Article

Neural dynamics underlying birdsong practice and performance

https://doi.org/10.1038/s41586-021-04004-1

Received: 8 December 2020

Accepted: 7 September 2021

Published online: 20 October 2021

Check for updates

Jonnathan Singh Alvarado[1], Jack Goffinet[2], Valerie Michael[1], William Liberti III[3], Jordan Hatfield[1], Timothy Gardner[4], John Pearson[1,5,6]✉ & Richard Mooney[1]✉

Musical and athletic skills are learned and maintained through intensive practice to enable precise and reliable performance for an audience. Consequently, understanding such complex behaviours requires insight into how the brain functions during both practice and performance. Male zebra finches learn to produce courtship songs that are more varied when alone and more stereotyped in the presence of females[1]. These differences are thought to reflect song practice and performance, respectively[2,3], providing a useful system in which to explore how neurons encode and regulate motor variability in these two states. Here we show that calcium signals in ensembles of spiny neurons (SNs) in the basal ganglia are highly variable relative to their cortical afferents during song practice. By contrast, SN calcium signals are strongly suppressed during female-directed performance, and optogenetically suppressing SNs during practice strongly reduces vocal variability. Unsupervised learning methods[4,5] show that specific SN activity patterns map onto distinct song practice variants. Finally, we establish that noradrenergic signalling reduces vocal variability by directly suppressing SN activity. Thus, SN ensembles encode and drive vocal exploration during practice, and the noradrenergic suppression of SN activity promotes stereotyped and precise song performance for an audience.

Adult male zebra finches sing more varied songs when alone (that is, during practice) and perform more stereotyped songs for females[1,2] that are also more effective courtship signals[6]. Notably, electrical and immediate early gene activity in a song-specialized part of the basal ganglia (BG) (area X; hereafter referred to as sBG) differs across these two states[3,7,8], and sBG lesions transiently reduce song variability[9]. Nonetheless, how SN ensembles encode variability during practice and how their activity is dynamically regulated to enable stereotyped song performance are unknown. To address these issues, we used a miniature microscope (miniscope) to image calcium activity of SNs expressing GCaMP7f under a CaMKII promoter[10] in freely singing adult male finches (Fig. 1a–c). The CaMKII promoter achieved highly selective expression of GCaMP and other constructs in SNs (Extended Data Figs. 1a, 4d, e). When males sang in social isolation, ensembles of SNs displayed dynamic activity patterns, with the timing and participation of active neurons changing across consecutive song renditions (Fig. 1d, Supplementary Video 1). Qualitatively, these dynamics were visualized by comparing the mean sequential activity in the ensemble across renditions to sequential activity in single renditions (Fig. 1e). We quantified these dynamics using within-day, within-neuron autocorrelations of song-aligned calcium activity, by computing the specificity and sensitivity of individual neuron activity, and by computing the probability that individual neurons participated within the ensemble across renditions (Extended Data Fig. 1b–d). Although SN ensemble activity was dynamic, it was specific to singing: SNs were largely silent during locomotion and other non-vocal movements (Extended Data Fig. 1e–h). Finally, SN activity could precede vocal onset (Fig. 1c, Extended Data Fig. 2), did not respond to song playback during non-vocal epochs, and was unaffected by singing-triggered noise (Extended Data Fig. 1i, j), indicating that SNs encode motor-related activity specifically during singing.

A major source of singing-related activity to the sBG are projection neurons (PNs) in the song premotor nucleus HVC[11]. In contrast to SNs, and consistent with prior work[12,13], the peak timing and participation of HVC PNs was relatively stable across song renditions within a single day (Extended Data Fig. 1c, d). Consequently, the singing-related activity patterns in a HVC PN ensemble within any single song rendition were similar to the mean across renditions (Fig. 1f). Thus, variable SN activity during practice was not an imaging artefact or inherited from HVC.

Imaging SNs as the male switched from practice to female-directed song performance revealed a stark decrease in activity, with calcium signals in the majority of SNs dropping below detection threshold (Fig. 2a–c, Extended Data Fig. 3a–g, Supplementary Video 2). This suppression was unrelated to the male's body movements (Extended Data Fig. 3h–j) and was not attributable to photobleaching, as interleaved presentations of the female reliably and reversibly diminished SN activity during singing (Extended Data Fig. 3e). To localize where this context-dependent switch originated, we used dual-fibre photometry to simultaneously image the activity of HVC neurons and their axons within the sBG (Fig. 2d, Extended Data Fig. 3k, l). Neither HVC cells

[1]Department of Neurobiology, Duke University, Durham, NC, USA. [2]Department of Computer Science, Duke University, Durham, NC, USA. [3]Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA, USA. [4]Phil and Penny Knight Campus for Accelerating Scientific Impact, University of Oregon, Eugene, OR, USA. [5]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA. [6]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. ✉e-mail: john.pearson@duke.edu; mooney@neuro.duke.edu

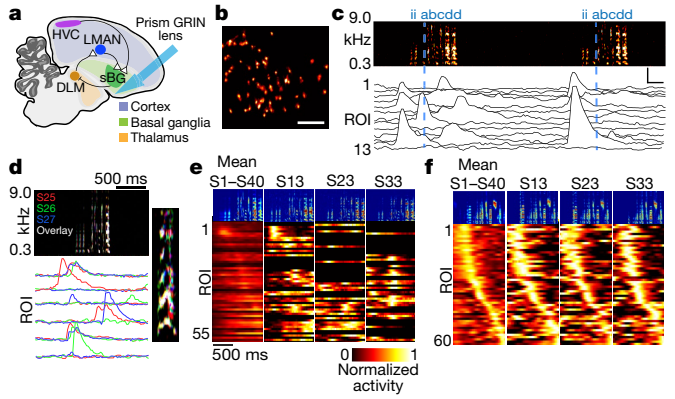


Fig. 1 | SN ensemble activity is song-specific and variable. a, Sagittal view of the finch brain showing prism graded-index (GRIN) lens location. DLM, medial portion of the dorsolateral nucleus of the anterior thalamus. **b**, Example extracted ROIs (regions of interest) from one field of view. Scale bar, 250 µm. **c**, Calcium traces from example SNs (*z*-scored) aligned to motif renditions (syllables: a, b, c, d; introductory notes: i). Dashed lines indicate motif onset. Scale bars = 4 SD < 200 ms. **d**, Example song-related activity for three song motifs performed within 30 s, and magnified view of an example syllable across renditions. Spectrograms and corresponding activity from three renditions pseudocoloured in red, green and blue. **e**, Average and single-trial song-aligned ROI activity for sBG SNs from one bird. ROIs with no detected events across the 40 analysed song renditions are thresholded to 0. **f**, Average and single-trial song-aligned ROI activity for HVC PNs from one bird.

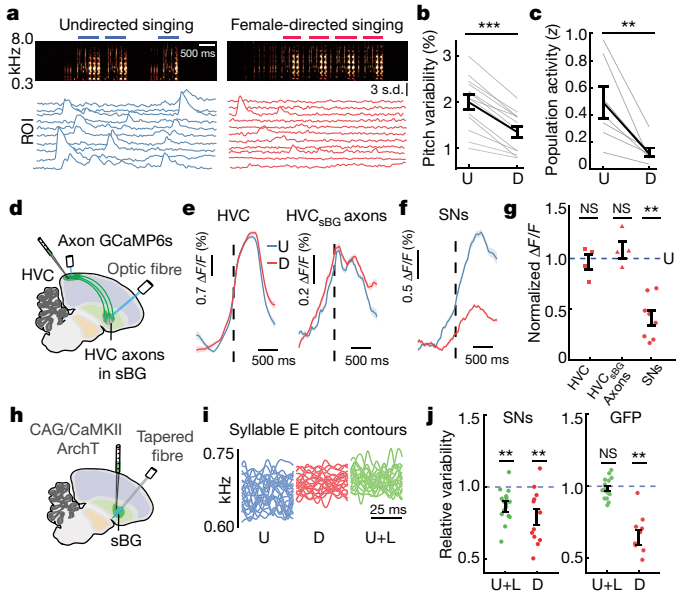


Fig. 2 | SN activity drives vocal variability. a, Example ROI SN activity aligned to undirected (blue) and directed (red) singing (blue and red bars, song motifs). **b**, Pitch variability across undirected (U) and directed (D) singing (Student's two-tailed paired *t*-test, $t_5 = 11.87$, ***$P = 0.000074$, $n = 6$ birds). **c**, Average SN population activity for undirected and directed singing (Student's two-tailed paired *t*-test, $t_6 = 3.74$, **$P = 0.0095$, $n = 6$ birds). **d**, Approach for dual photometric recordings of HVC and HVC$_{sBG}$ axon calcium signals. **e**, Representative photometry traces for HVC and HVC$_{sBG}$ axons for one bird. **f**, Representative photometry traces of SNs for one bird. **g**, Normalized activity for all three recording conditions (Student's two-tailed paired *t*-test; HVC: $t_3 = -0.98$, $P = 0.26$, $n = 4$ birds; HVC$_{sBG}$ axons: $t_3 = -1.38$, $P = 0.39$, $n = 4$ birds; SNs: $t_7 = 3.96$, **$P = 0.0054$, $n = 8$ birds). **h**, Approach for optogenetically suppressing SNs using CamKII.ArchT. **i**, Example pitch contours for one syllable in undirected (U), directed (D) and undirected laser (U + L) conditions. (25 renditions per condition). **j**, Group data showing pitch variability during directed singing normalized to undirected singing. Mixed-effects model, two-sided permutation test. SNs: laser effect size, 14.1% of baseline variability, ***$P = 0.000075$ ($n = 16$ syllables from 6 birds). Directed effect size: 20.9% of baseline, **$P = 0.0056$ ($n = 12$ syllables from 5 birds). GFP: laser effect size, 2% of baseline, $P = 0.2$ ($n = 15$ syllables from 5 birds). Directed effect size = 39.79% of baseline, ***$P = 0.0031$ ($n = 10$ syllables from 4 birds). See Supplementary Table 1 for a detailed description of model outputs. Data are mean ± s.e.m. NS, not significant.

nor their sBG axon terminals showed changes in their singing-related activity as the male alternated between song practice and performance (Fig. 2e, Extended Data Fig. 3l). However, similar to miniscope imaging of individual SNs, photometric recordings revealed that calcium signals in SN populations were strongly suppressed during performance (Fig. 2f, g, Extended Data Fig. 3m). Therefore, changes in SN ensemble activity directly parallel changes in song variability, and these changes are not obvious in HVC or its axons in the sBG.

To test whether diminished SN activity during female-directed performance drives more stereotyped song, we expressed the inhibitory opsin ArchT either selectively in SNs or pan-neuronally in sBG neurons, using CaMKII or CAG promoters, respectively. Post hoc histology confirmed that ArchT expression in pallidal projection neurons was driven by the CAG but not the CaMKII promoter and, with either promoter, illuminating the sBG with green light strongly suppressed neural activity (Fig. 2h, Extended Data Fig. 4a–e). Optogenetically suppressing SN or sBG neuron activity during a random subset (15–30%) of undirected song renditions (Extended Data Fig. 4f) reduced across-rendition variability in the fundamental frequency (pitch) of targeted syllables (Fig 2i, j, Extended Data Fig. 4g, h). Suppressing sBG or SN activity also reduced intra-syllable pitch variation, another feature of female-directed song, without altering syllable mean pitch[9] (Extended Data Fig. 4i, j). Lastly, optogenetic suppression slightly reduced syllable durations, which trended in the same direction as changes observed during female-directed performance[1] (Extended Data Fig. 4k). Thus, suppressing SN or sBG neuron activity during song practice recapitulates several of the acoustic changes that characterize female-directed singing.

These results show that SNs generate dynamic activity patterns that are causally linked to pitch variability. However, zebra finch song comprises several spectrally complex syllables organized into an orderly sequence, or motif, that cannot easily be summarized by this single metric. Thus, we examined the neural code underlying song variability by modelling the relationships between sBG ensemble activity and acoustic structure in practice song motifs. We first used the variational autoencoder[14,15] (VAE), an unsupervised machine learning model, to compress sound spectrogram data ($n \approx 16{,}000$ pixels) to low-dimensional ($n = 32$) latent representations (Extended Data Fig. 5a). These latent dimensions constitute a parsimonious description of song data that nonetheless preserves the rich structure of the original spectrograms[4] (for a similar approach, see ref. [5]). We found that pairs of motifs with more correlated region-of-interest (ROI) activity also tended to have higher acoustic similarity as quantified by shorter distances in the VAE's learned latent space, linking variability in SN activity patterns to variability in song (Extended Data Fig. 5b, c). We then used a VAE to jointly model both sound spectrograms and ensemble activity (ROI average fluorescence from about 60 ROIs per bird). Here, the addition of a global latent variable (Fig. 3a; Extended Data Fig. 5d) enabled us to generate paired observations of ensemble activity and vocalization and encouraged the two learned representations to conform to one another during training (after correcting for time of day; Extended Data Fig. 6a). This joint encoding model achieved positive predictive ability relative to control models (Fig. 3b, Extended Data Fig. 5d–g), indicating shared information between acoustic and neural data. Moreover, when paired, held-out neural and vocal data were independently encoded and projected along axes of shared variability in the VAE latent space, the resulting representations were more strongly

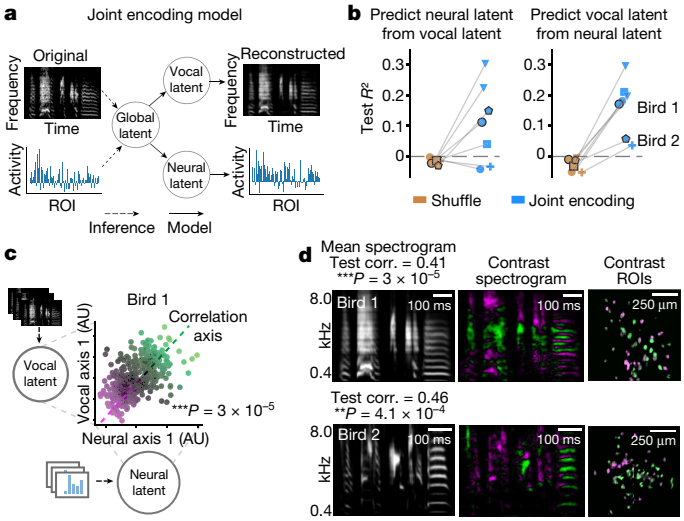


Fig. 3 | A joint neural–behavioural modelling approach relates SN population activity and song. a, Schematic of the joint modelling approach. Acoustic and neural data are modelled using separate encoders and decoders with a global latent variable capturing their shared variation. **b**, The joint encoding model captures shared variation in acoustic and neural data. The joint encoding model (blue) yields better predictions than a shuffle control (brown) on held-out data when predicting vocal latents from neural latents (left, 5 out of 7 sessions) and vice versa (right, 7 out of 7 sessions). Each marker represents a single recording session, with different shapes denoting different ($n = 5$) birds. The two birds in **d** are shown. **c**, Canonical directions linking vocal and neural latent spaces for a single session are strongly correlated on held-out test data (correlation = 0.40, $P = 3.0 \times 10^{-5}$, one-sided permutation test). Both test and train scatters are shown. Colour (magenta to green) indicates progression along the axis of shared variability. AU, arbitrary units. **d**, Contrast spectrograms (middle column) and ROI activities (right column) summarize shared neural and vocal variability for two birds. Each contrast is a weighted average of model-generated spectrograms and ROI activity pairs, with weights given by the pair's projection along the correlation (corr.) axis. *P*-values refer to corresponding correlations of held-out test data, as in **c**.

