1 receptors.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30, 5730–5743.
- Leonardo, A. (2004), “Experimental test of the birdsong error-correction model.” *Proceedings of the National Academy of Sciences*, 101, 16935–16940.
- Lewis, J. W., Ryan, S. M., ARNOLD, A. P., and Butcher, L. L. (1981), “Evidence for a catecholaminergic projection to area X in the zebra finch.” *The Journal of comparative neurology*, 196, 347–354.
- Liu, W.-C., Gardner, T. J., and Nottebohm, F. (2004), “Juvenile Zebra Finches Can Use Multiple Strategies to Learn the Same Song,” *Proceedings of the National Academy of Sciences of the United States of America*, 101, 18177–18182.
- Livingston, F. S. and Mooney, R. (2001), “Androgens and isolation from adult tutors differentially affect the development of songbird neurons critical to vocal plasticity.” *Journal of neurophysiology*, 85, 34–42.
- Lombardino, A. J. and Nottebohm, F. (2000), “Age at deafening affects the stability of learned song in adult male zebra finches.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20, 5054–5064.
- Long, M. A. and Fee, M. S. (2008), “Using temperature to analyse temporal dynamics in the songbird motor pathway,” *Nature*, 456, 189–194.

- Luo, M. and Perkel, D. J. (1999a), “A GABAergic, strongly inhibitory projection to a thalamic nucleus in the zebra finch song system,” *Journal of Neuroscience*, 19, 6700–6711.
- Luo, M. and Perkel, D. J. (1999b), “Long-range GABAergic projection in a circuit essential for vocal learning,” *Journal of Comparative Neurology*, 403, 68–84.
- Luo, M., Ding, L., and Perkel, D. J. (2001), “An avian basal ganglia pathway essential for vocal learning forms a closed topographic loop,” *Journal of Neuroscience*, 21, 6836–6845.
- Mackevicius, E. L., Bahle, A. H., Williams, A. H., Gu, S., Denisenko, N. I., Goldman, M. S., and Fee, M. S. (2019), “Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience.” *Elife*, 8.
- Mandelblat-Cerf, Y. and Fee, M. S. (2014), “An Automated Procedure for Evaluating Song Imitation,” *PLOS ONE*, 9, e96484–13.
- Mandelblat-Cerf, Y., Las, L., Denisenko, N., and Fee, M. S. (2014), “A role for descending auditory cortical projections in songbird vocal learning.” *Elife*, 3.
- Marler, P. (1970), “A comparative approach to vocal learning: Song development in white-crowned sparrows,” *Journal of Comparative and Physiological Psychology*, 71, 1–25.
- Marler, P. and Waser, M. S. (1977), “Role of auditory feedback in canary song development.” *Journal of Comparative and Physiological Psychology*, 91, 8–16.
- McInnes, L., Healy, J., and Melville, J. (2020), “UMAP: uniform manifold approximation and projection for dimension reduction,” *arxiv*.
- Miller, J. E., Hilliard, A. T., and White, S. A. (2010), “Song practice promotes acute vocal variability at a key stage of sensorimotor learning.” *PLOS ONE*, 5, e8592.
- Mooney, R. and Konishi, M. (1991), “Two Distinct Inputs to an Avian Song Nucleus Activate Different Glutamate Receptor Subtypes on Individual Neurons,” *Proceedings of the National Academy of Sciences of the United States of America*, 88, 4075–4079.
- Morris, D. (1954), “The Reproductive Behaviour of the Zebra Finch (*Poephila guttata*), with Special Reference to Pseudofemale Behaviour and Displacement Activities,” *Behaviour*, 6, 271–322.
- Murugan, M., Harward, S., Scharff, C., and Mooney, R. (2013), “Diminished FoxP2 levels affect dopaminergic modulation of corticostriatal signaling important to song variability.” *Neuron*, 80, 1464–1476.

