Latent state dynamics in female *Drosophila* during social interactions

Umesh K. Singla[1,*] **Albert Lin**[1,2] **Jonathan W. Pillow**[1] **Mala Murthy**[1]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[2]Center for the Physics of Biological Function, Princeton University, Princeton, NJ, USA
*Correspondence: usingla@princeton.edu

Abstract

Social interactions across animal species are governed by the interplay between multimodal sensory information and an animal's internal state. Here, we investigate this interplay in female *Drosophila* as she engages with the male during courtship, a highly dynamic social behavior. While male behaviors during courtship, such as song production, have been well characterized moment by moment, the female's actions have not been described with similar temporal precision, despite her central role in determining copulation outcomes. Her behavior displays high variability and is often viewed as volitional, raising questions about its structure and predictability at fine temporal resolution. To address this, we used a state-space model that combines Generalized Linear Models (GLMs) with a Hidden Markov Model (HMM) to uncover latent states that modulate the relationship between male sensory cues and female responses. We find that, overall male cues weakly predict female behavior, but that predictive power varies substantially across inferred states: some states exhibit clear cue-driven structure, while others show reduced sensitivity to external cues and more internally-driven behavior. At short timescales, female behavior appears only weakly predictable and highly variable, yet, at longer timescales, a rich latent state structure emerges, hinting at internal gating and evaluation of social signals over time. This work provides the first moment-by-moment characterization of female behavior during courtship, taking a crucial step toward closing the loop in social behavior modeling.

1 Introduction

Animal communication is shaped not only by the external signals individuals receive, but also by the internal states that modulate how those signals are interpreted, and acted upon. In *Drosophila melanogaster*, males produce structured songs and pursuit behaviors that have been quantified at subsecond resolution, providing a model system for studying social signals. During courtship, *Drosophila* males orient toward females, follow them closely, and produce species-specific courtship songs by extending and vibrating a single wing (Figure 1a). These songs consist of distinct motifs (pulses and sine waves) that are interspersed with other courtship elements including tapping the female with their forelegs and attempted copulation [1, 2, 3]. Male courtship behaviors have been extensively characterized and modeled, female responses during courtship have however received far less attention. Female trajectories during courtship vary widely—when grouped by male position and song onset, female trajectories show only a weak directional bias (Figure 1b,c). Female responses are often treated as noisy and volitional, despite her behavior being critical to the outcome of the courtship [4]. In this work, we explore the question of whether variability in female's behavior reflects underlying internal states that structure her engagement with the male cues (Figure 1d).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for non-human animal communication.

Here, we use a multi-state GLM-HMM framework to model how females integrate multimodal male cues (visual, auditory, and tactile) across multiple behavioral outputs: continuous locomotor responses (her forward, lateral, and angular velocities) and a discrete wing flicking behavior (Figure 1e-f). Across courtship interactions, we find that male cues weakly predict female actions in aggregate, but predictive power varies sharply across latent states. We identify states with clear cue-driven dynamics alongside states that appear internally-driven, revealing that courtship is structured by discrete modes of sensory-motor coupling rather than a uniform stimulus-response mapping. Overall, female's responses at millisecond timescales are highly variable but a structured latent dynamics emerge over longer windows (minutes).

2 Methodology

Social behavior quantification. To investigate how female flies respond to male cues during natural courtship, we used high-resolution pose tracking (SLEAP; [5]) to extract the trajectories and body keypoints of pairs of male and female flies as they freely interacted (Figure 1a). Simultaneously, we recorded the male's courtship song using a 9-mic array, allowing precise segmentation of his song into pulse and sine components (Figure 1a). We extracted detailed behavioral readouts by tracking 13 keypoints on the female's body, allowing us to quantify her forward, lateral, and angular velocity, as well as a discrete rejection behavior called wing flicking (Figure 3b). To relate female behaviors to male actions, we characterized the set of multimodal sensory inputs or feedback cues the female receives from the male, spanning visual, auditory, and tactile modalities (Figure 1f).

A multi-state model to predict female behavior. We fit a GLM–HMM in which the female transitions between discrete latent states, each with its own linear mapping from male sensory inputs to behavioral outputs. Specifically, male sensory features from the preceding 3 seconds were used to predict female forward, lateral, and angular velocities, as well as wing flicking, at a time t with regression weights that depend on the current latent state (Figure 2a; See Appendix). Each latent state z_t defines a separate GLM relating the male cue history s_t to the observed female behavior y_t. The female's movement variables, such as her forward, lateral, and angular velocity, are continuous and modeled as a Gaussian distribution whose parameters depend on her current latent state z_t and cue history s_t:

$$p(y_t \mid z_t, s_t) = \mathcal{N}(y_t \mid w_{z_t} s_t + b_{z_t}, \sigma^2_{z_t})$$

For wing flicking ($y_t = 1$) which is a discrete behavior, the model uses a logistic function to describe the probability of flicking:

$$p(y_t \mid z_t, s_t) = \sigma(w_{z_t} s_t + b_{z_t})$$

where $w_k \in \mathbb{R}^M$ denotes the GLM weights for latent state $k \in \{1 \ldots, K\}$. Note that since wing flicking behavior is rare, we excluded this output from the state estimation process to avoid it disproportionately influencing the latent states. Instead, we used state assignments inferred from the GLM–HMM fit to velocity data and trained separate GLMs to predict wing flicking within a state.

The full set of model parameters, $\theta \equiv \{\pi, A, w_k, b_k, \sigma^2_k\}$, is learned using the expectation–maximization (EM) algorithm (See Appendix).

3 Results

5-state GLM-HMM. We trained the GLM–HMM to predict female behavior using 80% of the fly pairs and evaluated its predictive performance on the rest. To assess model quality, we computed the difference in log-likelihood between the GLM–HMM and the Chance model (see Appendix). We fit GLM-HMMs with varying numbers of latent states. Note that the one-state GLM–HMM is simply a standard GLM with no internal state. We found that a five-state GLM–HMM achieved a substantial improvement in predictive performance on the held-out fly pairs: 220 bits/s over the Chance model, compared to 115 bits/s for the GLM. Increasing the number of states beyond 5 did not yield significant additional gains (Figure 2b). Models with 5–7 states consistently outperformed the GLM, suggesting that latent internal dynamics offer predictive value beyond what can be captured by sensory cues alone.

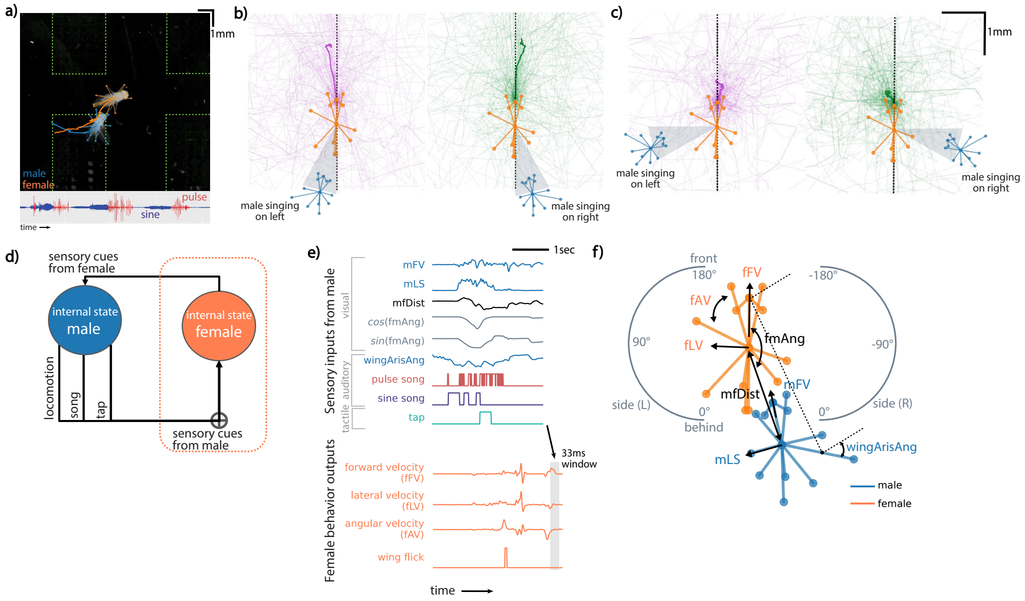

Figure 1: **Female behavior during courtship is variable and multimodal.** **(a)** Top: Example frame from a courtship video showing SLEAP-tracked male and female flies. Dashed green lines indicate 9 microphones used to record courtship song. Bottom: recorded courtship song from the male segmented into sine and pulse components. **(b, c)** Female trajectories following male song onset. Example 1-second trajectories of the female (orange) following male pulse song onset, separated by whether the male was singing from her left or right. Trajectory samples are aligned to the female's position and orientation at song onset (n=200). The male is positioned within the gray 30° sector. Trajectories reveal high variability in female responses. Bold indicates the average female trajectory in each condition. **(d)** Schematic of the bidirectional social loop in *Drosophila* courtship. The orange box highlights the portion of the loop studied in this work. **(e)** Example 3-second time series of male sensory cues (top) used to predict female behavior in the next 33ms time bin (below). Sensory cues: visual (mFV - male forward velocity, mLS - lateral speed, fmAng - male orientation w.r.t. female, and mfDist - distance), auditory (sine song, pulse song and wingArisAng - alignment between female's aristae and the singing male wing), and tactile (tap). Outputs: fFV - female forward velocity, fLV - lateral velocity, fAV - angular velocity, and wing flicking. **(f)** Schematic summarizing the relationship between sensory inputs and female outputs. See Figure 3 for details on tap and wing flick calculations.