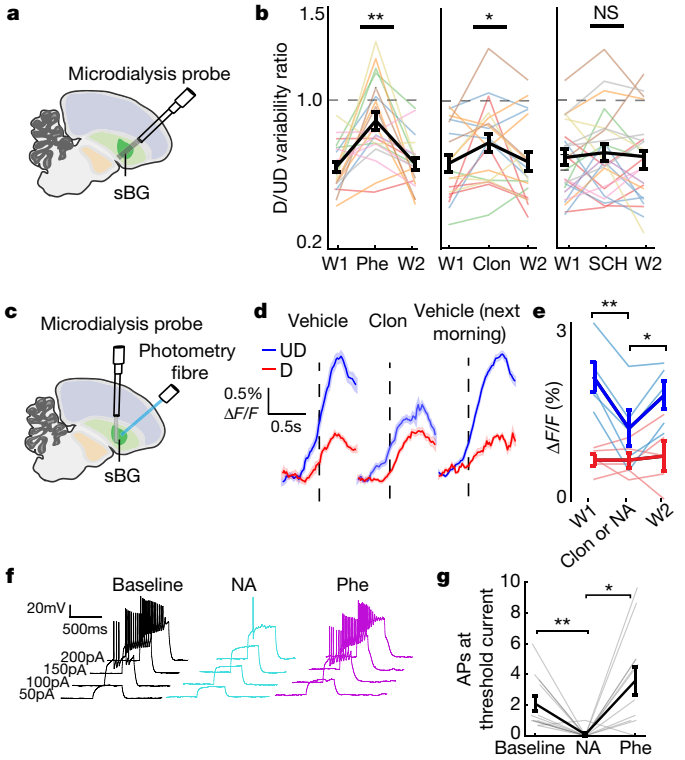


Fig. 4 | Noradrenergic signalling in the sBG reduces vocal variability by directly suppressing SN activity. a, Experimental approach for reverse microdialysis infusions into the sBG. **b**, Effect of noradrenergic and dopaminergic manipulations on directed (D) versus undirected (UD) pitch variability ratio (washes: W1 and W2). Phe, phentolamine (α-adrenergic receptor antagonist); clon, clonidine (α2-adrenergic receptor agonist); SCH, SCH23390 (D1 antagonist). Mixed-effects model, two-sided permutation test; see Supplementary Table 2 for detailed model output. Effect of drug presence on variability ratio: phentolamine, 0.25 + 0.060, $***P = 0.000094$ ($n = 20$ syllables from 8 birds); clonidine, 0.11 + 0.044, $*P = 0.01$ ($n = 19$ syllables from 6 birds); SCH23390, 0.05 + 0.043, $P = 0.19$ ($n = 23$ syllables from 9 birds). Colours indicate different birds. **c**, Experimental approach for simultaneous SN photometry and microdialysis experiments. **d**, Example SN photometry recordings from one bird during infusion of clonidine into the sBG; dashed line indicates first syllable of motif during infusion of clonidine into the sBG. **e**, Group data showing effects of α-adrenergic agonism (noradrenaline (NA) or clonidine) on SN calcium signals during undirected and directed song. One-way repeated measures analysis of variance (ANOVA) with Greenhouse–Geisser correction and post hoc Tukey test; $F_{(1.799, 7.197)} = 23.02$; wash 1 versus noradrenaline or clonidine $**P = 0.009$, noradrenaline or clonidine versus wash 2 $*P = 0.03$; clonidine, $n = 3$ birds; noradrenaline, $n = 2$ birds. **f**, Example ex vivo whole-cell current-clamp responses to positive current pulses at baseline, with noradrenaline application and with phentolamine application. **g**, Group data showing the number of action potentials (APs) elicited at threshold current measured at baseline, or at the same current in the presence of either noradrenaline or phentolamine (one-way repeated measures ANOVA with Greenhouse–Geisser correction and post hoc Tukey test; $F_{(1.458, 16.04)} = 10.01$; *P*-values are indicated; clonidine, $n = 12$ SNs). Data are mean ± s.e.m.

correlated than those produced by random data pairings for all cases (Fig. 3c, d). We also detected significant relationships between ROI activity, tempo and motif position within a song bout, as well as individual SNs that encoded the subsequent number of motifs in a song bout, suggesting a relationship between SN activity and song across multiple time scales (Extended Data Fig 6b–d). However, positive predictive performance could still be achieved when restricting analysis to only the first motif within a bout (Extended Data Fig. 5h). Therefore, the joint model's performance was not simply owing to correlations between relatively slow calcium signals and slow acoustic changes in motif structure across a song bout.

Next, we explored the relationships between neural activity and song discovered by the joint model. We depicted the effects of movement along these shared axes (Fig. 3c, magenta to green) by using simulated neural–vocal data pairs to produce contrast spectrograms and neural activity maps (Fig. 3d, Extended Data Fig. 6e). Despite the variability in ensemble participation and composition depicted in Fig. 1, the model learned a set of specific relationships between activity levels in select ROIs and specific song features, demonstrating that SN ensembles map distinct variants of neural ensemble activity to identifiable, highly structured changes in vocal output.

A remaining challenge is to identify the signals that regulate SN activity and thus influence vocal variability. Like the mammalian BG[16,17], the sBG receives input from noradrenaline-releasing neurons in the locus coeruleus (LC)[18] and dopamine-releasing neurons the in the ventral tegmental area (VTA)[19] (Extended Data Fig. 7a). Indeed, both noradrenaline and dopamine have been proposed to act in the sBG to reduce vocal variability during female-directed performance[20,21]. To explore this issue further, we infused either noradrenaline or dopamine antagonists into the sBG of adult male finches during practice and female-directed performance (Fig 4a, b). Blocking α-adrenergic receptors in the sBG abolished the differences in vocal variability that distinguish practice from performance by selectively increasing the variability of female-directed songs without affecting other song features (Fig. 4b, Extended Data Fig. 7b–d). Conversely, activating α-adrenergic receptors in the sBG selectively decreased the variability of practice songs

(Fig. 4b, Extended Data Fig. 7d). Lastly, blocking D1 receptors in the sBG did not alter song variability during practice or performance (Fig. 4b, Extended Data Fig. 7d), a different outcome from a prior study[21] that may reflect the briefer drug treatment schedules used here.

We then combined fibre photometry and reverse microdialysis to explore whether noradrenaline signalling underlies changes in SN activity that distinguish the practice and performance states (Fig. 4c, Extended Data Fig. 7e). First, we established that dialysing muscimol into the sBG strongly suppressed calcium signals in SNs during both states (Extended Data Fig.7f, g). We then determined that infusing noradrenaline or an α-adrenergic receptor agonist into the sBG strongly reduced SN activity during practice but not during female-directed performance (Fig. 4d, e). By contrast, blocking D1 receptors in the sBG did not alter SN activity levels during either practice or performance states (Extended Data Fig. 7h, i). In situ hybridization revealed that SN cell bodies expressed mRNA for the adrenergic receptor α2c (Extended Data Fig. 7j), which are G_i-coupled receptors also expressed in mammalian SNs[22]. Whole-cell current clamp recordings from identified SNs (Extended Data Fig. 8a) in brain slices showed that noradrenaline suppressed DC-evoked action potential activity in SNs, an effect that was reversed by blocking α-adrenergic receptors (Fig. 4f, g, Extended Data Fig. 8b–d). Furthermore, blocking these receptors in naive slices markedly increased DC-evoked action potential responses and input resistances in SNs (Extended Data Fig. 8e–h). Thus, noradrenaline signalling acts through α-adrenergic receptors to suppress SN activity and reduce vocal variability.

These findings suggest that LC neurons, the sBG's major source of noradrenaline, are more active when the male sings to a female than when alone. We tested this idea by first confirming that we could measure context-dependent differences in the expression of the immediate early gene *Fos* in the sBG (Extended Data Fig. 7k, l), as previously reported[3]. Extending this analysis to the LC revealed that *Fos* expression was higher during female-during singing (Extended Data Fig. 7m–o), consistent with an earlier study[23]. Further, *Fos* expression in the LC changed with social context but, unlike in HVC or the sBG, it did not depend on singing rates (Extended Data Fig. 7n), suggesting that it is not simply a consequence of female-directed singing.

Here we have linked social context-dependent changes in SN activity to state-dependent changes in song variability. Calcium signals in SNs but not in HVC PNs are increased during song practice and strongly suppressed during female-directed performance, reminiscent of the subtler changes that occur in SN but not HVC PN action potential activity across these states[8]. Apparently, small changes in action potentials in SNs correspond with large changes in calcium influx, which in turn may cause pronounced context-dependent changes in gene expression[3]. This study also establishes that noradrenaline suppresses SN activity to enable rapid switches between vocal practice and performance, consistent with previous immediate early gene studies[20] and in contrast to models involving dopamine signalling[8,21,24] (although non-D1 signalling cannot be excluded). The factors that contribute to enhanced performance to an audience are not fully understood, but probably include arousal, motivation and reward. More specifically, noradrenaline signalling is elevated in stressful and arousing situations[25,26], such as performing to an audience. Furthermore, male finches are more strongly motivated when singing to a female[27], and motivation can increase movement vigor and speed without sacrificing accuracy[28]. Finally, courtship is potentially rewarding, and rewarding contexts can suppress motor variability[29]. Regardless of how these factors contribute to enhanced song performance, our findings show that SNs drive song variability, rather than only passively receiving variability signals generated elsewhere in the cortico-BG pathway (that is, the lateral magnocellular nucleus of the anterior neostriatum (LMAN); Fig 1a). Moreover, while regions outside the sBG also contribute to song variability[2,30–32] and afford sites where noradrenaline can influence song[33], establishing that SNs generate and regulate song variability is important, given that the BG is a site where motor representations[34–36] and reinforcement signals converge to enable motor learning[34,37–39]. Indeed, the abundance of SNs is speculated to enable the exploration of a high-dimensional vocal–acoustic space on a millisecond timescale[34]. Joint modelling used here shows that natural variation in SN ensemble activity patterns can be linked to differences in song, revealing a complex but identifiable mapping between SN activity patterns and song variations. Thus, the dynamics of SN ensembles have meaningful effects on vocal exploration during practice, and the noradrenaline-dependent suppression of these dynamics enables stereotyped and precise song performance during courtship.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04004-1.

1. Sossinka, R. & Böhner, J. Song types in the zebra finch *Poephila guttata castanotis* 1. *Zeitschrift für Tierpsychologie* **53**, 123–132 (1980).
2. Kao, M. H., Doupe, A. J. & Brainard, M. S. Contributions of an avian basal ganglia–forebrain circuit to real-time modulation of song. *Nature* **433**, 638–643 (2005).
3. Jarvis, E. D., Scharff, C., Grossman, M. R., Ramos, J. A. & Nottebohm, F. For whom the bird sings: context-dependent gene expression. *Neuron* **21**, 775–788 (1998).
4. Goffinet, J., Brudner, S., Mooney, R. & Pearson, J. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *eLife* **10**, e67855 (2021).
5. Sainburg, T., Thielk, M. & Gentner, T. Q. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput. Biol.* **16**, e1008228 (2020).
6. Woolley, S. C. & Doupe, A. J. Social context-induced song variation affects female behavior and gene expression. *PLoS Biol.* **6**, e62 (2008).
7. Kao, M. H., Wright, B. D. & Doupe, A. J. Neurons in a forebrain nucleus required for vocal plasticity rapidly switch between precise firing and variable bursting depending on social context. *J. Neurosci.* **28**, 13232–13247 (2008).
8. Woolley, S. C., Rajan, R., Joshua, M. & Doupe, A. J. Emergence of context-dependent variability across a basal ganglia network. *Neuron* **82**, 208–223 (2014).
9. Kojima, S., Kao, M. H., Doupe, A. J. & Brainard, M. S. The avian basal ganglia are a source of rapid behavioral variation that enables vocal motor exploration. *J. Neurosci.* **38**, 9635–9647 (2018).
10. Hein, A. M., Sridharan, A., Nordeen, K. W. & Nordeen, E. J. Characterization of CaMKII-expressing neurons within a striatal region implicated in avian vocal learning. *Brain Res.* **1155**, 125–133 (2007).
11. Kozhevnikov, A. A. & Fee, M. S. Singing-related activity of identified HVC neurons in the zebra finch. *J. Neurophysiol.* **97**, 4271–4283 (2007).
12. Hahnloser, R. H. R., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002).
13. Liberti, W. A. 3rd et al. Unstable neurons underlie a stable learned behavior. *Nat. Neurosci.* **19**, 1665–1671 (2016).
14. Kingma D. P. & Welling M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2013).
15. Rezende D. J., Mohamed S. & Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. Preprint at http://arxiv.org/abs/1401.4082 (2014).
16. Björklund, A. & Dunnett, S. B. Dopamine neuron systems in the brain: an update. *Trends Neurosci.* **30**, 194–202 (2007).
17. Zerbi, V. et al. Rapid reconfiguration of the functional connectome after chemogenetic locus coeruleus activation. *Neuron* **103**, 702–718.e5 (2019).
18. Castelino, C. B., Diekamp, B. & Ball, G. F. Noradrenergic projections to the song control nucleus area X of the medial striatum in male zebra finches (*Taeniopygia guttata*). *J. Comp. Neurol.* **502**, 544–562 (2007).
19. Person, A. L., Gale, S. D., Farries, M. A. & Perkel, D. J. Organization of the songbird basal ganglia, including area X. *J. Comp. Neurol.* **508**, 840–866 (2008).
20. Castelino, C. B. & Ball, G. F. A role for norepinephrine in the regulation of context-dependent ZENK expression in male zebra finches (*Taeniopygia guttata*). *Eur. J. Neurosci.* **21**, 1962–1972 (2005).
21. Leblois, A., Wendel, B. J. & Perkel, D. J. Striatal dopamine modulates basal ganglia output and regulates social context-dependent behavioral variability through D1 receptors. *J. Neurosci.* **30**, 5730–5743 (2010).
22. Hara, M. et al. Role of adrenoceptors in the regulation of dopamine/DARPP-32 signaling in neostriatal neurons. *J. Neurochem.* **113**, 1046–1059 (2010).
23. Bharati, I. S. & Goodson, J. L. Fos responses of dopamine neurons to sociosexual stimuli in male zebra finches. *Neuroscience* **143**, 661–670 (2006).
24. Budzillo, A., Duffy, A., Miller, K. E., Fairhall, A. L. & Perkel, D. J. Dopaminergic modulation of basal ganglia output through coupled excitation-inhibition. *Proc. Natl Acad. Sci. USA* **114**, 5713–5718 (2017).
25. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).