- Nordeen, E. J. and Nordeen, K. W. (1988), “Sex and regional differences in the incorporation of neurons born during song learning in zebra finches.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 8, 2869–2874.
- Nordeen, K. W. and Nordeen, E. J. (1992), “Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches,” *Behavioral and neural biology*, 57, 58–66.
- Nottebohm, F., Stokes, T. M., and Leonard, C. M. (1976), “Central control of song in the canary, *Serinus canarius*.” *The Journal of comparative neurology*, 165, 457–486.
- Nottebohm, F., Kelley, D. B., and Paton, J. A. (1982), “Connections of vocal control nuclei in the canary telencephalon.” *The Journal of comparative neurology*, 207, 344–357.
- Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F., and Fee, M. S. (2015), “Growth and splitting of neural sequences in songbird vocal development,” *Nature*, 528, 352–357.
- Okuhata, S. and Saito, N. (1987), “Synaptic connections of thalamo-cerebral vocal nuclei of the canary.” *Brain research bulletin*, 18, 35–44.
- Ölveczky, B. P., Andalman, A. S., and Fee, M. S. (2005), “Vocal experimentation in the juvenile songbird requires a basal ganglia circuit,” *PLoS biology*.
- Ölveczky, B. P., Otchy, T. M., Goldberg, J. H., Aronov, D., and Fee, M. S. (2011), “Changes in the neural control of a complex motor sequence during learning.” *Journal of neurophysiology*, 106, 386–397.
- Price, P. H. (1979), “Developmental determinants of structure in zebra finch song.” *Journal of Comparative and Physiological Psychology*, 93, 260–277.
- Ravbar, P., Lipkind, D., Parra, L. C., and Tchernichovski, O. (2012), “Vocal exploration is locally regulated during song learning.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32, 3422–3432.
- Reiner, A., Laverghetta, A. V., Meade, C. A., Cuthbertson, S. L., and Bottjer, S. W. (2004), “An immunohistochemical and pathway tracing study of the striatopallidal organization of area X in the male zebra finch.” *The Journal of comparative neurology*, 469, 239–261.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014), “Stochastic backpropagation and variational inference in deep latent gaussian models,” *arxiv*.
- Roberts, T. F., Tschida, K. A., Klein, M. E., and Mooney, R. (2010), “Rapid spine stabilization and synaptic enhancement at the onset of behavioural learning.” *Nature*, 463, 948–952.

- Rocheffort, C., He, X., Scotto-Lomassese, S., and Scharff, C. (2007), “Recruitment of FoxP2-expressing neurons to area X varies during song development.” *Developmental neurobiology*, 67, 809–817.
- Scharff, C. and Nottebohm, F. (1991), “A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: Implications for vocal learning,” *Journal of Neuroscience*, 11, 2896–2913.
- Schultz, W., Dayan, P., and Montague, P. R. (1997), “A neural substrate of prediction and reward.” *Science*, 275, 1593–1599.
- Scott, B. B. and Lois, C. (2007), “Developmental origin and identity of song system neurons born during vocal learning in songbirds.” *The Journal of comparative neurology*, 502, 202–214.
- Singh, T. D., Nordeen, E. J., and Nordeen, K. W. (2005), “Song tutoring triggers CaMKII phosphorylation within a specialized portion of the avian basal ganglia,” *Journal of neurobiology*, 65, 179–191.
- Singh Alvarado, J., Goffinet, J., Michael, V., Liberti, W., Hatfield, J., Gardner, T., Pearson, J., and Mooney, R. (2021), “Neural dynamics underlying birdsong practice and performance.” *Nature*, pp. 1–5.
- Sober, S. J. and Brainard, M. S. (2009), “Adult birdsong is actively maintained by error correction,” *Nature Neuroscience*, 12, 927–931.
- Sober, S. J., Wohlgemuth, M. J., and Brainard, M. S. (2008), “Central contributions to acoustic variation in birdsong.” *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28, 10370–10379.
- Sohrabji, F., Nordeen, E. J., and Nordeen, K. W. (1990), “Selective Impairment of Song Learning Following Lesions of a Forebrain Nucleus in the Juvenile Zebra Finch,” *Behavioral and neural biology*, 53, 51–63.
- Sohrabji, F., Nordeen, E. J., and Nordeen, K. W. (1993), “Characterization of neurons born and incorporated into a vocal control nucleus during avian song learning.” *Brain research*, 620, 335–338.
- Sossinka, R. and Böhner, J. (1980), “Song Types in the Zebra Finch *Poephila guttata castanotis*,” *Zeitschrift für Tierpsychologie*, 53, 123–132.
- Striedter, G. F. (1997), “The telencephalon of tetrapods in evolution.” *Brain, behavior and evolution*, 49, 179–213.
- Tchernichovski, O., Nottebohm, F., Ho, C., Pesaran, B., and Mitra, P. (2000), “A procedure for an automated measurement of song similarity.” *Animal behaviour*, 59, 1167–1176.