Evaluating performance by behavior, by state and overall. We quantified how well the model captured moment-by-moment variation in female behavior by computing the Pearson correlation between predicted and actual velocities. As shown in Figure 2c, the model achieved consistent predictive performance across animals in both training and held-out datasets. A breakdown across all three velocity components—forward, lateral, and angular—revealed highest correlations for forward motion, with modest correlations for lateral and angular velocities as well as modest F1 scores for wing flicking behaviors (Figure 2d).

We next examined the properties of the latent states inferred by the GLM–HMM. Individual flies occupied multiple states during courtship, with most animals spending substantial time in States 3–5 (Figure 2f). To better understand how predictive performance varied by state, we computed the correlation between predicted and actual velocities within each state. Some states (e.g., State 2) were more predictable than others on average, however, there was a considerable amount of variability within flies (Figure 2g). While correlation scores decreased slightly on held-out data (Figure 2g, right), the relative pattern was preserved, suggesting that these differences reflect some meaningful pattern rather than overfitting.

Retrieved latent states are structured across timescales. We next investigated what behavioral features the GLM–HMM states correspond to and when they occur during courtship. State 1 (colored blue in (Figure 4)) appeared infrequently, typically at the start of sessions, and was marked by high

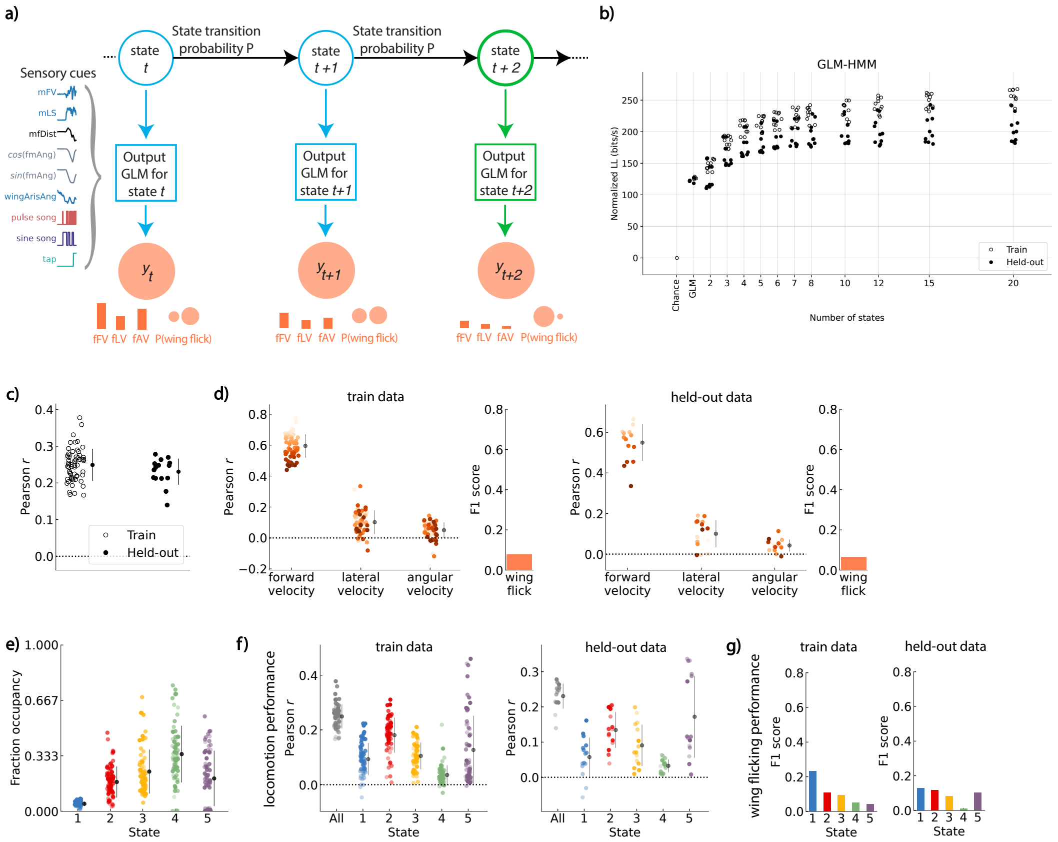

Figure 2: **Model architecture and performance.** **(a)** Schematic of the GLM-HMM framework. At each time point *t*, the female is assumed to be in a latent state that determines how multimodal male sensory cues are linearly weighted to predict her motor outputs: forward (fFV), lateral (fLV), and angular velocity (fAV), along with probability of a wing flick (0/1). **(b)** Normalized log-likelihood (LL) on held-out fly pairs (in bits/s; see Appendix). Each circle represents one courtship pair. **(c)** Pearson correlation scores between predicted and observed female behavioral outputs, for both training (60 fly pairs) and held-out test data (15 fly pairs), using the 5-state GLM-HMM. Each open or filled circle represents one courtship pair. **(d)** Pearson correlation scores broken down by behavioral variable—forward velocity (fFV), lateral velocity (fLV), and angular velocity (fAV)—for training (left) and held-out (right) fly pairs, as well as model performance on predicting wing-flicking evaluated by F1 score metric. Each dot represents one courtship pair. Within each panel, flies are color-coded consistently: the same shade of orange denotes the same fly pair across behaviors. **(e)** Fractional occupancy of each latent state across fly pairs in the 5-state GLM-HMM. Each dot denotes the proportion of time a given fly pair spent in a particular state. **(f)** Female behavioral predictability by latent state, using soft state assignments from the GLM-HMM, on training data (left) and held-out (right). Pearson correlation between predicted and observed outputs is computed within each state by weighting time points proportionally to their inferred state probabilities (See Appendix). **(f)** Wing flicking F1 scores by state. Black markers indicate the mean ± 1 s.d. in all plots.

velocities in both animals, large inter-fly distances, and minimal song and tap cues (Figure 4a-b). We interpret this as a chamber familiarity state at the start of a session (Figure 4e-f) and not reflective of the courtship, and we exclude it from further analysis.

State 2 (red) reflects an active locomotor state of the female, marked by her moderate to fast movement with the male following the female at some distance, circling and producing both song and tapping cues. This state occurs throughout but declines over time. State 3 (yellow) represents moderate engagement, with the female moving slowly in close proximity to the male, who circles behind her with consistent song and tapping. Manual video inspection suggests the female responds subtly through side-stepping, turning, or small shifts in position during this state. This state is sustained across the session (Figure 4c). State 4 (green) corresponds to a low-activity period in which the female is largely stationary and often engaged in grooming, while the male remains nearby with minimal movement and singing or tapping.

Finally, State 5 (purple) reflects full stillness of the female, with neither locomotion nor grooming, as also confirmed by video inspection. The courtship song and tap cues are minimal in this state, the male remains positioned closely to the female however with little motion. This state often persists for several seconds and becomes increasingly common as courtship progresses (Figure 4e), though it drops sharply just before copulation, when the female typically transitions back into State 3. This prolonged stillness may reflect a critical phase of "assessing" the male or decision-making, but that interpretation remains speculative. It aligns with prior work showing that average female speed decreases in response to courtship song over timescales of seconds to minutes [3, 6, 7].