26. Breton-Provencher, V. & Sur, M. Active control of arousal by a locus coeruleus GABAergic circuit. *Nat. Neurosci.* **22**, 218–228 (2019).
27. Cooper, B. G. & Goller, F. Physiological insights into the social-context-dependent changes in the rhythm of the song motor program. *J. Neurophysiol.* **95**, 3798–3809 (2006).
28. Wong, A. L., Lindquist, M. A., Haith, A. M. & Krakauer, J. W. Explicit knowledge enhances motor vigor and performance: motivation versus practice in sequence tasks. *J. Neurophysiol.* **114**, 219–232 (2015).
29. Pekny, S. E., Izawa, J. & Shadmehr, R. Reward-dependent modulation of movement variability. *J. Neurosci.* **35**, 4015–4024 (2015).
30. Jaffe, P. I. & Brainard, M. S. Acetylcholine acts on songbird premotor circuitry to invigorate vocal output. *eLife* **9**, e53288 (2020).
31. Olveczky, B. P., Andalman, A. S. & Fee, M. S. Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS Biol.* **3**, e153 (2005).
32. Sober, S. J., Wohlgemuth, M. J. & Brainard, M. S. Central contributions to acoustic variation in birdsong. *J. Neurosci.* **28**, 10370–10379 (2008).
33. Sheldon, Z. P. et al. Regulation of vocal precision by noradrenergic modulation of a motor nucleus. *J. Neurophysiol.* **124**, 458–470 (2020).
34. Fee, M. S. & Goldberg, J. H. A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. *Neuroscience* **198**, 152–170 (2011).
35. Markowitz, J. E. et al. The striatum organizes 3D behavior via moment-to-moment action selection. *Cell* **174**, 44–58.e17 (2018).
36. Klaus, A. et al. The spatiotemporal organization of the striatum encodes action space. *Neuron* **95**, 1171–1180.e7 (2017).
37. Hisey, E., Kearney, M. G. & Mooney, R. A common neural circuit mechanism for internally guided and externally reinforced forms of motor learning. *Nat. Neurosci.* **21**, 589–597 (2018).
38. Xiao, L. et al. A basal ganglia circuit sufficient to guide birdsong learning. *Neuron* **98**, 208–221.e5 (2018).
39. Coddington, L. T. & Dudman, J. T. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nat. Neurosci.* **21**, 1563–1573 (2018).

Methods

Animals

All experiments were performed in accordance with a protocol approved by the Duke University Institutional Animal Care and Use Committee. Results were collected from a total of 64 adult (>80 days post hatch (dph)) male zebra finches (*Taeniopygia guttata*).

Viral injection surgeries

Adult zebra finches (80–163 dph) were anaesthetized with 2% isoflurane gas before being placed in a custom stereotaxic apparatus. After applying a topical anaesthetic (0.25% bupivacaine) and making a vertical incision in the skin over the skull, we made ~1-mm craniotomies in the skull at a predetermined distance from the bifurcation of a major blood vessel (the 'Y sinus'; sBG: 43 degrees from horizontal head angle, 5.1 mm anterior, 1.6 mm lateral, sitch to 72 degrees, move 1.6 mm anterior, 3.0 mm ventral from brain surface. This angle was used to avoid damage to the song nucleus LMAN. Using a glass pipette attached to a pressure injection system (Drummond Nanoject II), we gave bilateral injections of dextrans or virus (constructs specified below for each experiment). After injections, craniotomies were sealed with bone wax and the incision site in the skin was closed with a tissue adhesive (VetBond).

Miniscope imaging

Surgery. Using the surgical procedures described above, we induced the expression of genetic calcium indicators of neural activity in either HVC or sBG. To target our injections to the sBG, we mapped the anterior boundary of the nucleus as defined by the presence of high tonic activity. We then injected virus 400 µm posterior from the anterior boundary of the nucleus with 3 injection sites (2,950, 2,750 and 2,550 µm ventral, 15 injections of 9.2 nl each, 10 min wait per site). Two viral strategies were used to infect sBG (AAV2/9.FLEX.CAG.GCaMP7f + AAV2/9 0.4 CamKII. Cre, 6 birds; AAV2/9.CamKII.GCaMP7f, 2 birds). In the same surgery, a 1-mm diameter prism gradient-index lens (Inscopix) was then slowly implanted (~5 µm s^{-1}) at the identified anterior boundary of sBG, with the purpose of targeting our imaging plane 250 µm into the sBG. For HVC injections, birds received three 250 nl injections of lentivirus, containing a Rous sarcoma virus (RSV) promoter driven genetically encoded calcium indicator, GCaMP6f, into the song premotor cortex. This virus has been shown to have a strong tropism for excitatory neurons with a preference for projection neurons[13].

After waiting ~4 weeks for viral expression, a magnetic (Inscopix) or 3D printed baseplate was implanted on the bird's skull to hold a miniature microscope[13,40] for imaging. After a period of recovery (~3 d), birds were placed in a recording chamber and the miniature microscope was attached to the baseplate for imaging. The activity of sBG or HVC neurons was then imaged, and data were collected using either custom or commercial acquisition software (Inscopix) and synchronized with custom written software (synchronizing audio and frame times). The LED power was maintained between 0.12 and 0.24 mW mm^{-2}. The maximum field of view was approximately either 900 µm × 650 µm or 700 × 525 µm, Sound was sampled at 44 kHz and imaging data were sampled between 10 Hz and 30 Hz, corresponding to exposure times of 99.84 ms to 33.25 ms. To capture as many singing trials as possible, 30–60-s trials were automatically triggered via custom scripts to detect vocalizations.

ROI extraction and event detection. To extract regions of interest corresponding to putative neurons, we used an adaptation of the constrained non-negative matrix factorization approach (CNMF) for miniscope data[41] (CNMF-E). This framework can reliably deal with the large fluctuating background in endoscopic data and the highly correlated patterns present in song-related areas. The following parameters were used for all our data, with slight variations across birds due to variability in signal quality: mean-subtracted 2D gaussian smoothing kernel (gSig) = 3–5 pixels, maximum soma diameter (gSiz) = 10–14 pixels, minimum pixel-to-noise ratio for seeding a neuron (PNR) = 6–8, minimum spike size (smin) = 5 and minimum local correlation: 0.85. These parameters were chosen to minimize the inclusion of neurons with weak SNRs. A ring model was used to estimate background signal for each neuron (ring_radius = 18 pixels), and a maximum spatial overlap of 0.75 was used for merging overlapping neurons. The noise level for each neuron was defined by subtracting the reconvolved calcium activity from the raw calcium activity (neuron.C_raw – neuron.C). This noise level was used to *z*-score all neurons' activity. To detect calcium events, we used a binarized version of the inferred spiking events (neuron.S, all inferred events set to 1).

Female-directed song collection. One or two females were presented in a transparent plexiglass cage (~1 min per presentation). Females were presented regularly every ~45 min to collect female-directed song, and undirected song was collected in between presentations. Directed song was used for further analyses if it occurred with low latency (<20 s).

Speaker playback experiments. Songs were recorded from each bird in isolation. These recordings were amplified and low-pass filtered at 10 kHz, and further bandpass filtered between 350 and 10,000 Hz for playback. A 10-ms ramp function was applied at the beginning and end of each stimulus using a cosine function (custom MATLAB, custom LabVIEW) to suppress acoustical transients. Stimuli were presented from a speaker located ~20 cm in front of the bird, and calibrated to a sound level similar to normal song performance (~20–30 song playbacks per experiment).

Real-time auditory feedback experiments. We generated an acoustic template for online syllable detection that identified no less than 80% of the renditions of the targeted syllable with no more than a 10-ms jitter in detection onset. We then used this template to trigger a 25 ms burst of WN on 50% of identified syllables. 'Catch' and 'hit' trials were then pooled for the analyses shown.

Webcam video analyses. The bird's position was measured using custom MATLAB codes (M. Ben-Tov, Technion University) that detected and tracked the centroid of the body position across video frames (Logitech webcam, 30 frames per second), and speed of movement was calculated as the change in position across pairs of frames. A threshold was then defined as the 80th percentile of the distribution of instantaneous velocities displayed by the bird over a 5-min period, and any crossings of >2 consecutive frames were designated as locomotion periods.

Song-triggered optogenetic inhibition

Adult male birds (age range 83–162 dph) were bilaterally injected sBG with a virus containing an archaerhodopsin construct (2/9. AAV-CAG-ArchT-GFP or 2/1.AAV-CamKII-ArchT-GFP) or a control construct (2/1 AAV-CamKII-GFP or 2/9 AAV-CAG-Cre-GFP). After waiting ~21 days to allow for viral expression, 6 of the 12 birds (3 CAG, 3 CamKII) were tested for terminal field optogenetic responses in sBG with an optrode through which 50- to 500-ms pulses of light were delivered and neural activity was recorded simultaneously (Differential A-C Amplifier 1700, A-M Systems). All birds were then implanted bilaterally over sBG with either tapered fibreoptic ferrules[42] (9 birds, Optogenix), or regular ferrules (3 birds, RWD). Craniotomies were then sealed with melted bone wax; ferrules were secured in place with MetaBond and then covered with a layer of VetBond. After birds recovered from anaesthesia under a heat lamp, fibreoptic cables (Thorlabs, 200-µm core, 0.37 NA) were connected to the newly implanted ferrules by ferrule sleeves. The other ends of the fibreoptic cables were attached to a two-channel optical commutator (FRJ_1×2i_FC-2FC, Doric), allowing the bird to move about its cage freely. The commutator was then connected by a patch cable (Thorlabs) to a DPSS laser (BL473T3-100, Shanghai Lasers). After waiting 3–5 days for habituation and singing

to start, a syllable template was designed that detected no less than 90% of the renditions of the targeted syllable with no more than a 7-ms jitter in detection onset. This template was used to trigger delivery of a continuous pulse of green light (400–1,000 ms duration, 532 nm, 8–15 mW emitted at each ferrule) to the sBG on a random 30% of trials. On the same day, or in the following 1–3 days, a female zebra finch was presented intermittently (every 20–30 min, ~30 s per presentation) to collect directed song. Undirected song was collected for the same days. Upon ending the experiment, histology was performed, and only birds that had robust viral expression and accurate placement of ferrules (taper approximately spanning the length of sBG, or blunt end near the dorsal edge of sBG). Exclusion of birds was blind to behavioural results.

Joint encoding VAE

Motif preprocessing. Song motifs were manually segmented. Spectrograms were taken as the log modulus of a short time Fourier transform (Hann windows, sample rate: 44.1 kHz, segment length: 512, overlap: 384), manually scaled and clipped to an appropriate range. The spectrograms were then interpolated to 128 target frequencies linearly spaced between 0.4 and 8 kHz. Linear time warping was then performed on the spectrogram power time series (spectrograms summed over the frequency dimension, median absolute deviation-normalized) to account for tempo variation. More specifically, for each power time series $f(t)$, a shift β_0 and log inverse tempo value $\log \beta_1$ were optimized using Powell's method with the following objective:

$$\min_{\beta_0,\beta_1} \left\| f\left(\beta_0 + \beta_1 t\right) - f_{\text{target}}(t) \right\|_2^2 + \lambda_0 \beta_0^2 + \lambda_1 \left(\log \beta_1\right)^2$$

with initial conditions $\beta_0 = \log \beta_1 = 0$, where $f(\beta_0 + \beta_1 t)$ is defined by linear interpolation and f_{target} is taken to be the average warped time series. This procedure was repeated, with f_{target} being updated every iteration and λ_0 and λ_1 slowly annealed to zero. Each spectrogram was then linearly interpolated to 128 warped target times using its associated warping parameters β_0 and β_1, resulting in 128-by-128 time-warped spectrograms. Spectrograms in Fig. 3, Extended Data Figs. 5, 6 are upsampled by a factor of 3 using cubic spline interpolation for visualization.

Calcium preprocessing. After ROI extraction (described above), the raw calcium trace for each ROI was independently z-scored. For each song motif, a vector of ROI activities was calculated by averaging the raw calcium trace of each ROI from 300 ms before motif onset to 150 ms after motif offset.

Minimizing time confounders. To prevent our analysis from finding shared information between neural activity and vocal behaviour that is simply due to non-causal time-in-day effects such as photobleaching and diurnal changes in song, we iteratively removed the components of our calcium activity vectors and song spectrograms that could be reliably predicted by time. Specifically, we first used lasso to perform linear feature selection (predicting calcium activity vectors and spectrogram vectors from time in recording). More specifically, if $x_i(t)$ are the data, we performed

$$\min_{\beta} \sum_t (x_i(t) - \beta_i t)^2 + \lambda \sum_i |\beta_i|$$

with λ chosen manually for spectrograms and calcium activity. This allowed us to consider only elements of the data (spectrogram pixels or ROIs) that showed a clear linear dependence on time. After this, we performed a single kernel ridge regression[43] that attempted to fit $\sum_i \beta_i x_i(t)$, the projection of the data along the axis learned in the first step, as a (potentially nonlinear) function of time (kernel bandwidth parameters chosen to maximize average test-set predictive performance over seven folds of the data). Third, we removed these effects from the data, with the residuals used for subsequent analysis. We then repeated this procedure until the lasso regression produced the 0-vector feature, indicating time was no longer able to explain training set variance.

Separate encoding models. For the separate VAEs encoding spectrograms and calcium activity, we used a standard VAE model[14,15]. If x denotes the data and y the learned latent variables with unit normal priors, we assumed:

$$y_i \sim N(0, I), x_i \sim N(\mu_i(y_i), \varsigma_i I)$$

where $i = 0,1$ indicates neural or vocal modality, and μ is a nonlinear mapping parameterized by a neural network. We performed approximate inference in this network by assuming a posterior also defined by neural networks $f_{\mu,i}$ and $f_{\sigma^2,i}$:

$$q_i(y_i \mid x_i) \propto N(y_i; 0, I) N(y_i; f_{\mu,i}(x_i), f_{\sigma^2,i}(x_i) I)$$

For the case of Gaussians, the normalization constant for this distribution can be calculated analytically. The parameters of μ_i, $f_{\mu,i}$, and $f_{\sigma^2,i}$ for $i = 0,1$ are optimized via stochastic gradient ascent to maximize the standard evidence lower bound (ELBO):

$$L_{\text{separate}} = \sum_{i=0,1} \sum_n E_{q(y_i \mid x_i^n)} \log \left[\frac{p(x_i^n, y_i)}{q(y_i \mid x_i^n)} \right]$$

where p is defined by the generative process and $n = 1,\ldots,N$ indexes motif number. This is equivalent to training two independent VAEs.