- Tchernichovski, O., Mitra, P. P., Lints, T., and Nottebohm, F. (2001), “Dynamics of the Vocal Imitation Process: How a Zebra Finch Learns Its Song,” *Science*, 291, 2564–2569.
- Tchernichovski, O., Eisenberg-Edidin, S., and Jarvis, E. D. (2021), “Balanced imitation sustains song culture in zebra finches.” *Nature Communications*, 12, 2562.
- Teramitsu, I., Poopatanapong, A., Torrisi, S., and White, S. A. (2010), “Striatal FoxP2 is actively regulated during songbird sensorimotor learning.” *PLOS ONE*, 5, e8548.
- Teşileanu, T., Ölveczky, B., and Balasubramanian, V. (2017), “Rules and mechanisms for efficient two-stage learning in neural circuits.” *Elife*, 6.
- Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., and Deisseroth, K. (2009), “Phasic Firing in Dopaminergic Neurons Is Sufficient for Behavioral Conditioning,” *Science*, 324, 1080–1084.
- Tschida, K. A. and Mooney, R. (2012), “Deafening Drives Cell-Type-Specific Changes to Dendritic Spines in a Sensorimotor Nucleus Important to Learned Vocalizations,” *Neuron*, 73, 1028–1039.
- Tumer, E. C. and Brainard, M. S. (2007), “Performance variability enables adaptive plasticity of —[lsquo]—crystallized—[rsquo]— adult birdsong,” *Nature*, 450, 1240–1244.
- Vates, G. E. and Nottebohm, F. (1995), “Feedback circuitry within a song-learning pathway,” *Proceedings of the National Academy of Sciences of the United States of America*, 92, 5139–5143.
- Veit, L., Aronov, D., and Fee, M. S. (2011), “Learning to breathe and sing: development of respiratory-vocal coordination in young songbirds,” *Journal of neurophysiology*, 106, 1747–1765.
- Vicario, D. S. and Nottebohm, F. (1988), “Organization of the zebra finch song control system: I. Representation of syringeal muscles in the hypoglossal nucleus.” *The Journal of comparative neurology*, 271, 346–354.
- Walker, M. P., Brakefield, T., Hobson, J. A., and Stickgold, R. (2003), “Dissociable stages of human memory consolidation and reconsolidation.” *Nature*, 425, 616–620.
- Warren, T. L., Tumer, E. C., Charlesworth, J. D., and Brainard, M. S. (2011), “Mechanisms and time course of vocal learning and consolidation in the adult songbird,” *Journal of neurophysiology*, 106, 1806–1821.
- Wild, J. M. (1993), “Descending projections of the songbird nucleus robustus archistriatalis.” *The Journal of comparative neurology*, 338, 225–241.

- Woolley, S. C. and Doupe, A. J. (2008), “Social context-induced song variation affects female behavior and gene expression.” *PLoS Biol*, 6, e62–0537.
- Woolley, S. C., Rajan, R., Joshua, M., and Doupe, A. J. (2014), “Emergence of context-dependent variability across a basal ganglia network.” *Neuron*, 82, 208–223.
- Woolley, S. M. N. (2012), “Early experience shapes vocal neural coding and perception in songbirds,” *Developmental Psychobiology*, 54, 612–631.
- Xiao, L., Chattree, G., Oscos, F. G., Cao, M., Wanat, M. J., and Roberts, T. F. (2018), “A Basal Ganglia Circuit Sufficient to Guide Birdsong Learning.” *Neuron*, 98, 208–221.e5.
- Xiao, L., Merullo, D. P., Koch, T. M. I., Cao, M., Co, M., Kulkarni, A., Konopka, G., and Roberts, T. F. (2021), “Expression of FoxP2 in the basal ganglia regulates vocal motor sequences in the adult songbird.” *Nature Communications*, 12, 2617–18.
- Yu, A. C. and Margoliash, D. (1996), “Temporal hierarchical control of singing in birds,” *Science*, 273, 1871–1875.
- Zann, R. (1990), “Song and call learning in wild zebra finches in south-east Australia,” *Animal behaviour*, 40, 811–828.
- Zhao, W., Garcia-Oscos, F., Dinh, D., and Roberts, T. F. (2019), “Inception of memories that guide vocal learning in the songbird.” *Science*, 366, 83–89.
- Zhou, B., Hofmann, D., Pinkoviezky, I., Sober, S. J., and Nemenman, I. (2018), “Chance, long tails, and inference in a non-Gaussian, Bayesian theory of vocal learning in songbirds.” *Proceedings of the National Academy of Sciences of the United States of America*, 115, E8538–E8546.
- Zuschratter, W. and Scheich, H. (1990), “Distribution of choline acetyltransferase and acetylcholinesterase in the vocal motor system of zebra finches.” *Brain research*, 513, 193–201.

Biography

Samuel Navickas Brudner grew up in Brooklyn, New York. He received a BA in Cognitive Science from Yale University in 2012. At Duke University, he coauthored a paper (“Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires”), and is preparing a first-author manuscript based on Chapter 3 and 4 of this thesis work. While at Duke, he received support through the James B. Duke Award, and through a Ruth L. Kirschstein Predoctoral Individual National Research Service Award from NICHD (F31 HD098772-02).