Latent states are defined by distinct mappings between feedback cues and female locomotion behavior. Although States 3, 4, and 5 all have low female velocity (Figure 4a-b, Figure 7b), they differ markedly in how the female behavior couples to male cues in each state. A summary of obtained GLM filters is visualized in Figure 5a, broken down by behavior, confirm this distinction, revealing low cue sensitivity in States 3–4 but tuning to male cues in State 2 and 5. In State 3, the female shows weak but detectable responsiveness, occasionally adjusting her position in response to male circling, song, or tapping—reflected in low but non-zero filter weights (Figure 5). In State 4, interpreted as a grooming state, her motion appears largely self-generated and decoupled from male behavior, with filters near zero regardless of male cue. In contrast, State 5 shows strong sensorimotor coupling: although the female remains stationary, she is poised to respond to male cues and capable of producing large responses, as shown by its high, input-dependent filter weights (Figure 5). To test whether individual states were sufficient to explain behavior and whether state differences reflected more than just velocity scaling, we generated female velocities using either the full GLM–HMM model or only the filters from one of the five states (Figure 6a). We found that States 1, 2, and 5 could generate a wide range of responses depending on the input, whereas States 3 and 4 failed to do so.

4 Conclusion

The fruit fly courtship system offers several advantages for developing machine learning approaches to animal communication: precisely quantifiable behaviors at subsecond resolution, multimodal signals that can be systematically measured, and sufficient data volume for fitting expressive probabilistic models. These features make it an ideal testbed for methods that may eventually scale to more complex communication systems in other species. In this work, we analyze the behavioral responses of the female fruit fly in response to the partner social cues in a task-free, naturalistic experimental setup. While prior studies have successfully characterized the sensory-driven latent structure of male song behavior [3, 8], the corresponding dynamics on the female side have remained unexplored. To advance toward a comprehensive understanding of the bidirectional structure of this social interaction, our work takes the critical next step by modeling female behavior through the similar lens of latent state inference. This established female model will serve as a prerequisite for future work aimed at jointly modeling both animals' internal states and determining their mutual influence within the closed-loop social system.

Overall, our approach exemplifies how statistical and machine learning methods can help reveal hidden communicative structures in animal behavior—a key challenge in applying artificial intelligence to non-human communication systems, where ground truth labels for internal states or communicative intent are unavailable.

Acknowledgments and Disclosure of Funding

We thank Shruthi Ravindranath, Talmo Pereira, and Junyu Li for sharing their experimental data. We thank the members of the Murthy and Pillow labs for their advice and support on the project. This project is supported by a Simons Collaboration on the Global Brain (SCGB) award to MM and JP, and the NIH Targeted BRAIN Circuits Projects grant to MM (R01NS110060). AL was supported in part by the NSF through the Center for the Physics of Biological Function (PHY-1734030). The authors declare that they have no competing interests.

References

[1] Herman T Spieth. Courtship behavior in drosophila. *Annual review of entomology*, 19(1): 385–405, 1974.

[2] Christelle Lasbleiz, Jean-François Ferveur, and Claude Everaerts. Courtship behaviour of drosophila melanogaster revisited. *Animal Behaviour*, 72(5):1001–1012, 2006.

[3] Philip Coen, Jan Clemens, Andrew J. Weinstein, Diego A. Pacheco, Yi Deng, and Mala Murthy. Dynamic sensory cues shape song structure in Drosophila. *Nature*, 507(7491): 233–237, March 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13131. URL `http://www.nature.com/articles/nature13131`.

[4] Kaiyu Wang, Fei Wang, Nora Forknall, Tansy Yang, Christopher Patrick, Ruchi Parekh, and Barry J. Dickson. Neural circuit mechanisms of sexual receptivity in Drosophila females. *Nature*, 589(7843):577–581, January 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2972-7. URL `https://www.nature.com/articles/s41586-020-2972-7`.

[5] Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D'Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods*, 19(4):486–495, April 2022. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-022-01426-1. URL `https://www.nature.com/articles/s41592-022-01426-1`.

[6] Jan Clemens, Cyrille C. Girardin, Philip Coen, Xiao-Juan Guan, Barry J. Dickson, and Mala Murthy. Connecting Neural Codes with Behavior in the Auditory System of Drosophila. *Neuron*, 87(6):1332–1343, September 2015. ISSN 08966273. doi: 10.1016/j.neuron.2015.08.014. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627315007084`.

[7] Rich Pang, Christa A. Baker, Mala Murthy, and Jonathan Pillow. Inferring neural population codes for *Drosophila* acoustic communication. *Proceedings of the National Academy of Sciences*, 122(21):e2417733122, May 2025. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2417733122. URL `https://pnas.org/doi/10.1073/pnas.2417733122`.

[8] Adam J. Calhoun, Jonathan W. Pillow, and Mala Murthy. Unsupervised identification of the internal states that shape natural behavior. *Nature Neuroscience*, 22(12):2040–2049, December 2019. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-019-0533-x. URL `http://www.nature.com/articles/s41593-019-0533-x`.

[9] janclemenslab/glm_utils. glm_utils. `https://github.com/janclemenslab/glm_utils`. Accessed: July 2025.

[10] Sean Escola. Hidden Markov Models for the Stimulus-Response Relationships of Multistate Neural Systems. 2011.

[11] Zoe C. Ashwood, Nicholas A. Roy, Iris R. Stone, The International Brain Laboratory, Anne E. Urai, Anne K. Churchland, Alexandre Pouget, and Jonathan W. Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2):201–212, February 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-01007-z. URL `https://www.nature.com/articles/s41593-021-01007-z`.

[12] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[13] Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. *Advances in neural information processing systems*, 7, 1994.

[14] Scott W. Linderman, Peter Chang, Giles Harper-Donnelly, Aleyna Kara, Xinglong Li, Gerardo Duran-Martin, and Kevin Murphy. Dynamax: A python package for probabilistic state space modeling with jax. *Journal of Open Source Software*, 10(108):7069, 2025. doi: 10.21105/joss.07069. URL `https://doi.org/10.21105/joss.07069`.

A Dataset

We confirm that our research complies with all relevant ethical regulations.

Behavioral chambers. We analyzed behavioral data from 75 male–female pairs of *Drosophila melanogaster* engaged in natural courtship. Behavioral experiments followed the protocol described in the SLEAP software [5]. In brief, flies interacted in custom-fabricated behavioral chambers with a 30mm × 30mm 3D-printed base (Formlabs Form 2, Black V3) and a clear PETG vacuum-molded dome (WidgetWorks Unlimited). Overhead video was captured using a Blackfly S 13YM3-M USB3 camera (FLIR) equipped with an MVL35M23 35mm FL C-mount lens (Thorlabs) and a 25-mm premium 850nm longpass filter (Thorlabs FELH0850). Illumination was provided by 850nm infrared LED strips positioned for side lighting. The arena floor included nine embedded microphone inlets arranged in a 3×3 grid beneath a fine 3D-printed mesh, allowing simultaneous acoustic and behavioral recording. Data acquisition was handled by custom-built workstations with Intel i7-8700K CPUs, 64GB RAM, 4TB Samsung 860 Evo SSDs, and EVGA GeForce GTX 1080 Ti (11GB) GPUs. Videos were recorded from above at 150 frames per second (fps) with a 5ms exposure time and a frame size of 1024×1024 pixels (1 channel), yielding a spatial resolution of 30.3 pixels/mm. Real-time image compression was performed using the Motif recording system and API (Loopbio GmbH), with GPU-accelerated H.264 encoding via the libx264 library (superfast preset). This setup produced nearly lossless videos with independently seekable frames.

Flies. All behavioral experiments were conducted using virgin male and female *Drosophila melanogaster* (wild-type strain NM91), aged 3–5 days post-eclosion, following the protocol in Coen et al., 2014 [3]. Fly bottles were kept at 25°C and 60% relative humidity. Experiments were initiated within two hours of incubator lights turning on. Males were single-housed, while females were group-housed prior to experiments. To prevent ceiling walking, the plastic dome of the behavioral chamber was coated with Sigmacote (SL2, Sigma-Aldrich) and allowed to dry under a fume hood for at least 30 minutes before use. Flies were gently introduced into the behavioral chamber using a custom-made aspirator. Recordings were terminated upon copulation or after 30 minutes, whichever occurred first. In total, we recorded 75 NM91 male–female pairs, yielding approximately 22 hours of courtship behavior.

Fly pose estimation and tracking via SLEAP. Fly poses were automatically tracked and manually proofread in all videos using SLEAP. We used the pre-trained 'flies13' model published in Pereira et al., 2022 [5], which defines a 13-node skeleton capturing prominent anatomical landmarks: head, thorax, abdomen, left and right wings (wingL, wingR), forelegs (forelegL4, forelegR4), midlegs (midlegL4, midlegR4), hindlegs (hindlegL4, hindlegR4), and eyes (eyeL, eyeR). The skeleton includes 12 edges connecting: thorax to head; thorax to abdomen; each wing to thorax; each leg to thorax; and head to each eye. Male and female identities were tracked using SLEAP's flow-shift-based identity tracking, followed by manual proofreading and correction of identity switches using the SLEAP GUI. Final joint coordinates and associated confidence scores were exported to .h5 files via the SLEAP API and used for all subsequent analyses.