Joint encoding model. For the joint encoding model, we assume an additional global latent variable z that gives rise to the latent variables for each data type via linear map:

$$z \sim N(0, I), y_i \sim N(A_i z, \sigma^2 I), x_i \sim N(\mu_i(y_i), \varsigma_i I)$$

Again, we perform approximate inference, this time attempting to directly estimate the posterior over z given the data:

$$q_i(z \mid x_i) \propto N(z; 0, I) N(z; f_{\mu,i}(x_i), f_{\sigma^2,i}(x_i) I).$$

As in previous work by Wu and Goodman[44], we additionally define

$$q(z \mid x_1, x_2) \propto N(z; 0, I) \prod_{i=1,2} N\left(z; f_{\mu,i}(x_i), f_{\sigma^2,i}(x_i) I\right),$$

which combines evidence in our beliefs about z across the two modalities using a product of experts construction. The parameters A_i, σ^2, μ_i, $f_{\mu,i}$ and $f_{\sigma^2,i}$ for $i = 0,1$ are optimized to maximize:

$$L_{\text{joint}} \equiv L_{\text{separate}} + \sum_n E_{q(z \mid x_0^n, x_1^n)} \log \left[\frac{p(x_0^n, x_1^n, z)}{q(z \mid x_0^n, x_1^n)} \right]$$

Note that this is the sum of three separate ELBOs, one joint neural/vocal ELBO and two separate ELBOs, which reflects the multi-objective nature of joint encoding models.

Network and training details. The latent dimension was fixed at 32, as in a previous study[4], and the same deep convolutional neural network architectures were used to parameterize the model and recognition for the vocal arm of the joint and separate encoding models. The neural arm model and recognition model were parameterized by two- and three-layer fully connected networks with ReLU activation, respectively. Approximate posteriors were restricted to the family of Gaussians with diagonal covariance. All parameters were optimized using Adam optimization with a learning rate of 0.001.

Neural activity correlation versus VAE latent distance analysis

For each experimental session, a standard VAE was trained to model motif spectrograms, corrected for time-of-day trends (see 'Minimizing time confounders'). Then, the collection of pairs of motifs was segregated by the similarity of the corresponding ROI activity vectors (see 'Calcium preprocessing'). Specifically, each pair of motifs was sorted by the correlation of the corresponding ROI activity vectors. In Extended Data Fig. 5b, the cumulative distribution of VAE latent distances between pairs of motifs is plotted for several quantiles of ROI activity correlations for a single example session. In Extended Data Fig. 5c, the median VAE latent distance between pairs of motifs in each ROI correlation decile is calculated for each bird. For each experimental session, these median distances subtracted by their mean value are plotted in order to adjust for scale differences across birds.

Spectrogram and ROI contrasts. For each experimental session, a single joint encoding model was trained with a random 20/80 test/train split for 300 epochs. The neural and vocal latent means were then inferred for each training datapoint using the modality-specific recognition models $q(z \mid x_i)$ for $i = 0,1$. Latent dimensions explaining negligible variance were discarded by transforming the latent means into their principal components and truncating at the number of latent dimensions needed to explain 99% of variance.

To verify that the trained model captured shared structure between the two latent spaces, we performed canonical correlation analysis to identify highly correlated axes in each latent space. We then used the individual encoders to infer z from each data type separately, performed the same dimension reduction via principal component analysis, and finally projected onto each identified canonical coordinate to produce scatterplots as in Fig. 3c. We report the correlation of held-out test data projected onto the pairs of canonical coordinates and estimate significance by randomly shuffling the spectrogram–calcium activity pairing.

To visualize the effects of movement along these canonical coordinates on ROI fluorescence and spectrograms, we created pairs of representative spectrograms as follows: First, we generated 160,000 spectrogram–neural activity vector pairs using the trained model. These pairs were then weighted by their projected distance from the midpoint along the identified canonical coordinate axis, such that spectrograms farther from the midpoint were more likely to contribute. Finally, we calculated and plotted weighted averages of the model-generated data using these weights to produce a summary of contrast along each axis.

Model performance comparison

For each experimental session, we split the data into seven random tranches for cross validation. For each training run, five of these tranches were used for training, one was used for hyperparameter selection for the VAE model, and the last was used for both test performance of the VAE model and for cross-validation of models that predicted vocal latents from neural latents and vice versa (Extended Data Fig. 5f). For each combination of model noises $\Sigma_{\text{neural}} \in \{0.2,0.4\}$, $\Sigma_{\text{vocal}} \in \{0.02,0.04\}$, and objective (predicting neural latents from vocal latents and vice versa), we trained a separate model, pausing every 10 epochs to evaluate our objective on the validation set. We trained for a maximum of 1,000 epochs and terminated a run if there was no improvement on the validation set objective in 200 epochs. At the end of a set of runs, we used the first (VAE) validation set to identify the combination of Σ_{neural}, Σ_{vocal}, and training epoch that produced the best results. We report training performance as the corresponding results on the final test set. We report the average cross-validation test performance over the seven test tranches.

To assess the model's ability to predict neural latents from vocal latents and vice versa, we used ridge regression (and also report kernel ridge regression in the supplement). We set the regularization parameter for this model using a leave-one-out cross-validation procedure on the second validation set: for each data point in the set, we selected the parameter value that produced the best fit on all but the selected data point and assessed performance on the single held-out data point. For a given multivariate latent Y, we calculated the average of the R^2 objective: $R^2 = 1 - \frac{\left\| Y_{\text{test}}^{\text{true}} - Y_{\text{test}}^{\text{predicted}} \right\|_2^2}{\left\| Y_{\text{test}}^{\text{true}} - Y_{\text{train}}^{\text{avg}} \right\|_2^2}$ across these data. We report the average cross-validation value of this metric over the 7 test tranches.

In vivo microdialysis

Custom-made reverse microdialysis probes were implanted into the sBG of adult birds. The probes were secured in place first using Meta-Bond and then a coating of VetBond. Birds were then removed from the apparatus and recovered under a heat lamp. After recovery, birds were placed in a sound isolation box and left for two to three days until their singing rate recovered to normal levels.

For data collection, the following drug schedule was sequentially repeated for selected drugs: two days of saline, followed by one day of drug and then two days of saline. All infusions were done in the morning before cage lights came on. Analysis was restricted to a window up to 5 h after infusion. Drugs were washed out every day at 17:00, multiple drugs were used sequentially on the same bird until probes broke or failed. The following concentrations and drugs were used for behavioural and photometry experiments: SCH23390 (~5 mM), phentolamine (~5 mM), clonidine (1.5 mM) and muscimol (~1 mM).

Fibre photometry recordings

A three-channel multi-fibre photometry system (Neurophotometrics) was used for these experiments. In brief, light from three different wavelength LEDs (470 nm and 560 nm in phase, and 415 nm out of phase) were bandpass filtered and directed down a fibreoptic patch cord via a 20× objective. This was coupled to a fibreoptic cannula implanted in the animal. Emitted GCaMP fluorescence was collected through the same cannula and patch cord (Doric MFP_200/220/900-0.37 FC_MF1.25), split by a 532 long pass dichroic, bandpass filtered, and focused onto opposite sides of CMOS camera sensor. Data were acquired using the open-source software Bonsai by drawing a region of interest around the two images (green and red) of the patch cord and calculating the mean pixel value. Both the blue and green channels were then median filtered with a window of four frames, and the blue channel was fit to and subtracted from the green signal using the Matlab polyfit function. Baseline fluorescence was defined as the average signal during a 60 s period of silence.

Dual photometry recordings in HVC and sBG

Following an injection of AAV2/9.hSyn.AxGCaMP6m into HVC, we waited ~28 days for axonal GCaMP expression. A second surgery was then performed in which the first photometry fibre was implanted dorsal to HVC to image local HVC axons, and the second was implanted in the ipsilateral sBG to image HVC_{sBG} axons.

Simultaneous pharmacology and photometry recordings

Fibre photometry implants were placed in sBG, but a small 300 × 300 µm region of cranial surface was left exposed without metabond and covered temporarily with Kwik-sil. After waiting ~21 days for viral expression and observing song-related signals, a second surgery was performed in which a microdialysis infusion probe (described above) was lowered through the exposed region at a head angle of 43 degrees, 5.1 AP, 1.45 mm ML, in order to avoid damaging LMAN, or displacing the sBG photometry fibre. In 4 out of 6 birds, 2 mM muscimol was infused to validate stereotactic targeting and drug diffusion. For the experiments in Fig. 4d, e and Extended Data Fig. 7h, two birds were shared between the adrenergic and SCH23390 experiment. In an additional experimental session, one received noradrenaline and the other received clonidine.

In-situ hybridization

In situ hybridization was performed using hybridization chain reaction (HCR v3.0, Molecular Instruments). Dissected brain samples were post-fixed overnight in 4% PFA at 4 °C, cryoprotected in a 30% sucrose solution in RNAse-free PBS (DEPC-PBS) at 4 °C for 48 h, frozen in Tissue-Tek OCT Compound (Sakura), and stored at −80 °C until sectioning. Eighty-micrometre-thick coronal floating sections were collected into a sterile 24-well plate in DEPC- PBS, and fixed again briefly for 5 min in 4% PFA. Sections were rinsed in DEPC-PBS, incubated for 45 min in 5% SDS in DEPC-PBS, rinsed and incubated in 2× SSCT, pre-incubated in HCR hybridization buffer at 37 °C, and then placed in HCR hybridization buffer containing RNA probes overnight at 37 °C. The next day, sections were rinsed 4 × 15 min at 37 °C in HCR probe wash buffer, rinsed with 2× SSCT, pre-incubated with HCR amplification buffer, then incubated in HCR amplification buffer containing HCR amplifiers at room temperature for ~48 h. On the final day, sections were rinsed in 2× SSCT, rinsed again with 2× SSCT, then mounted on slides and coverslipped with Fluoromount-G (Southern Biotech). After drying, slides were imaged on a Zeiss inverted 710 laser scanning confocal microscope.

Fos measurements in LC, sBG and HVC

To minimize off-target *Fos* mRNA detection, birds were perfused 30 min after cage lights first turned on in the morning. For directed singing, 4–5 females were presented sequentially over 30 min to maximize motif amounts. Live video was monitored to ensure no significant undirected (facing away from the female, disengaged) singing occurred during female presentations. To enforce the silent condition, the sound booth door where the cage resides was left open, which has previously been noted to effectively suppress singing[3]. For the undirected condition, birds were allowed to sing freely in the morning for the 30-min window. Lights were then turned off and birds were immediately perfused.

For each bird, a ~25-µm-thick *z*-stack (~0.85-µm optical slices, 350 × 350 µm) encompassing locus coeruleus was collected at 40× power to accurately visualize *Fos* signal, along with VGAT and TH channels. All image processing was done with ImageJ. First, all channels were noise-subtracted (20 µm rolling-ball radius), and VGAT and TH channels were smoothed, automatically thresholded (Otsu method), and converted to a binarized mask. The mask was then transferred to the *Fos* channel, which was then also thresholded. *Fos* particles within each mask were then quantified for intensity and total area (expressed as a fraction of the VGAT or TH masks). For VGAT quantification, a bounding rectangle was manually drawn around the LC TH-positive signal, and only the VGAT signal within that area was used for masking. For HVC and sBG measurements, the mean intensity value of an entire 60-µm-thick *z*-stack (20× power, taken in the center of the sBG or HVC) was used.

Whole-cell recordings

Birds (65–120 dph) were deeply anaesthetized with isoflurane and were used to prepare 250 µm thick coronal slices. A subset of birds received injections 2–3 weeks prior of an AAV encoding a fluorescent reporter under the CamKII promoter in order to label MSNs. The brain was dissected in ice-cold artificial cerebrospinal fluid (ACSF) containing the following (in mM): 119 NaCl, 2.5 KCl, 1.30 $MgCl_2$, 2.5 $CaCl_2$, 26.2 $NaHCO_3$, 1.0 $NaHPO_4$-H_2O and 11.0 dextrose, and bubbled with 95% O_2, 5% CO_2. The brain was mounted on an agar block and sliced in ice-cold ACSF with a vibrating-blade microtome (Leica). Slices were incubated for 15 min at 32 °C in a bath of NMDG recovery solution containing the following (in mM): 93.0 NMDG, 2.5 KCl, 1.2 NaH_2PO_4, 30.0 $NaHCO_3$, 20.0 HEPES, 25.0 glucose, 2.0 thiourea, 5.0 sodium L-ascorbate, 2.0 sodium pyruvate, 10.0 $MgSO_4$ 7 H_2O, 0.5 $CaCl_2$ and 95.0 HCl. Slices were then moved to a bath of ACSF as above and allowed to gradually reach room temperature over the course of 30 min, where they remained for the duration. Recordings were performed in ACSF at a temperature of 32 °C. For current clamp experiments patch electrodes (7–10 MΩ) were filled with potassium gluconate internal solution containing the following (in mM): 124 potassium gluconate, 4 NaCl, 10 HEPES, 2 EGTA, 2 $MgCl_2$, 2 Mg-ATP salt 0.3 Na-GTP salt and 10 sodium phosphocreatine. Neurons were targeted using interference contrast and epifluorescence to visualize fluorescent indicators (GFP or tdTomato) previously expressed via viral injection. Recordings were made using a Multiclamp 700B amplifier whose output was digitized at 10 kHz (Digidata 1440A). Liquid junction potential was measured at (+5 mV) and was not compensated. The drugs used (20 µM noradrenaline, 10 µM phentolamine (Sigma-Aldrich)) were added to the ACSF and perfused onto slices. Pharmacological agents were bath applied for 10 min before making recordings. Signals were analysed using Igor Pro (Wavemetrics). Spike threshold was defined as the first 500-ms current step (in 25-pA increments) that elicited one or more spikes in the baseline condition. Input resistance was defined as the slope of the regression line fitted to the current–voltage curve at membrane potentials more negative than −50 mV, as described previously[45]. Sag ratio was defined as the difference between the peak deflection at the beginning of a hyperpolarizing current injection and the resting membrane potential divided by the difference between the steady state potential reached after a hyperpolarizing current injection and the resting membrane potential. A sag ratio of 1 therefore indicates no sag, while a value greater than 1 indicates the presence and degree of sag. The 10–90% rise time of the membrane response to a hyperpolarizing current injection was calculated as the time (in ms) needed for the membrane potential to go from 10% to 90% of the difference between the resting membrane potential and the potential after current injection.

Specificity and sensitivity

To compute a single neuron's specificity (true negative rate) and sensitivity (true positive rate), we first defined a window of −1 s to +1 s around each bout. A bout is defined as continuous periods of song with no inter-syllable pauses longer than 100 ms. Bouts were only selected if they were flanked by two or more seconds of silence. After extracting these singing windows, we found an equal amount of equivalent windows for periods of silence (2 s window of silence flanked by at least 4 s of silence) and computed the following for each identified neuron, where a song where a neuron 'participated' is defined by the presence of an inferred spiking event (neuron.S, see ROI extraction and event detection).

Sensitivity = no. of participated song windows/no. of total song windows

Specificity = no. of non-participated silence windows/no. of total silence windows

All-to-all correlation coefficient

To determine how stable each neuron's response time course is over different songs, we computed all pairwise Pearson correlations for each individual neuron's active trials. We then calculated the median of this distribution for each neuron, which were combined across birds and plotted as in Extended Data Fig. 1b–d.