Song segmentation. Audio was segmented into courtship song using previously described methods [8, 3], with an added pose-based filter to reduce false-positive sine detections. For each audio recording, the segmentation algorithm provided the onset and offset times of pulse bouts and sine trains, as well as the center of each detected pulse. Due to acoustic limitations in the behavioral setup used here, sine detection was prone to noise. Specifically, we retained only those sine bouts where at least one of the male's wing angles exceeded a threshold of $3\mathrm{deg}$ during the bout.

B Designing sensory inputs

To model female locomotion and wing flicking behavior, we transformed the tracked fly trajectories into a set of f behavioral feedback cues, which served as inputs to the GLM–HMM. For each cue, we extracted a 3 s window of history preceding the current time bin, sampled at 150 Hz, resulting in 450 time points per cue. These temporal windows were projected onto a set of four raised cosine basis functions, yielding four filter coefficients per cue. This produced a $4\times f$-dimensional feature vector (f

cues × 4 basis functions). We appended a constant term to this vector to model an intercept, resulting in a final $4 \times f+1$-D input vector per time bin.

For each fly, we constructed a design matrix of size $T \times (4 \times f+1)$, where T is the number of time bins after discarding the first 3 s (used for constructing the temporal history). Design matrices from all flies were concatenated to form a population-level dataset, enabling us to fit a single GLM–HMM model across animals.

We used a raised cosine basis set to capture temporal structure in the feedback cues. These basis functions are approximately orthonormal and spaced to provide smooth, overlapping temporal filters over the 3-second window, allowing the model to learn coarse-to-fine temporal dependencies while reducing dimensionality. We used the glm_utils [9] library to construct raised cosine basis functions and transform or inverse-transform the design matrices.

Output-specific feedback cues. We fit a single GLM–HMM model with a shared set of latent states across all outputs. However, the set of input features used to predict each behavioral output—forward velocity (fFV), lateral velocity (fLV), angular velocity (fAV) and wing flicking—was distinct, reflecting the sensory cues most relevant to that dimension of behavior. In particular, the input sets for fLV and fAV included interaction terms (e.g., cue × side) to capture directional effects that are not applicable to fFV or wing flicking behavior. This was implemented by applying an output-specific mask during the M-step in the EM algorithm, such that only the designated subset of input features contributed to the GLM weights for each output.

Inputs used to predict the female's forward velocity (fFV) included f=7 unsigned cues (Figure 1): male forward velocity (mFV), male–female distance (mfDist), male lateral speed (mLS), female heading to male thorax angle (fmAng), binary pulse and sine song, and male tapping. Each of these cues was temporally smoothed using four raised cosine basis functions, resulting in 4 coefficients per cue. This produced a feature vector of size 4×7+1=29, where the final element is a bias offset term.

For lateral (fLV) and angular velocity (fAV), the same base cues were used, but with directional (signed) information introduced through cue × side interaction terms. These allowed the model to capture the effects of asymmetric stimuli—such as song played to the left or right of the female—on her turning and lateral movement. The full set of signed cues included: mFV × side, mLS × side, mfDist × side, wingArisAng × side, pulse × side, sine × side, and tap × side. This resulted in a feature vector of size 4×7+1=29 each. The sine component, $\sin(\text{fmAng})$, corresponds to the "side" variable here (as shown in Figure 1 and Figure 5), indicating whether the male is to the left or right of the female.

To predict wing flicking behavior, f=8 unsigned cues were used: mFV, mLS, mfDist, fmAng, wingArisAng, male tapping, and binary pulse and sine song. This resulted in a feature vector of size 4×8+1=33.

Encoding male position relative to female. To represent the relative angular position of the male with respect to the female, we used both the cosine and sine of the angle "fmAng" between their orientations (Figure 1). This circular encoding captures the full 360° directional relationship in a smooth and continuous way, avoiding discontinuities at the angle wraparound (e.g., near $0°/180°$).

Z-scoring sensory inputs. All input features were z-scored independently for each fly to ensure comparability across individuals and to standardize the scales of different cues. For binary features such as pulse and sine song, we applied "safe" z-scoring: if a feature had near-zero variance (std < 1e-2, e.g., present in only a few frames), it was set to zero entirely to avoid instability during model fitting. Otherwise, the feature was z-scored as usual.

Z-scoring female behavioral outputs. All behavioral output variables—forward velocity (fFV), lateral velocity (fLV), and angular velocity (fAV)—were standardized independently for each fly by subtracting the mean and dividing by the standard deviation. Binary outputs, such as wing flicking, were left untransformed (0 or 1).

Smoothing. All male sensory input variable timeseries as well as female behavioral outputs were smoothed using a causal half-Gaussian kernel ($\sigma = 3$ frames (20 ms)), truncated at 4σ (12 frames, 80ms). Female outputs were then downsampled to 30 Hz by averaging within non-overlapping 33 ms windows (5 frames at 150 Hz). This preprocessing reduced high-frequency noise while preserving fast behavioral dynamics relevant to model prediction.

C Modeling

Chance model

We constructed the Chance baseline model using data from the entire courtship dataset, regardless of state or sensory input. This model estimates the emission distribution as a Gaussian with mean and covariance computed from all behavioral observations pooled together, for each continuous emission. For the discrete output (wing flicking), the Chance model uses a Bernoulli distribution with probability equal to the fraction of wing flicks observed in the dataset (i.e., n/N where n is the number of time points during the courtship with wing flicking and N is the total number of time points during the entire courtship).

GLM-HMM

The relationship between internal state, sensory input, and behavioral output is effectively modeled by a Generalized Linear Model–Hidden Markov Model (GLM-HMM) [8, 10, 11], which captures discrete latent states corresponding to distinct mappings from sensory cues to behavior (Figure 2). In this framework, each latent state z_t defines a separate generalized linear model relating the male cue history s_t to the observed female behavior y_t, including forward, lateral, and angular velocities, as well as wing flicking. The emission distribution at time t is:

$$p(y_t \mid z_t, s_t) = \mathcal{N}(y_t \mid w_{z_t} s_t + b_{z_t}, \sigma^2_{z_t}) \qquad \text{for continuous outputs (velocities)} \tag{1}$$

$$p(y_t \mid z_t, s_t) = \sigma(w_{z_t} s_t + b_{z_t}) \qquad \text{for binary outputs (wing flicking)} \tag{2}$$

Inference of GLM-HMM parameters. When fitting the GLM–HMM, the goal is to estimate the parameters that govern both the latent state dynamics and the emission model of female behavior. These parameters include the initial state distribution $\pi \in \mathbb{R}^K$, the state transition matrix $A \in \mathbb{R}^{K \times K}$, and the set of emission weights $w_k^{(j)} \in \mathbb{R}^M$, biases $b_k^{(j)}$, and (for continuous emissions) covariances $\sigma_k^{(j)} \in \mathbb{R}$ for each latent state k. Here j indexes emission variables (e.g., forward, lateral, angular, wing flicking). We denote the full set of parameters as $\theta = \{\pi, A, w_k^{(j)}, b_k^{(j)}, \sigma_k^{2(j)}\}$.

These parameters were fit to the female behavioral data using maximum a posteriori (MAP) estimation, implemented via the Expectation-Maximization (EM) algorithm. The EM algorithm has previously been adapted to fit hidden Markov models with external inputs [10, 12, 8, 11, 13]. However, since several implementation details are application-specific, we include a full description of the procedure here for completeness. The EM algorithm seeks to maximize the log-posterior of the model parameters given the female behavior data Y and sensory input features S. The log-posterior is given by, up to an unknown constant:

$$\begin{aligned}
\log p(\theta \mid Y, S) &= \log p(Y \mid S, \theta) + \log p(\theta) \\
&= \log \sum_Z p(Y, Z \mid S, \theta) + \log p(\theta) && \text{(marginalization)} \\
&= \log \sum_{z_{1:T}} p(y_{1:T}, z_{1:T} \mid s_{1:T}, \theta) + \log p(\theta) && \text{(expanding)} \qquad (3)
\end{aligned}$$

where the sum is taken over all K^T possible latent state sequences $z_{1:T}$. The first term represents the log-likelihood or the log-posterior of the observed data under the model, and the second term is a prior on the parameters.

Priors. The prior distribution over the model parameters θ was assumed to factorize as follows:

$$
\begin{aligned}
p(\theta) &= p(\{w_k^{(j)}\}) \cdot p(A) \cdot p(\pi) \\
&= \left[\prod_{k=1}^{K} \prod_{j} \mathcal{N}(w_k^{(j)} \mid 0, \lambda_j^{-1}) \right] \left[\prod_{k=1}^{K} \text{Dirichlet}(A_k \mid \boldsymbol{\alpha}_k) \right] \text{Dirichlet}(\boldsymbol{\alpha}_\pi) \qquad (4)
\end{aligned}
$$

We placed a zero-mean Gaussian prior on each GLM weight vector $w_k^{(j)}$ where λ_j is the inverse variance and controls the strength of regularization for emission variable j. Larger values of λ_j have a shrinking effect on the fitted weights, biasing them toward zero. For the continuous emission variables (forward, lateral, and angular velocity), we set the regularization parameter $\lambda_j = 10^{-6}$. For the discrete wing flicking emission, which uses a Bernoulli emission model, we used a stronger prior with $\lambda_j = 1$.

The transition matrix and initial state distribution were each given Dirichlet priors with symmetric concentration parameters. For the transition matrix A, we used a structured Dirichlet prior over each row A_k that encourages self-transitions (i.e., persistence within states) where the concentration parameters $\boldsymbol{\alpha}_k \in \mathbb{R}^K$ were set as $\boldsymbol{\alpha}_k = \alpha \cdot \mathbf{1}_K + \kappa \cdot \mathbf{e}_k$. Here, $\alpha = 1.1$ is a weakly informative base concentration applied to all transitions, $\kappa = 100$ is a stickiness parameter that adds mass to the diagonal (self-transition) entry, $\mathbf{1}_K$ is a vector of ones, and $\mathbf{e}_k$ is a one-hot vector indicating the k-th state. This form biases the prior toward self-transitions while still allowing transitions to other states. For the initial state distribution $\boldsymbol{\alpha}_\pi = \alpha \cdot \mathbf{1}_K$, with $\alpha = 1.1$. This weakly informative prior encourages a broadly uniform initial state distribution while still allowing the model to learn the estimate of π from the data.

Fitting using Expectation-Maximization (EM) algorithm. We used the EM algorithm to maximize the log-posterior given in Eq. 3 with respect to the GLM-HMM parameters. As the sum involves an exponential number of terms—O(K^T) to be specific—we do not maximize this expression directly. Instead, the EM algorithm provides an efficient way to compute this term using a single forward and backward pass over the data. During the E-step of the EM algorithm, we compute the 'expected complete data log-likelihood' (ECLL), which is a lower bound on the right-hand side of Eq. 3. Then, during the 'maximization' or M-step of the algorithm, we maximize the ECLL with respect to the model parameters θ. It can be shown that this procedure has the effect of always improving the log-posterior in each step of the algorithm and converges to a local optimum of the log-likelihood [10, 12].

The 'complete data log-likelihood' (CLL) for a session is written as $\log P(Y, Z|S; \theta)$:

$$
\begin{aligned}
CLL(\theta) &= \log P(Y, Z|S; \theta) \\
&= \log P(y_{1:T}, z_{1:T}|s_{1:T}; \theta) \\
&= \log \left[P(z_1 \mid \pi) \prod_{t=2}^{T} P(z_t|z_{t-1}, A) \prod_{t=1}^{T} P(y_t|z_t, s_t, w_k, b_k, \Sigma_k) \right] \\
&= \log \pi_{z_1} + \sum_{t=2}^{T} \log A_{z_{t-1}, z_t} + \sum_{t=1}^{T} \log B_{z_t}(y_t, s_t)
\end{aligned}
$$

where $B_{z_t}(y_t, s_t)$ is the Gaussian and Bernouli distribution given by the emission model equation (Eq. 2).

The Expected-CLL or the ECLL for a session, where the expectation is with respect to the distribution over the latents $\sum_Z p(Z \mid Y, S; \theta_{old})$ computed during the E-step, can now be written as:

$$
\begin{aligned}
ECLL(\theta) &= \sum_Z P(Z \mid Y, S; \theta^{old})\, CLL \\
&= \sum_Z P(Z \mid Y, S; \theta^{old})\, \log P(Y, Z|S; \theta) \\
&= \sum_Z \log \pi_{z_1} P(Z \mid Y, S; \theta^{old}) + \sum_Z \sum_{t=2}^{T} \log A_{z_{t-1}, z_t} P(Z \mid Y, S; \theta^{old}) + \sum_Z \sum_{t=1}^{T} \log B_{z_t}(y_t, s_t) P(Z \mid Y, S; \theta^{old}) \\
&\vdots \\
&= \sum_{k=1}^{K} \log \pi_k \gamma_k(1) + \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{t=2}^{T} \log A_{jk} \xi_{j,k}(t) + \sum_{k=1}^{K} \sum_{t=1}^{T} \log B_k(y_t, s_t) \gamma_k(t) \qquad (5)
\end{aligned}
$$

Here, we denote $\gamma_k(t) = P(z_t = k|Y_{1:T}, s_{1:T}, \theta^{old})$ for the posterior state probability of being in state k at time point t, and $\xi_{j,k}(t) = P(z_{t-1} = j, z_t = k \mid Y, S; \theta^{old})$ is the joint posterior state distribution for two consecutive latents z_t and z_{t-1}. We compute these two posterior distributions γ and ξ in the E-step as below:

E-step

The E-step of the EM algorithm involves computing the posterior distribution $P(Z|Y, \theta_{old})$ over the hidden variables given the data and the current setting of the GLM-HMM parameters θ_{old} using the forward–backward algorithm. The forward–backward algorithm makes use of recursion and memoization to allow these posterior probabilities to be calculated efficiently, with the forward and backward passes of the algorithm each requiring just a single pass through the whole session.

The goal of the forward pass is to obtain, for each time point t within a session and each state k, the quantity $a_i(t) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_t = y_t, z_t = i \mid s_{1:t})$ of observing $Y = y_1, y_2, ..., y_t$ which represents the posterior probability of the female behavior data up until time t and the latent state at time t being state k. Assuming there are K total states, it can be recursively computed as:

$$
a_j(t+1) = \sum_{k=1}^{K} a_k(t)\, A_{jk}\, B_j(y_t, s_t)
$$

where $a_j(1) = \pi_j\, P(y_1|z_t = j, s_1)$ and $B_j(y_t, s_t) = P(y_t|z_t = j, s_t)$ is the usual Gaussian or Bernoulli GLM distribution.

During the backward pass, the goal is to calculate the posterior probability of the future behavior data given the latent state $b_j(t) = P(Y_{t+1} = y_{t+1}, ..., Y_T = y_T \mid z_t = j, s_{t+1:T})$, as follows:

$$
b_j(t) = \sum_{k=1}^{K} b_k(t+1)\, A_{jk}\, B_k(y_{t+1}, s_{t+1})
$$

where $b_j(T) = 1$.

From the $a_j(t)$ and $b_j(t)$ quantities obtained from the forward-backward algorithm, we can compute the posterior state distribution γ over the latent state at every time step (this uses data from the whole session):

$$
\begin{aligned}
\gamma_k(t) &= P(z_t = k \mid Y_{1:T}, s_{1:T}, \theta_{old}) \\
&= \frac{P(z_t = k, Y_{1:T} \mid s_{1:T}, \theta_{old})}{P(Y_{1:T} \mid s_{1:T}, \theta_{old})} \\
&= \frac{P(Y_{1:t}, z_t = k \mid s_{1:t}, \theta_{old}) \cdot P(Y_{t+1:T} \mid z_t = k, s_{t+1:T}, \theta_{old})}{P(Y_{1:T} \mid s_{1:T}, \theta_{old})} \\
&= \frac{a_k(t) \cdot b_k(t)}{\sum_{i=1}^{K} a_i(t) \cdot b_i(t)} \qquad (6)
\end{aligned}
$$

Similarly, we can obtain the joint posterior state distribution ξ for the consecutive latents:

$$\begin{aligned}\xi_{j,k}(t) &= P(z_{t-1} = j, z_t = k \mid Y, S; \theta_{old}) \\ &= \frac{a_j(t)\, A_{jk}\, b_j(t+1)\, B_k(y_{t+1}, s_{t+1})}{\sum_{i=1}^{K} a_i(t) \cdot b_i(t)}\end{aligned} \tag{7}$$

Haivng now computed γ and ξ, ECLL is now a simply a function of model parameters θ with every other term known (Eq. 5).