Shared population between trials

To better understand how the identity of the song-related neural ensembles varied over song renditions, we calculated the list of active neurons for any given pair of songs as defined by the presence of an inferred event. We then determined what fraction of this total list was active during both songs, and repeated this procedure for every pair of trials. All pairwise comparisons were then combined across birds and plotted in Extended Data Fig. 1d.

Pitch variability

The percent variability in the pitch of a given syllable was calculated by measuring the pitch of the entire small stable component (>15 ms) of the target syllable. The standard deviation was then computed, divided by the average cross-rendition pitch, and multiplied by 100 to get a per cent variability measure. Only renditions performed 30 min after drug infusion and up to 5 h after were used for further analyses.

Calcium event-triggered average spectrogram

To generate the colour-coded event-triggered spectrogram averages (Extended Data Fig. 2c), we collected, for each neuron, all events (binarized inferred spikes) across the entire time series of concatenated calcium videos (time-series durations ranged between 10 and 30 min per bird), regardless of whether the event happened during or outside of song.

Statistics and quantification

Data are presented as the mean ± s.e.m. One-way repeated measures ANOVAs were corrected for sphericity (Greenhouse–Geisser) to ensure accurate *P* values. *P* values of 0.05 or below were considered significant. Star values as: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Mixed-effects models

For all data in which multiple syllables were collected from the same bird and pooled we used linear mixed effects models to account for the hierarchical dependencies in the data. Specifically, we accounted for correlations between measurements that were sampled from the same bird by including a random effect term for drug or laser condition grouped by bird ID, as well as a random effect term for syllable ID. Fixed effect coefficient ± standard error, confidence intervals and *P* values are reported in Supplementary Tables 1, 2.

Outlier and experiment inclusion criteria

We excluded acoustic measurements that exceeded 4 × s.d. of a given dataset, as well as measurements where pitch could not be computed reliably. These criteria served to remove erroneous measurements resulting from faulty segmentation and cage noise. Cut-off for minimum number of song renditions varied by experiment type owing to the different challenges in obtaining singing data for each experiment. Minimum cut-offs were as follows: for miniscope imaging of undirected-only days, 30 undirected bouts; for miniscope and photometry imaging of undirected and directed songs, 10 bouts; for optogenetic experiments, 100 catch bouts, and 30 laser-stimulated bouts; for microdialysis experiments, 50 undirected bouts, 15 directed bouts; and for simultaneous microdialysis and photometry experiments, 10 bout initiations per condition.

For all analyses in Fig. 1 and Extended Data Fig. 1, each song was selected to have the same core motif structure (for example, 'abcd'), and a window of −1 s to +1 relative to song onset was selected for further analyses. For each bird, data used for any given plot were collected within a single day.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Core datasets have been posted to the Duke University Library Research Data Repository (https://research.repository.duke.edu). Source data are provided with this paper.

Code availability

Custom code and software are available at https://github.com/pearsonlab/autoencoded-vocal-analysis and https://github.com/pearsonlab/finch-vae.

40. Ghosh, K. K. et al. Miniaturized integration of a fluorescence microscope. *Nat. Methods* **8**, 871–878 (2011).
41. Zhou, P. et al. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife* **7**, e28728 (2018).
42. Pisanello, M. et al. Tailoring light delivery for optogenetics by modal demultiplexing in tapered optical fibers. *Sci. Rep.* **8**, 4467 (2018).
43. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).
44. Wu, M. & Goodman, N. Multimodal generative models for scalable weakly-supervised learning. *Adv. Neural Info. Process. Syst.* **31**, 5575–5585 (2018).
45. Farries, M. A., Ding, L. & Perkel, D. J. Evidence for "direct" and "indirect" pathways through the song system basal ganglia. *J. Comp. Neurol.* **484**, 93–104 (2005).

Acknowledgements The authors thank M. Booze for animal husbandry and K. Franks, F. Wang and D. Purves for editorial comments on an earlier version of this manuscript. This work was supported by NIH R01 NS099288 (R.M.), R01 NS118424 (R.M., J.P. and T.G.), the George Barth Geller Fund (R.M.), a Broad Predoctoral Fellowship (J.S.A.) and NIH Predoctoral Fellowship F31 DC017879 (V.M.).

Author contributions J.S.A. and R.M. designed all experiments except the HVC miniscope imaging experiments, which were designed by W.L. and T.G.; J.G. and J.P. developed VAE methods to analyse acoustic and neural data; J.S.A. performed all in vivo imaging and behavioural experiments and analysed all related data, except for HVC miniscope imaging experiments, which were executed by W.L. and T.G.; V.M. performed in vitro recordings and V.M. and J.S.A. analysed resulting data; J.S.A. and J.H. performed and analysed histological experiments; J.G. analysed acoustic and neural data using VAEs; J.S.A., J.G., J.P. and R.M. wrote the manuscript; J.S.A., J.G., W.L., T.G., J.P. and R.M. edited the manuscript.

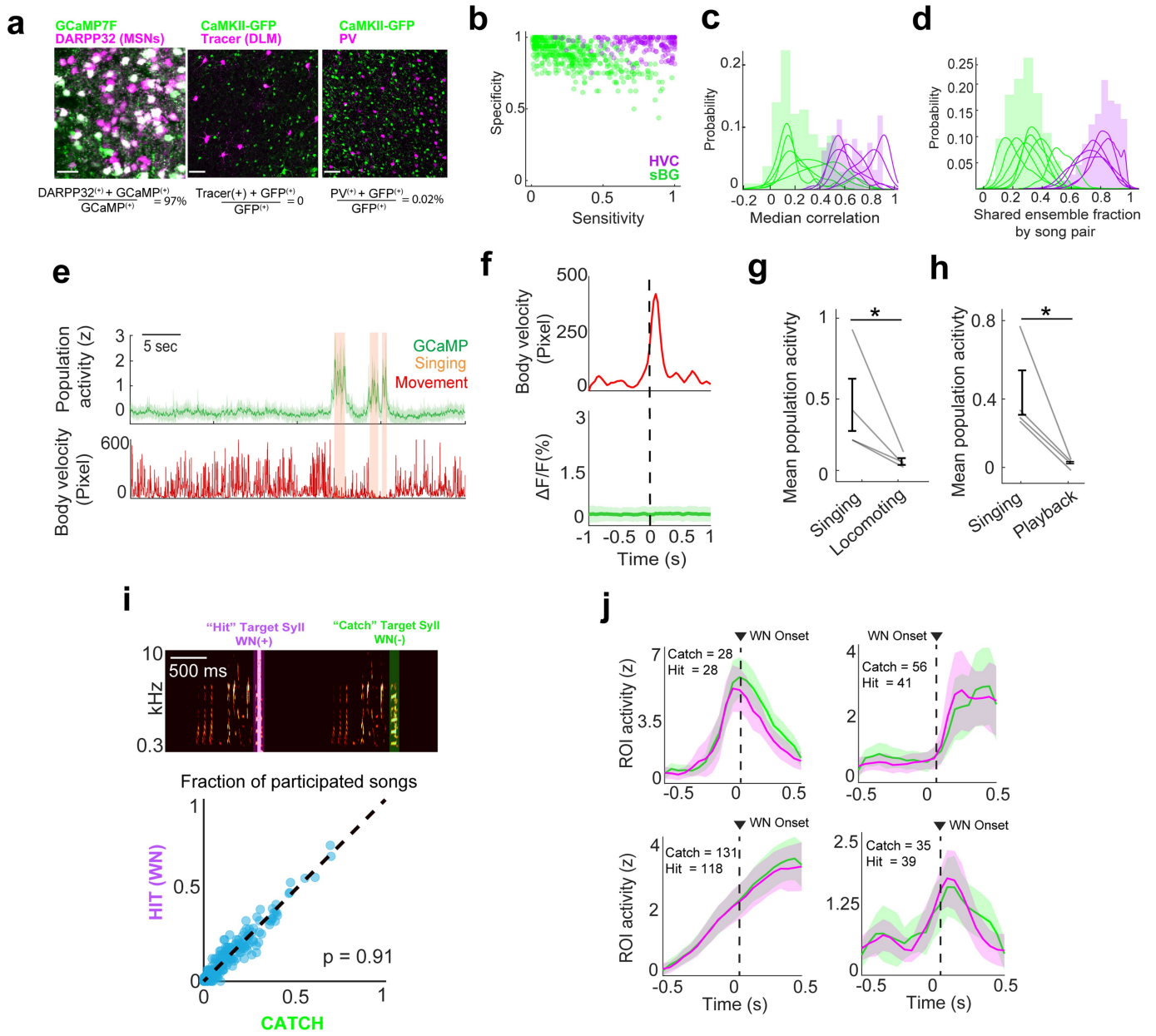
Competing interests The authors declare no competing interests.

Additional information

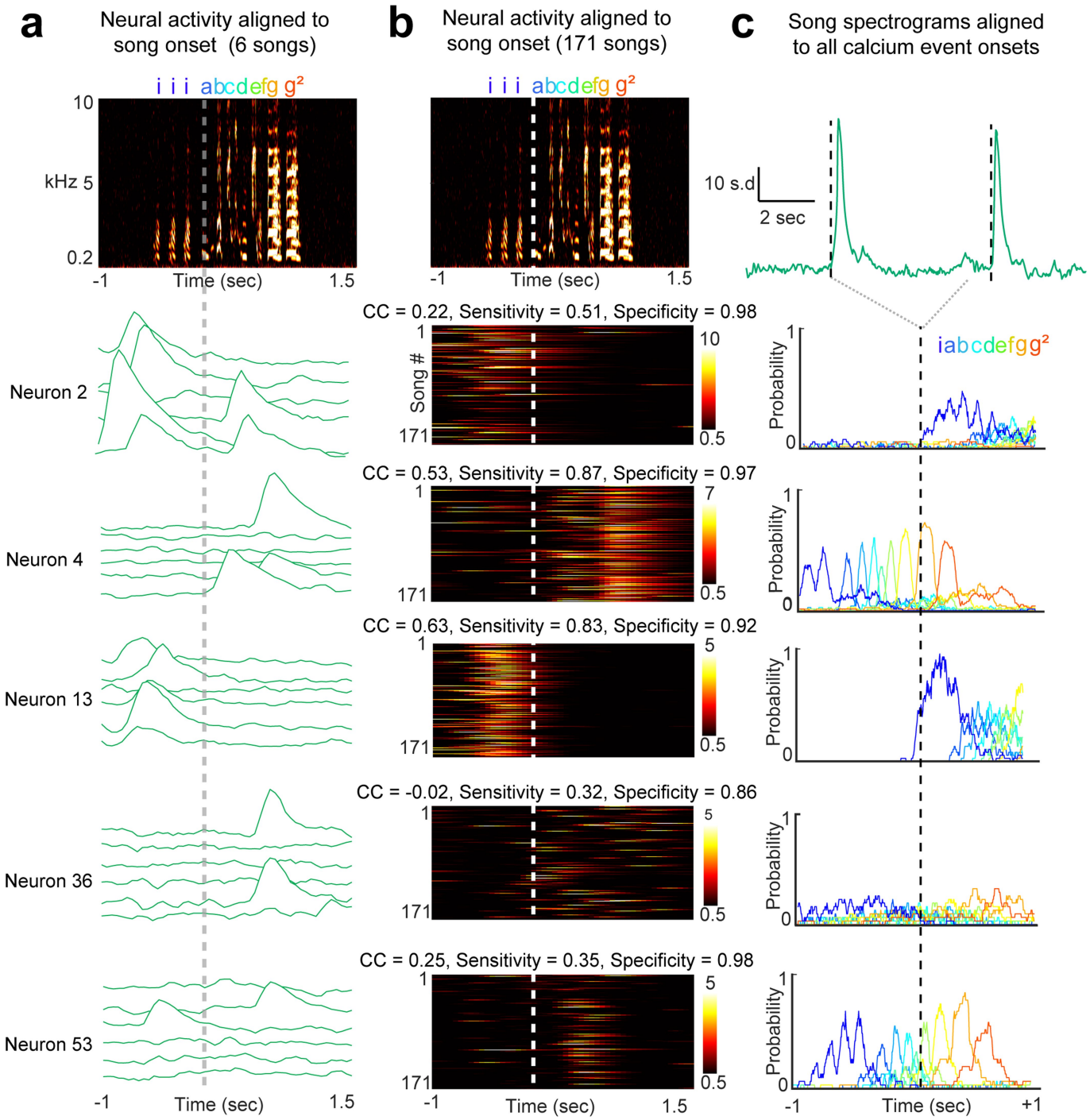
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04004-1.

Correspondence and requests for materials should be addressed to John Pearson or Richard Mooney.