M-step

After running the forward–backward algorithm, we can compute the total ECLL by summing over the per-session ECLLs (Eq. 5) and adding the log-prior (Eq. 4). During the M-step, we maximize the ECLL with respect to the GLM-HMM parameters θ. This uses the smoothed state probabilities $\gamma_t(k)$ and $\xi_{j,k}(t)$ computed during the E-step (Eqs. 6 and 7). For the initial state distribution π, transition matrix A and GLM weights for continuous emissions, this results in closed-form updates. The initial state probability π_k is updated as:

$$\pi_k^{new} = \frac{\sum_{e=1}^{E} \gamma_1(k)}{E},$$

The updated transition probabilities A_{ij} are given by the mode of the posterior Dirichlet distribution:

$$A_{ij}^{new} = \frac{\alpha - 1 + \sum_{e=1}^{E} \sum_{t=2}^{T} \xi_{i,j}(t)}{\sum_{j'=1}^{K} (\alpha - 1 + \sum_{e=1}^{E} \sum_{t=2}^{T} \xi_{i,j'}(t))},$$

Because these GLMs contribute independently to the terms A, π and emission terms B, we can optimize the filters for each output dimension separately.

In case of continuous emissions (forward, lateral and angular velocity), each state-specific emission model assumes a Gaussian distribution over the output $y_t^{(j)} \in \mathbb{R}$ $(j = 1 \ldots 3)$ with mean linearly dependent on the input vector $s_t \in \mathbb{R}^M$:

$$y_t^{(j)} \mid z_t = k, s_t, \theta \sim \mathcal{N}(w_k^{(j)} \cdot s_t, \sigma_k^{2(j)}) \quad \text{(for forward, lateral and angular velocity emissions)}$$

To estimate the GLM weights $\{w_k^{(j)}, b_k^{(j)}, \sigma_k^{2(j)}\}$, we solve the weighted multivariate linear regression problem for each state k. For notational simplicity, we assume $w_k^{(j)}$ includes the bias term $b_k^{(j)}$, with s_t augmented by a constant 1. To estimate the weights for each state k in the GLM-HMM, we pooled sufficient statistics across all sessions; thus, the variables $\gamma_k(t)$, y_t, and s_t below represent data concatenated across sessions.

The state-specific linear weights $w_k^{(j)} \in \mathbb{R}^{M+1}$ and the emission covariances $\sigma_k^{2(j)} \in \mathbb{R}$ have closed-form solution for the updates given by:

$$w_k^{new(j)} = \left(\sum_{t=1}^{T} \gamma_t(k)\, y_t s_t^\top \right) \left(\sum_{t=1}^{T} \gamma_t(k)\, s_t s_t^\top + \lambda_j I \right)^{-1}$$

$$\sigma_k^{2new(j)} = \frac{1}{\sum_{t=1}^{T} \gamma_t(k)} \sum_{t=1}^{T} \gamma_t(k) \left(y_t^{(j)} - w_k^{(j)} \cdot s_t \right) \left(y_t^{(j)} - w_k^{(j)} \cdot s_t \right)^\top$$

For numerical stability, we added a small constant to the estimated covariance $\sigma_k^{2new(j)} \leftarrow \sigma_k^{2new(j)} + 10^{-8}$.

For binary outputs such as wing flicking ($y_t^{(j)} \in \{0, 1\}$), Bernoulli GLM weights have no such closed-form update.

$$P(y_t^{(j)} \mid z_t = k, s_t, \theta) = \sigma(w_k^{(j)} \cdot s_t) \quad \text{(for wing flicking)}$$

We use the Dynamax Python package [14] to minimize the negative ECLL for all emission models. For continuous emissions, closed-form updates are implemented in the package. For Bernoulli emissions, Dynamax performs gradient-based optimization using the Adam optimizer, which is implemented via the Optax library.

Initializing GLM-HMM weights. We first fit a single-state linear regression model (i.e., a GLM without latent states) to each behavioral output. The estimated weights were then used to initialize the emission parameters of the GLM-HMM, with small random noise added independently to each state's parameters to break symmetry and encourage state specialization.

D Assessing model performance

Normalized Test LogLikelihood. We assessed model performance by calculating the log-likelihood of data held-out from training. We held-out out entire sessions of courtship data for assessing test set performance. That is, when fitting the model, the ECLL in Eq.5 are modified to include only 80% of sessions (because we use five-fold cross-validation throughout this work); and the log-likelihood of the held-out 20% of sessions E' is calculated using the fit parameters θ and a single run of the forward pass on the held-out sessions. In particular, we assessed how well the model predicted the next output given knowledge of all the data up to the present moment. In practice, it can be computed as:

$$\text{LL}_{\text{forward}}(\text{model}) = \sum^{\text{Test set}} \log \sum_{k=1}^{K} a_k(T) \tag{8}$$

that is, the sum of the last column of the a matrix obtained after doing a single forward pass on a test session.

To report the log-likelihood in more interpretable units, we normalized by subtracting the log-likelihood under the Chance model (described above; Figure 2), as follows:

$$\text{LL}_{\text{norm}}(\text{model}) = \text{LL}_{\text{forward}}(\text{model}) - \text{LL}_{\text{forward}}(\text{chance})$$

The chance model was drawn from the full distribution of behavior across all courtship recordings (Figure 2). To express this in interpretable units, we report $\text{LL}_{\text{norm}}(\text{model})$ as bits per second, by dividing it by the total duration of courtship in seconds. The normalized log-likelihood of the forward model thus reports the improvement in predicting female behavior over the Chance model, based on knowledge of her history to better estimate the current state.

State Inference. Latent states were inferred using forward filtering in the GLM–HMM framework. For each time point t, we computed the predictive state distribution $\hat{\gamma}_t$ over states using all observations up to the previous time point $t - 1$:

$$\hat{\gamma}_k(t) = P(z_t = k \mid s_{1:t}, y_{1:t-1})$$

This procedure is applied after training the GLM–HMM, using the learned transition and emission parameters to decode state sequences on the training and held-out data. It can be computed and stored using intermediate values during the calculations of the matrix a and γ during a forward pass in the E-step. For visualization purposes (e.g., state sequences over time), we assigned each time point to the most probable state ($\arg\max_k \hat{\gamma}_t(k)$). However, for model predictions, we used the soft state probabilities to compute a weighted sum of outputs across all states.

Behavior prediction. Rather than relying on hard state assignments, the model prediction at each time point was taken as a weighted sum over the predictions from all latent states, with weights given by the predictive state probabilities (from the forward filtering algorithm). Formally, for a behavioral output y_t, our model prediction $\hat{y}_t$ is given by:

$$
\begin{aligned}
\hat{y}_t &= \sum_{k=1}^{K} p(z_t = k \mid s_{1:t}, y_{1:t-1}) \cdot \hat{y}_{t,k} \qquad\qquad \text{where} \quad \hat{y}_{t,k} = w_k \cdot s_t + b_k \\
&= \sum_{k=1}^{K} \hat{\gamma}_k(t) \cdot \hat{y}_{t,k} \qquad (9)
\end{aligned}
$$

where $\hat{y}_{t,k}$ is the GLM prediction from state k for one of the continuous velocity emission predictions. However, for wing flick predictions, we used hard state assignments (obtained from the forward filtering step and using the state with the maximum probability at each time step; importantly, we did not use the Viterbi algorithm for state inference) and included an additional sigmoid nonlinearity (assuming animal in state k at time t):

$$P(\hat{y}_t) = \sigma(w_k \cdot s_t + b_k)$$

Pearson correlation coefficient To evaluate model performance for the continuous velocity emissions, we computed the Pearson correlation score between the observed behavioral output y and the model's soft predictions $\hat{y}$ as defined above. We compute:

$$r = \frac{\sum_t (y_t - \bar{y})(\hat{y}_t - \bar{\hat{y}})}{\sqrt{\sum_t (y_t - \bar{y})^2} \sqrt{\sum_t (\hat{y}_t - \bar{\hat{y}})^2}}.$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the mean of the observed y_t and predicted $\hat{y}_t$ outputs, respectively. This correlation captures the linear relationship between predicted and observed signals while incorporating uncertainty in latent state identity. The correlation score was computed separately for each fly and for each behavioral output variable (forward, lateral, and angular velocity). In Figure 2c, we report the mean Pearson correlation score averaged across these three output dimensions.