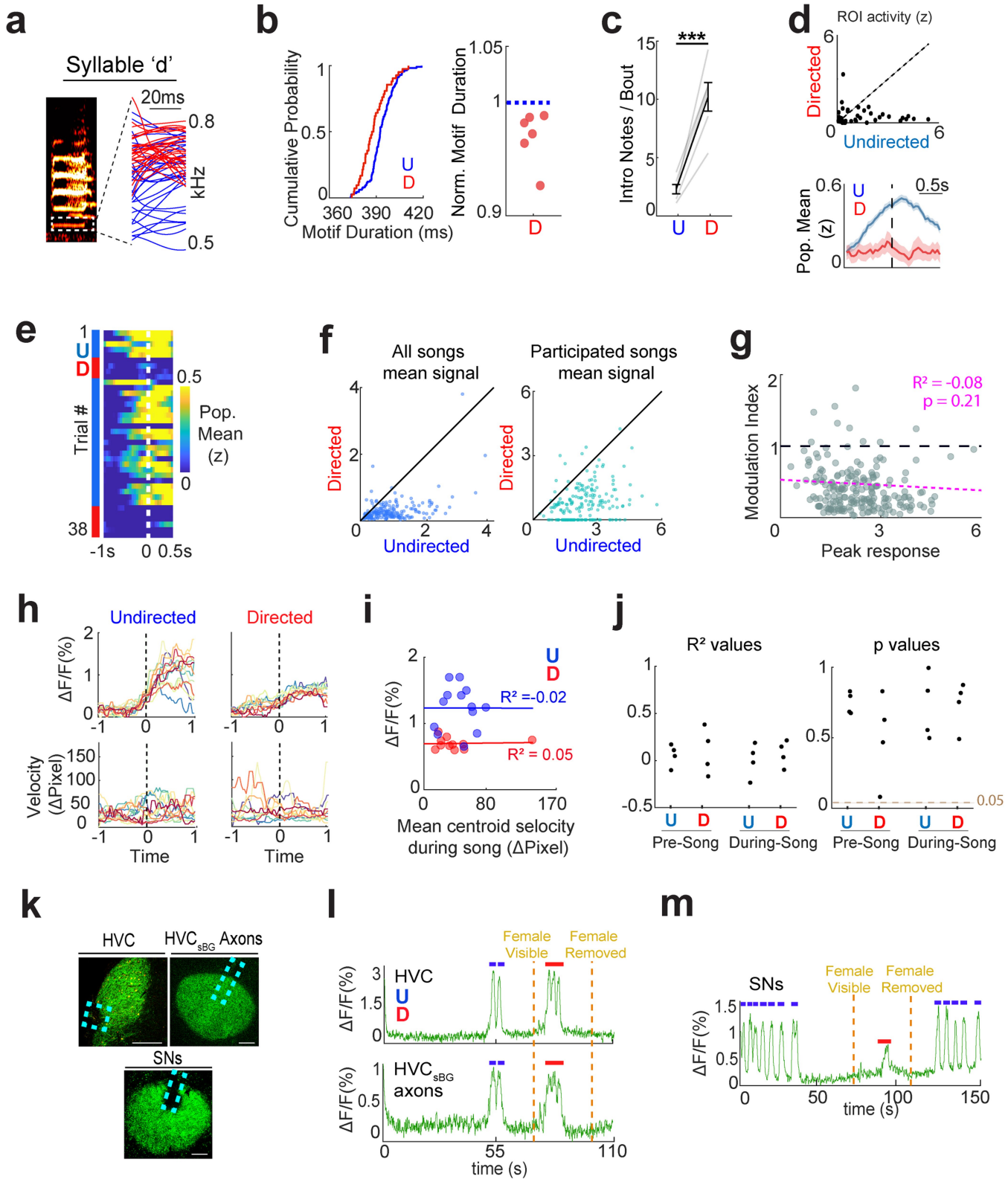
Peer review information *Nature* thanks David Robbe, Kazuhiro Wada and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Targeting and characterization of SN activity. **a**) CaMKII promoter strategy selectively labels SNs in the sBG. Left: Overlap of CaMKII-GCaMP and the SN marker DARPP-32. Middle: Superimposed image reveals no overlap between retrogradely labeled globus pallidus internus neurons and CaMKII-GFP (0/41 Tracer(+) neurons were co-labeled with GFP, n = 2 birds). Right: Superimposed images reveal almost no overlap between parvalbumin (PV) and CaMKII-GFP (5/251 PV(+) neurons were co-labeled with GFP, n = 2 birds). Scale bars = 100 µm. **b**) Specificity and sensitivity of HVC PNs and sBG SNs. **c**) Median autocorrelation for all recorded SNs and HVC PNs (median HVC autocorrelation: 0.71, SNs: 0.16). **d**) Shared fraction of active ensemble for SNs and HVC PNs across song renditions (median HVC shared fraction: 0.86, SNs: 0.28). For (**b**–**d**), SNs: n = 529 neurons from 7 birds, HVC: n = 165 neurons from 5 birds. **e**) Example of mean SN activity aligned to body velocity; orange shading denotes singing periods. Data are displayed as mean ± s.e.m. **f**) Detected movement initiations (1311 detected initiations from 1 recording session, top) aligned to SN activity from photometry recordings (bottom). **g**) Group data comparing mean SN activity during singing vs. non-singing locomotion (Student's one-sided paired *t*-test, t_3=2.464, *p = 0.0453; n = 4 birds). **h**) Group data comparing mean SN activity during singing vs. playback of the bird's own song (Student's one-sided paired *t*-test; t_3= 3.31, *p = 0.0226; n = 4 birds). **i**) Disrupted auditory feedback during singing does not acutely affect SN activity. A random 50% of song renditions were targeted for syllable-triggered white noise (top). Participation probability was not affected by the playback of white noise (Student's two-sided paired t-test; p = 0.91; n = 184 neurons from 3 birds). **j**) Example traces shown for 4 SNs comparing activity during normal singing and during singing-triggered noise. t = 0 denotes target syllable onset, dashed line is white noise onset. Only song renditions in which the cell participated were included. 0/184 neurons were found to be significantly modulated by white noise (two-sided Mann-Whitney U-test with Hochberg correction, 0/184 significantly modulated neurons from 3 birds). All error bars denote mean ± s.e.m.

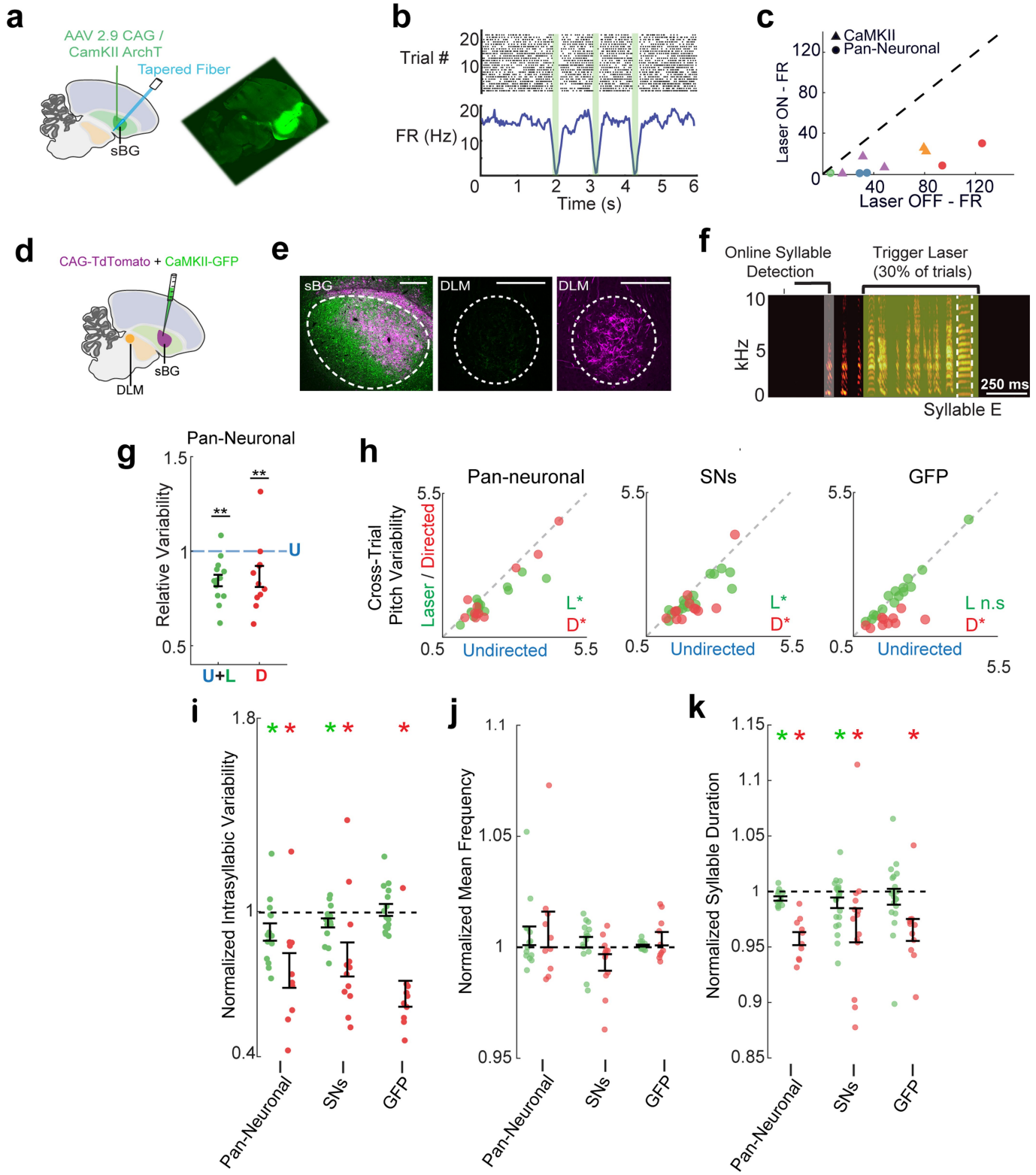


Extended Data Fig. 2 | Example song-related SN activity. a) Representative motif spectrogram (top) aligned to sample activity traces from the first 6 undirected song renditions for 5 ROIs, aligned to song motif onset (vertical dashed line; a-g, syllables; i, introductory notes). **b**) Same representative motif as (a), with activity heatmaps for all 171 trials collected throughout the day, along with the corresponding values for all-to-all correlation and sensitivity. Color scale represents z-scored fluorescence. **c**) Fluorescence trace for one neuron showing two example calcium events (top). Event-triggered probability of song syllable for the 5 neurons (bottom, see methods). All detected calcium events in the time series (27.7 minutes of concatenated recordings, 4.9 minutes with vocalizations) were used to generate the average spectrogram, which is visually represented in terms of the probability of occurrence for each syllable.



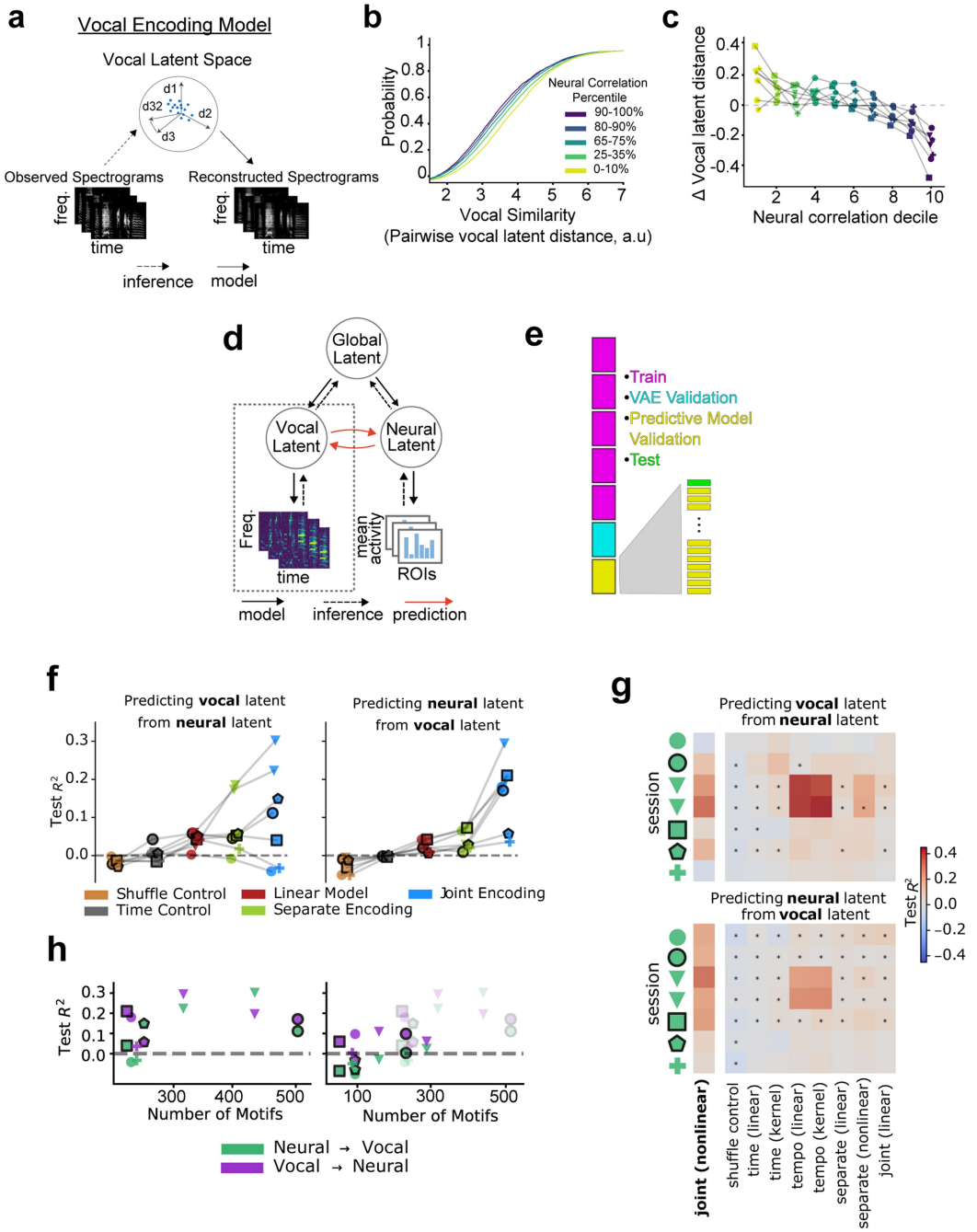
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Supplemental analyses of song, movement, and neural activity during directed and undirected song. a) Example frequency contours of syllable 'd' in undirected (blue) and directed (red) renditions. **b**) Birds with head-mounted miniscopes exhibit typical directed song features in addition to decreased pitch variability, such as faster directed motifs. Left: Cumulative distribution plot for motif durations in 1 bird. Right: group data for 6 birds (Student's two-sided paired t-test, $t5 = 1.87$ $p = 0.12$, $n = 6$ birds). **c**) Directed motifs are preceded by more introductory notes than undirected motifs. (Student's two-sided paired t-test, $t5 = -7.69$, ***$p = 0.00094$, $n = 6$ birds). **d**) Top: mean activity of 53 ROIs during directed and undirected singing from one bird. Bottom: mean SN population activity aligned to song onset. **e**) Heatmap of mean population activity for interleaved undirected and directed singing. Dashed line = onset of first syllable in motif. **f**) Left: Mean z-scored activity in undirected and directed conditions, plotted for all ROIs that were collected in directed and undirected conditions, averaged across all collected songs Right: Similar to left, but using only trials in which each neuron had a detected event ($n = 215$ neurons from 6 birds). **g**) Relationship between ROI signal (peak of averaged active trials) and the ratio between its directed and undirected activity ($n = 215$ neurons from 6 birds). Dashed line indicates no modulation ($D/U = 1$). **h)** Photometry (top) and velocity (bottom) color-matched traces aligned to undirected ($n = 13$) and directed ($n = 11$) songs. Dashed line indicates the onset of the first motif syllable. i) R values between average locomotion during song (500 ms time window) and DF/F for one bird, computed from data in (h) (f). **j**) Left: Group data showing R values comparing average song-related neural activity to movement in two conditions: averaging locomotion values over a window of 500 ms before motif onset (pre-song) or 500 ms after motif onset (during-song). Right: Corresponding p values. **k**) Representative histology of photometry recordings. Left: Histology of AAV 2/9 AxGCaMP6m.p2a.nls.tdTomato injection into HVC. Middle: HVC axons in sBG from the same bird. Right: Local injection of AAV 2/9.CaMKII.GCaMP6s into sBG. Scale bar = 50 µm. **l**) Sample recording session for dual recordings from HVC and HVC_{sBG} axons. Undirected singing, (blue) female presentation and directed singing (red) are collected in the same session. **m**) Same as (**l**), but for SN photometry. All error bars denote mean ± s.e.m.



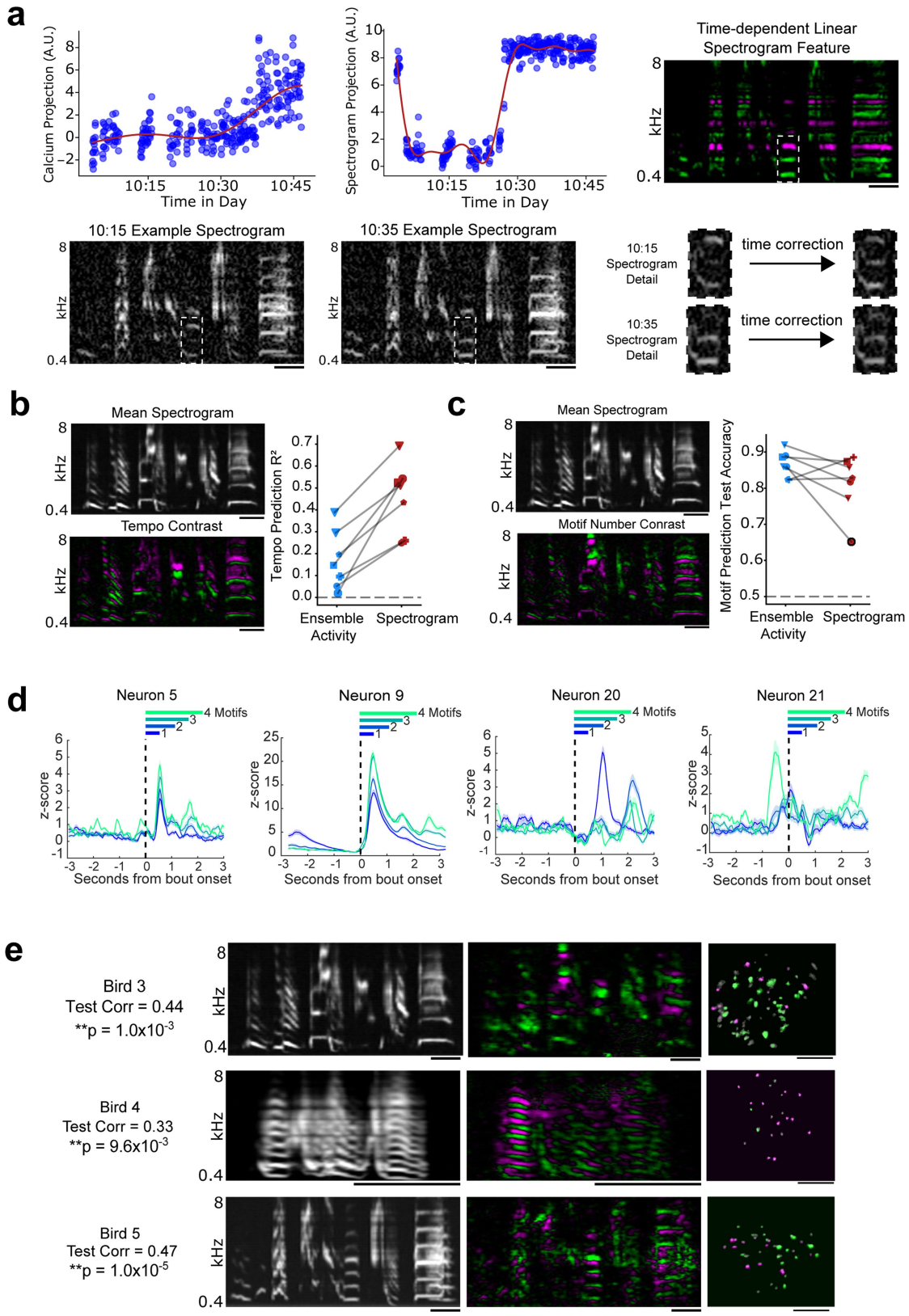
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Additional analyses of optogenetic suppression experiments. a) Experimental approach with representative histology from a CAG.ArchT injection into the sBG, both shown in sagittal view. **b**) Optrode recording in a CAG-ArchT bird showing suppressive effect of green light illumination on spontaneous action potential activity of a single sBG unit. **c**) Group data showing suppressive effects of green light illumination across neurons (CamKII, 5 neurons, CAG, 5 neurons). **d**) Sagittal schematic for coinjections of Pan-neuronal (CAG) and SN (CaMKII) fluorescent proteins into the sBG. **e**) CAG-driven expression of TdTomato (magenta) and CaMKII-driven expression of GFP (green) shown superimposed in the sBG (left) and in separate green (middle) and magenta (right) channels in the pallido-recipient thalamic nucleus DLM. Scale bar = 250 µm. **f**) Experimental approach for syllable-triggered optogenetic inhibition. **g**) Group data showing pitch variability during directed and laser-stimulated singing normalized to undirected singing for pan-neuronal inhibition. Mixed effects model, 2-sided permutation test. Laser effect size (relative to baseline: -17.8%, ***p = 0.0011, n = 14 syllables from 6 birds. Directed singing effect size: 13.8%, *p = 0.01, n = 12 syllables from 5 birds.**h**) Pitch variability group data (same data as Fig. 2j and Extended Data Fig. 4g), non-normalized, comparing values during undirected song versus either undirected + laser (L) or directed (D) conditions. **i**) Intrasyllabic variability data normalized to undirected levels. Mixed effects, 2-sided permutation test. Model fit to non-normalized data, comparing undirected and experimental (undirected + laser (green), or directed (red)) conditions (for model output details, see Tables 1 and 2 in Supplementary Information for model details, in all cases significance was assessed using a two-sided permutation test). **Pan-Neuronal**: Estimated laser effect size: −0.00071 (−10.11% of baseline) ± 0.0002, *p = 0.015. Estimated directed singing effect size: -0.0013 (-20.65%) ± 6.94, **p = 0.0098. **SNs**: Estimated laser effect size: −0.00034 (-4.09%) ± 0.00011, **p = 0.007. Estimated directed singing effect size: -0.0033 (−22.18%) ± 0.00063. **GFP**: Estimated laser effect size: −0.000054 (0.60%) ± 0.00043, p = 0.80. Estimated directed singing effect size = −0.0027 (−36.00%) ± 0.00050, ***p = 0.000082. Pan-neuronal Laser n = 14 syllables from 6 birds, directed n = 12 syllables from 5 birds; SNs: Laser n = 16 syllables from 6 birds, directed n = 12 syllables from 5 birds; GFP: Laser n = 15 syllables from 5 birds for laser, directed n = 10 syllables from 4 birds. **j**) Mean syllable frequency group data normalized to undirected levels. **Pan-Neuronal**: Estimated laser effect size: 3.55 ± 3.14 Hz, p = 0.27. Estimated directed singing effect size: 9.91 ± 6.94 Hz, p = 0.17. **SNs**: Estimated laser effect size: -8.28 ± 4.54 Hz, p = 0.079. Estimated directed singing effect size: −13.95 ± 9.13 Hz, p = 0.14. **GFP**: Estimated laser effect size: 0.60 ± 0.37 Hz, p = 0.12. Estimated directed singing effect size = 3.041 ± 2.46 Hz, p = 0.23. Pan-neuronal laser n = 14 syllables from 6 birds, directed n = 12 syllables from 6 birds; SNs: Laser n = 16 syllables from 6 birds, directed n = 12 syllables from 5 birds; GFP: Laser N = 15 syllables from 5 birds for laser, directed n = 10 syllables from 4 birds. **k**) Mean syllable duration group data normalized to undirected levels. **Pan-Neuronal**: Estimated laser effect size: -0.58 ± 0.23 ms, *p = 0.016. Estimated directed singing effect size: −0.65 ± 0.28 msec, *p = 0.035. **SNs**: Estimated laser effect size: -0.82 ± 0.36 ms, **p = 0.029. Estimated directed singing effect size −2.76 ± 0.62 ms, ***p = 0.00016. **GFP**: Estimated laser effect size: -0.84 ± 0.63 ms, p = 0.19. Estimated directed singing effect size = −4.37 ± 0.80 ms, ***p = 0.000030. Pan-neuronal: Laser N = 14 syllables from 6 birds, directed N = 12 syllables from 6 birds; SNs: Laser N = 16 syllables from 6 birds, directed n = 12 syllables from 5 birds; GFP: Laser n = 15 syllables from 5 birds for laser, directed n = 10 syllables from 4 birds. Data are displayed as mean ± sem. All error bars denote mean ± s.e.m. Pan-neuronal: Laser N = 14 syllables from 6 birds, Directed N = 12 syllables from 6 birds; SNs: Laser N = 16 syllables from 6 birds, Dir N = 12 syllables from 5 birds; GFP: Laser N = 15 syllables from 5 birds for laser, Dir N = 10 syllables from 4 birds. Data are displayed as mean ± sem. All error bars denote mean ± s.e.m.



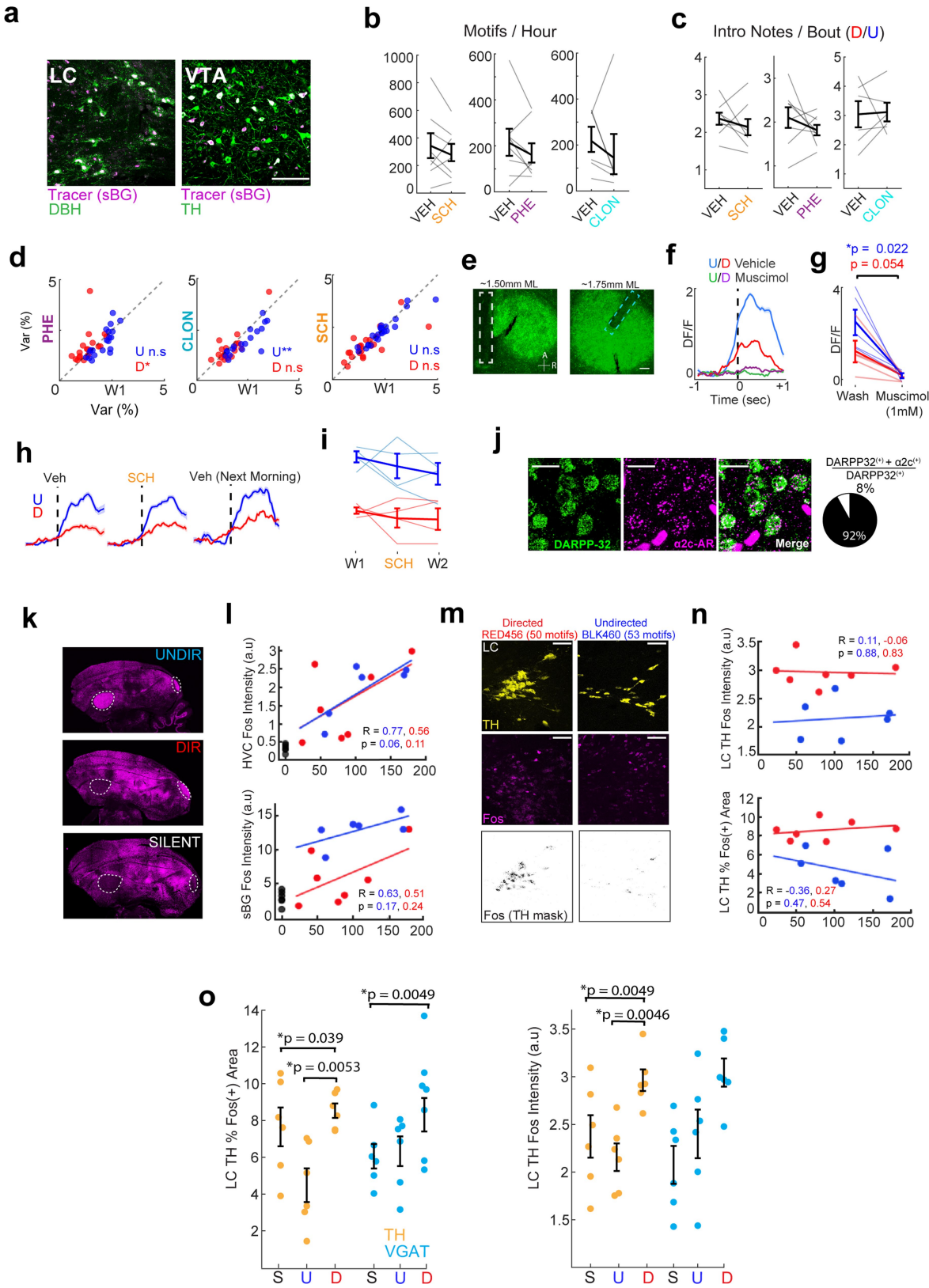
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Joint encoding model details and comparison to alternate models. a) Schematic for learning low dimensional latent features of motif spectrograms using a variational autoencoder (VAE) approach. The model learns a compressed representation of the data that is sufficient to reconstruct the original. **b**) Cumulative distribution of pairwise song distances in VAE latent space, grouped by the similarity of the associated neural patterns (neural correlation percentiles, yellow to blue). For more dissimilar neural activity (yellow), songs are farther apart in VAE space, while more highly correlated neural activity (blue) shifts the distribution to the left, implying overall more similar songs, as indicated by smaller VAE vocal latent distances. **c**) Group data showing median VAE distance (relative to the mean) within each neural correlation decile for all 7 sessions from 5 birds; pairs of trials with highly correlated neural activity patterns are closer in VAE latent space. Marker shape denotes bird identity. **d**) Schematic of the joint modeling approach. Acoustic data is modeled using a VAE as before (boxed region) and a second VAE is used to model the neural data. A global latent variable is then used to capture shared variation in the two modalities. **e**) Schematic of model training and validation. VAE models were trained using sevenfold cross-validation. Within each fold, data were partitioned into seven tranches, five for VAE model training (magenta), one for VAE model validation and hyperparameter selection (cyan), and one for assessing model performance (yellow). For the VAE model, average performance on the yellow test set across the seven cross-validation folds is reported. For predictive models trained to predict one set of latents from another, a "leave-one-out" strategy on the yellow data set (right) was used to select predictive model hyperparameters and assess performance. **f**) Joint encoding outperforms a collection of control models. The shuffle control randomly pairs spectrograms and ROI activity vectors. The time control uses time-in-session to predict the joint encoding model's neural latents (left) and vocal latents (right). The linear model comprises independently trained neural and vocal variational autoencoders (as in Fig. 3a without the global latent), with emission and recognition networks restricted to linear mappings. The separate encoding model comprises independently trained neural and vocal variational autoencoders with emission and recognition models parameterized by deep neural networks. The joint encoding model is the full model as presented in Fig. 3a. For all models, prediction is performed using ridge regression and test performance is evaluated using the cross-validation procedure described in Methods. Average test set performance over 7 cross-validation folds of each of 7 sessions from 5 birds is shown. Each line represents a single bird-session. **g**) Model comparison split by experimental session. Performance (measured by R^2) for the task of predicting vocal latents from neural latents (top) and vice versa (bottom) for each of 7 sessions from 5 birds. In addition to the models presented in b, the comparison includes models using motif tempo to predict joint encoding neural latents (top) and vocal latents (bottom); using kernel ridge regression in place of linear ridge regression (with leave-one-out regularization strength and radial basis function bandwidth selection); and a version of the joint encoding model with emission and recognition networks restricted to linear mappings. Joint encoding predictive performance is compared with each control model for each experimental session (one-sided Wilcoxon signed-rank test, * denotes $p < 0.05$). For both imaging sessions of one bird (bird 5, denoted by triangles in panels b–d), both neural latents and vocal latents could be robustly predicted from song tempo. **h**) Left: Predictive performance versus number of song motifs (left) for each of 7 experimental sessions. Poor predictive performance is observed for experimental sessions with fewer than 300 motifs and fewer than 50 ROIs (not shown). Symbols denote birds, as in panels b and c. Right: Similar to left. Opaque markers indicate performance using only first motifs in each bout, faded markers performance indicate performance using all motifs.