Pearson correlation coefficient per state. To evaluate how well each latent state predicts continuous behavioral outputs, we computed a state-specific version of the Pearson correlation coefficient $\hat{r}^{(k)}$ using soft assignments (Figure 2g). Specifically, we computed a soft-assignment weighted Pearson correlation coefficient between the true behavioral output y_t and the state-specific prediction $\hat{y}_{t,k}$, for each state k. This method incorporates the posterior state probabilities $\hat{\gamma}_t(k)$ (predictive state distribution using data up to time point $t-1$) as weights. For each state k, the steps were as follows:

- Compute the weighted means:

$$\mu_y^{(k)} = \frac{\sum_t \hat{\gamma}_t(k) y_t}{\sum_t \hat{\gamma}_t(k)}, \quad \mu_{\hat{y}}^{(k)} = \frac{\sum_t \hat{\gamma}_t(k) \hat{y}_{t,k}}{\sum_t \hat{\gamma}_t(k)}$$

- Compute the weighted covariance:

$$\mathrm{Cov}^{(k)} = \frac{\sum_t \hat{\gamma}_t(k)(y_t - \mu_y^{(k)})(\hat{y}_t^{(k)} - \mu_{\hat{y}}^{(k)})}{\sum_t \hat{\gamma}_t(k)}$$

- Compute the weighted variances:

$$\mathrm{Var}_y^{(k)} = \frac{\sum_t \hat{\gamma}_t(k)(y_t - \mu_y^{(k)})^2}{\sum_t \hat{\gamma}_t(k)}, \quad \mathrm{Var}_{\hat{y}}^{(k)} = \frac{\sum_t \hat{\gamma}_t(k)(\hat{y}_t^{(k)} - \mu_{\hat{y}}^{(k)})^2}{\sum_t \hat{\gamma}_t(k)}$$

- Finally, compute the weighted Pearson correlation:

$$\hat{r}^{(k)} = \frac{\mathrm{Cov}^{(k)}}{\sqrt{\mathrm{Var}_y^{(k)} \cdot \mathrm{Var}_{\hat{y}}^{(k)} + \varepsilon}}$$

  where ε is a small constant added for numerical stability.

F1 score To evaluate model performance on wing flicking (Figure 2d), we computed the F1 score between the observed binary behavioral output y and the model's predicted output $\hat{y}$. The predicted output was thresholded at 0.5 to yield a binary classification, and F1 score was computed across all time points.

F1 score per state. To evaluate model performance on wing flicking in each state (Figure 2h), we used hard state assignments obtained by taking the most probable state at each time point $(\arg\max_k \hat{\gamma}_t(k))$. For each state, we then computed the standard F1 score between the observed and predicted binary outputs, using only the time points assigned to that state.

Cross-validation. To select the appropriate number of latent states, we performed cross-validation by splitting the dataset into training and test sets. For each candidate model (with a different number of states), we fit the model parameters on the training data and evaluated performance on held-out test data Figure 2b.

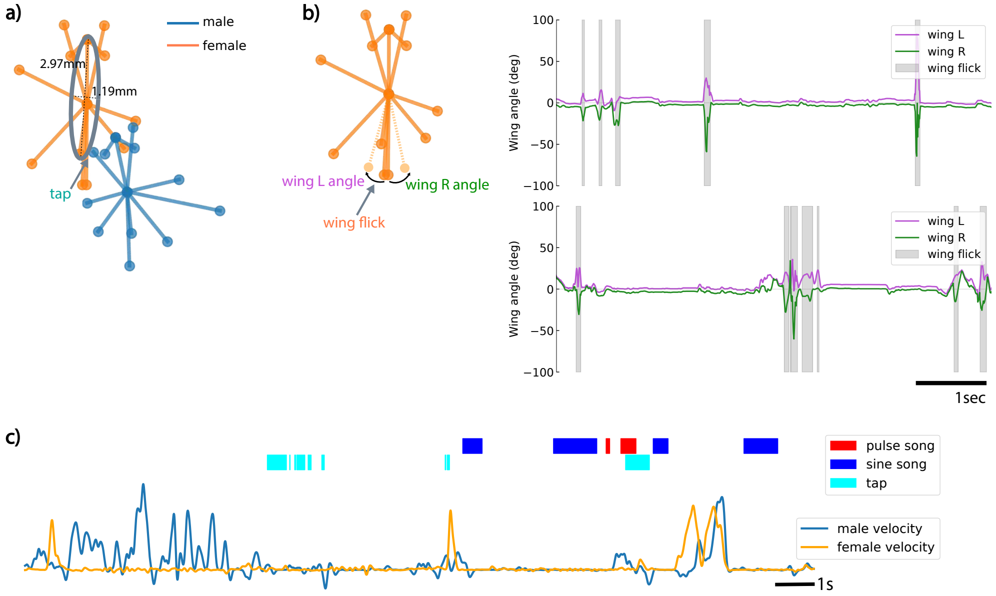

Figure 3: **Sensory features.** **(a)** Taps are identified using a heuristic based on proximity: a tap is registered when the male foreleg tip enters the ellipsoidal region containing the female from head to abdomen. Shown is an example frame where a male tap is detected. **(b)** Left: Female wing flicking is detected when the left and right wing angles deviate by more than 20° from each other. Wing angles are measured relative to the body axis using the tracked wing tip positions. Right: Two panels showing examples of 5-second segments of left wing (magenta) and right wing (green) angle traces from two different flies. Gray shaded bars indicate detected wing flicks. **(c)** A representative bout showing male and female forward velocity over time, overlaid with male song and tap. Female responses do not show consistent moment-to-moment coupling with male cues.

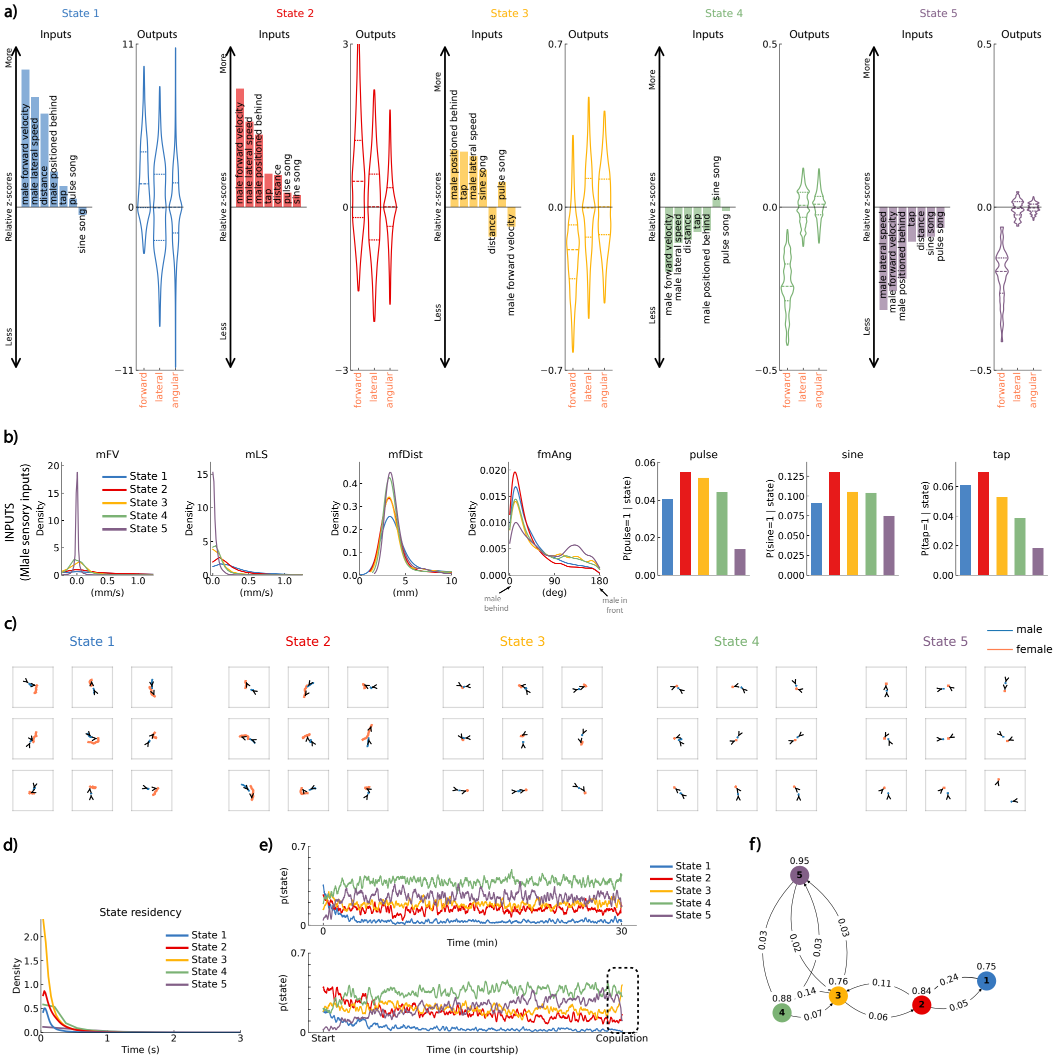

Figure 4: **Retrieved latent states uncover structure in female behavior across timescales. (a)** Left: For each state, sensory input features are ordered by their relative difference from the across-state mean. Right: Distributions of female forward, lateral, and angular velocities within each state. Note that the axis scales differ across states. **(b)** Distributions of values for some of the sensory cues in each state. Although a state may have features that are larger or smaller than average, the distributions are highly overlapping. **(c)** Representative traces of male and female movement trajectories in each state. Arrows indicate fly orientation at the end of 300ms. **(d)** The dwell times of the 5 states across all of the fly pairs. **(e)** The mean probability across flies of being in each state fluctuated only slightly over time when aligned to the absolute time (top) or the time of copulation (bottom). Data are from all 75 fly pairs. **(f)** Fitted state transition diagram representing the inferred dynamics of the 5-state GLM–HMM. Arrow labels indicate the probability of transitioning from one state to another.

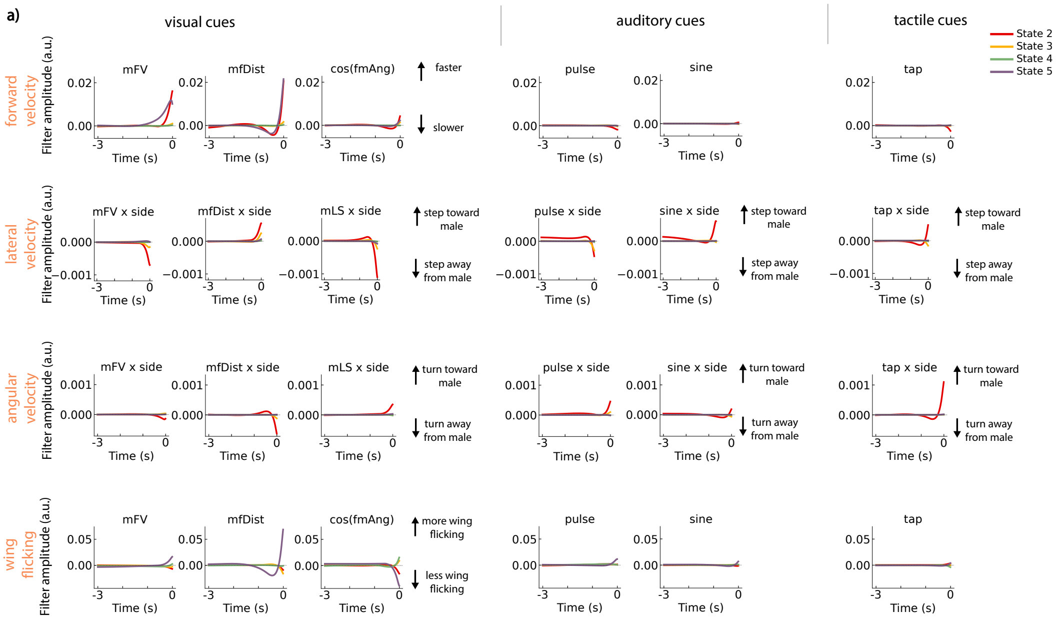

Figure 5: **State-dependent sensory filters reveal modality-specific drive of female locomotion.** **(a)** Output filters that predict female velocities for some of the feedback cues. For each latent state (State 2-5) inferred by the five-state GLM-HMM, we plot the GLM filter that converts a 3-second history of male cues into predicted female velocities. Columns are grouped by visual (left), auditory (center), and tactile (right) cues; rows show the effect on female forward velocity (top), lateral velocity (middle), angular velocity (middle), and wing flicking (bottom). States 2 (red) and 5 (purple) show strong filters, whereas the states 3 (yellow) and 4 (green) filters are nearly flat. Cue abbreviations are in Figure 1. Interaction terms (× *side*) in the lateral and angular filters capture cue laterality: for mFV, mfDist, mLS and tap cues, *side* indicates which side of the female the male's thorax occupies, whereas for pulse and sine song, it specifies whether the singing wing is on her left or right.

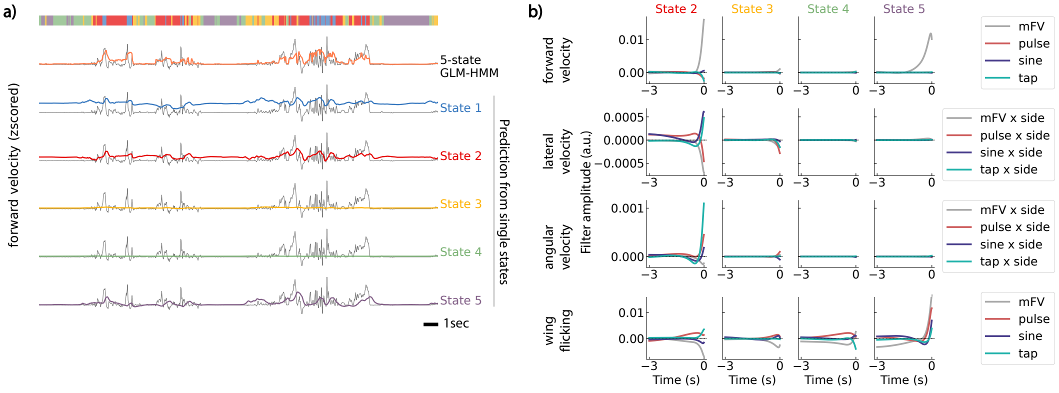

Figure 6: **Latent states are defined by distinct mappings between feedback cues and female locomotion behavior.** **(a)** A 30-second segment of female's forward velocity (fFV; zscored) from the natural courtship dataset, with the prediction of states indicated above in colored bar (first row). The prediction of the full GLM–HMM model (gray; second row) is very different from the prediction if we assume that the animal is always in one of the states 1-5. The output using the fFV filters from only that state is illustrated in the lower 5 rows. **(b)** Example output filters for each state revealed that even for the same feedback cues, the GLM–HMM shows distinct patterns of integration across states. Plotted here are the filters for male forward velocity (mFV), pulse song, sine song and the tap.

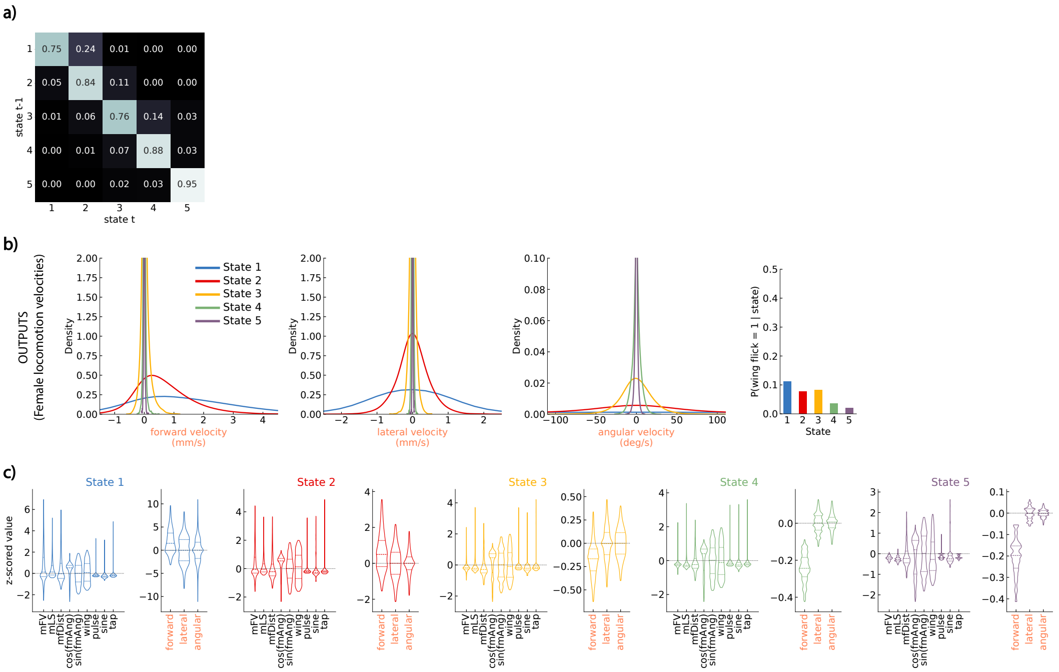

Figure 7: **State-wise distributions of behavioral outputs and sensory inputs.** **(a)** Inferred full transition matrix of the 5-state GLM–HMM, showing the probability of transitioning from each state (rows) to every other state (columns). **(b)** Distributions of each female behavioral output variable within each latent state. **(c)** Distributions of male sensory cues within each latent state, complementing Figure 4 where only the mean z-scored value per cue was shown. For the analysis in panel **(b)** and **(c)** here, we use hard-state assignments obtained using a forward-filtering step.