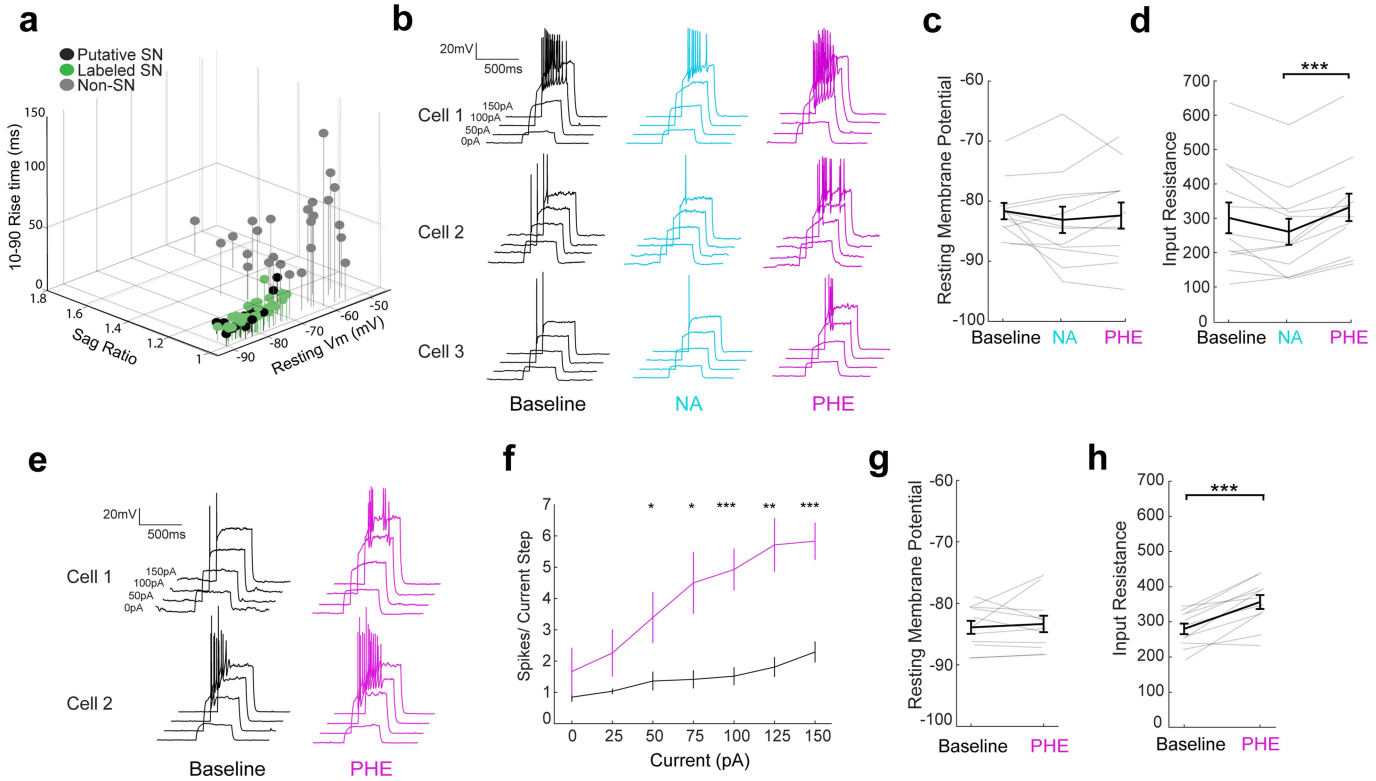
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Joint encoding model preprocessing and additional examples. a) To minimize time confounders, components of calcium activity vectors (top left) and spectrograms (top middle) that could reliably be predicted by time-of-day were removed (red lines; see Methods). The calcium and spectrogram residuals after prediction are used for further analysis in place of the original data (top right). Positive weights are shown in green and negative weights in magenta. Note that the effects are restricted to regions with vocalization. For two example spectrograms from 10:15 (bottom left) and 10:35 (bottom middle), time-of-day correction makes the resulting syllables more similar to one another. Scale bar for right column: 100 ms. **b**) Left: Despite time warping, spectrograms show consistent tempo-related changes. Difference plot between the average faster-than-median spectrograms and the average slower-than-median spectrograms (bottom, positive values in green, negative in magenta) for one example bird (Bird 3, squares). The consistent horizontal bands throughout the motif indicate upward pitch shifts associated with faster tempos, which were observed for almost all experimental sessions. Scale bars denote 100 ms. Right: Both ensemble activity and warped spectrograms contain information about tempo. For each experimental session, tempo can be predicted from ensemble activity vectors (blue) and spectrograms (red) after both signals have been corrected for time-of-day. Dotted line denotes chance performance. Scale bars denote 100 ms. **c**) Spectrograms also show consistent motif-number-related changes. For the same example bird as in b, the average of the first motifs in every bout and the average of all other motifs exhibit clear differences (bottom, positive values in green, negative in magenta). Right: Both ensemble activity and time-warped spectrograms contain information about motif number. For each experimental session, motif number (first motif vs. rest) could be reliably predicted from ensemble activity vectors (blue) and spectrograms (red) using the same procedure described for tempo prediction (reporting test accuracy, weighted by class so that chance performance is 0.5). Dotted line denotes chance performance. Scale bars denote 100 ms. **d**) Example average ROI activity aligned to the first syllable of bouts consisting of 1, 2, 3 or 4 motifs. Note that ROIs 20 and 21 display qualitatively different activities in bouts of different lengths. **e**) Weighted average generated spectrograms and ROI activity pairs, with weights given by their projection along the correlation axis, describe how song spectrograms (middle column) and neural activity (right column) vary together. P-values refer to corresponding correlations of held-out test data, as in Figure 3c. Scale bars for left and middle columns: 100 ms. Scale bars for right column: 250 μm.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Effects of adrenergic signaling manipulations on song and neural activity. a) Retrograde labelling of dopamine beta hydroxylase (DBH) and tyrosine hydroxylase (TH) positive cell bodies in the locus coeruleus (LC) and ventral tegmental area (VTA), respectively, following retrograde tracer injections into the sBG. Scale bar = xx microns and applies to both panels. **b**) SCH, PHE, and CLON infusion do not significantly affect singing rates (Student's two-tailed paired t-test, CLON: $t_5 = 0.81$, p = 0.46, n = 6 birds; PHE: $t_7 = 0.97$, p = 0.28, n = 8 birds; SCH: $t_7 = 1.17$, p = 0.36, n = 8 birds). **c**) SCH, PHE, and CLON infusion do not significantly affect number of introductory notes per bout (Student's two-tailed paired t-test, CLON: $t_5 = -0.16$, p = 0.88, n = 6 birds; PHE: $t_7 = 1.05$, p = 0.33, n = 8 birds; SCH: $t_7 = 0.91$, p = 0.30, n = 8 birds). **d**) Left: Effects of PHE on pitch variability in directed and undirected song. Mixed effects model, 2-sided permutation test, see Tables 1 and 2 in Supplementary Information for model details. Estimated effect of drug presence on directed: 0.36 (26.1% of baseline) ± 0.17, *p = 0.02. Estimated effect of drug presence on undirected: −0.14 (6.7% of baseline) ± 0.12, p = 0.24. Middle: Effects of CLON on pitch variability in directed and undirected song. Estimated effect of drug presence on DIR: −0.025 (1.5% of baseline) ± 0.10, p = 0.1. Estimated effect of drug presence on UNDIR: −0.36 (14.8% of baseline) ± 0.12, **p = 0.0049. Right: Effects of SCH on pitch variability in directed and undirected song. Estimated effect of drug presence on directed: 0.081 (4.9% of baseline) ± 0.22, p = 0.71. Estimated effect of drug presence on undirected: 0.048 (2% of baseline) ± 0.84, p = 0.57. **e**) Representative histology showing photometry probe and microdialysis probe placement into sBG for simultaneous drug delivery and photometry. Scale bar = 100 µm. **f**) Representative DF/F measurements during directed and undirected singing for one bird before and after (>1 hour) beginning muscimol infusion. **g**) Muscimol infusion suppresses calcium signals recorded in the sBG during both directed and undirected singing (Student's two-tailed paired t-test; $t_4 = 3.63$, p-values are indicated; n = 5 birds). **h**) Sample traces for SN imaging during infusion of SCH23390 into the sBG. **i**) Group data showing mean SN photometry signals during SCH23390 infusion in undirected and directed conditions (n = 4 birds). **j**) DARPP32 and α2c-AR mRNA co-expression sBG SNs (n = 3 birds). Scale bar = 20 µm. **k**) Low power confocal images showing Fos mRNA expression in a sagittal section of the finch brain across behavioral conditions. Dashed white outlines highlight the sBG and HVC. **l**) Fos intensity levels in HVC and the sBG plotted against motif count (30-minute window) in either directed (red) or undirected (blue) singing conditions. For all immediate early gene experiments, undirected n = 6 birds, directed n = 7 birds, silent n = 6 birds. **m**) Example confocal image z-stack collected in the LC. The intensity and area of Fos puncta (magenta) were quantified within the TH-positive mask (yellow). Scale bar = 50 µm. **n**) Mean Fos intensity and area within LC TH mask plotted against for directed (red) and undirected (blue) motif counts. **o**) Group data for Fos intensity (left) or area (right) plotted for TH and VGAT masks during either directed (red, N = 7 birds) or undirected (blue, N = 6 birds) singing conditions. One-way ANOVAs with post hoc Tukey tests were performed separately for TH and VGAT masks under each condition. Post hoc comparisons for significant ANOVAs are displayed. Fos mRNA puncta Intensity: TH mask, $F_{(2,16)} = 7.46$, **p = 0.0051, VGAT mask, $F_{(2,16)} = 7.4$, **p = 0.0053. Fos mRNA Area: TH mask, $F_{(2,16)} = 9.02$, p = 0.0024, VGAT mask, $F_{(2,16)} = 3.6$, p = 0.051.



Extended Data Fig. 8 | Effects of adrenergic signaling on SN excitability. **a**) Rise time, sag, and resting membrane potential can be used to distinguish SNs from non-SNs in the sBG (see Methods). **b**) Three more example SNs recorded during baseline, NA, and PHE. **c**) Effect of NA and PHE on SN resting membrane potential (One-way repeated measures ANOVA with Greenhouse-Geisser correction. $F_{(1.487, 16.36)} = 0.5950$ $p = 0.51$; $n = 11$ cells). **d**) Effect of NA and PHE on SN input resistance (One-way repeated measures ANOVA with Greenhouse-Geisser correction and post-hoc Tukey test. $F_{(1.229, 13.52)} = 7.980$; Baseline vs NA: $p = 0.054$; NA vs PHE: ***$p = 0.0003$; $n = 11$ cells). **e**) 2 example SNs recorded during baseline and PHE, from a different experiment than **a**–**d**. **f**) F-I curves showing increased action potentials in response to positive current injection for baseline and PHE conditions ($n = 12$ cells). **g**) Effect of PHE on SN resting membrane potential (Student's two-tailed paired t-test, $t_{10} = 0.55$, $p = 0.59$; $n = 11$ cells, separate from those shown in panel **c**). **h**) Effect of PHE on SN input resistance (Student's two-tailed paired t-test, $t_{10} = 5.42$, ***$p = 0.0003$; $n = 11$ cells, separate from those shown in panel **d**).

nature research

Corresponding author(s): Richard Mooney

Last updated by author(s): August 19, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For miniscope recordings, software by Inscopix (NVISTA HD) was used in combination with MATLAB (MathWorks) to acquire calcium imaging data. For photometry recordings, Bonsai (open-ephys) was used to collect webcam video, audio, and calcium signals. For image collection with a confocal microscopy, ZEN software version 3.3 (Zeiss) was used. Custom LABVIEW code was used to collect singing data in optogenetic and microdialysis experiments.
Data analysis	Autoencoded Vocal Analysis (v0.3), the Python package used to generate and warp spectrograms for the VAE analysis, is freely available online: https://github.com/pearsonlab/autoencoded-vocal-analysis". For image analysis, ImageJ with Fiji version 2.1.0/1.53c (https://fiji.sc/) was used. For other data analysis and generation of figures, custom scripts and functions in MATLAB were used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We are in the process of mounting core data on the Research Data Repository, a service managed by the Duke University Libraries (https://research.repository.duke.edu).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size estimation was performed beforehand. Our sample sizes were determined based on sample SD in the earlier sets of experiments, and largely conform to convention in the field (Hisey E, Kearney MG, Mooney R. A common neural circuit mechanism for internally guided and externally reinforced forms of motor learning. (Nat Neurosci 2018; 21: 589–597. Liberti WA 3rd, Markowitz JE, Perkins LN, Liberti DC, Leman DP, Guitchounts G et al. Unstable neurons underlie a stable learned behavior. Nat Neurosci 2016; 19: 1665–1671.).
Data exclusions	Experiments with unsuccessful surgery, injection, implantation, and expression were excluded from the data.
Replication	The results are based on behavior and recordings from multiple birds and multiple neurons (as described in the text) and the reproducibility of the findings are shown in the scatter plots and other accompanying figures.
Randomization	Animals were randomly allocated into experimental groups.
Blinding	No blinding was performed because longitudinal experiments in individual birds, each with unique songs, rendered blinding difficult to perform.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Rabbit primary antibody for TH (AB152; MilliporeSigma). Rabbit primary antibody for DBH (22806; ImmunoStar). Mouse primary antibody for GFP (A11120; Thermofisher). Anti-rabbit secondary antibody (111545144; Jackson ImmunoResearch).
Validation	The primary antibodies have been widely used in rodents, and have been validated in birds. http://www.emdmillipore.com/US/en/product/Anti-Tyrosine-Hydroxylase-Antibody,MM_NF-AB152 http://www.immunostar.com/shop/antibody-catalog/dbh-dopamine-beta-hydroxylase-antibody/ https://www.thermofisher.com/antibody/product/GFP-Antibody-clone-3E6-Monoclonal/A-11120 In our study, the antibody for TH labeled large cell bodies in VTA, indicating it labels dopaminergic neurons. In our study, the antibody for DBH labeled many neurons in the locus coeruleus, demonstrating its validity as a marker for noradrenergic neurons.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Results were collected from adult (>80 d post hatch) male zebra finches (Taeniopygia guttata).
Wild animals	Not used.

Field-collected samples	Not used.
Ethics oversight	All experiments were performed in accordance with a protocol approved by Duke University Institutional Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.