Structure and individuality of navigation in zebrafish larvae

Mattéo Dommanget-Kott, Jorge Fernandez-De-Cossio-Diaz, Monica Coraggioso, Volker Bormuth, Rémi Monasson, Georges Debrégeas, Simona Cocco

▶ **To cite this version:**

Mattéo Dommanget-Kott, Jorge Fernandez-De-Cossio-Diaz, Monica Coraggioso, Volker Bormuth, Rémi Monasson, et al.. Structure and individuality of navigation in zebrafish larvae. 2024. hal-04445557

HAL Id: hal-04445557
https://hal.science/hal-04445557

Preprint submitted on 8 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Structure and individuality of navigation in zebrafish larvae

Mattéo Dommanget-Kott,[1,2,*] Jorge Fernandez-de-Cossio-Diaz,[3,*] Monica Coraggioso,[1] Volker Bormuth,[1] Rémi Monasson,[3] Georges Debrégeas,[1,†] and Simona Cocco[3,‡]

[1] *Institut de Biologie Paris-Seine (IBPS), Laboratoire Jean Perrin, Sorbonne Université, CNRS, France*
[2] *Université Paris Cité, France*
[3] *Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 PSL Research, Sorbonne Université, Université de Paris, France*

(Dated: February 8, 2024)

Zebrafish larvae use stereotyped discrete swim bouts of various types to navigate their environment. Their temporal sequence displays a complex structure, whose characteristics are modulated by external factors, such as the water temperature. Here, we show that the use of Hidden Markov Models allows one to parse the exploratory kinematics of larval zebrafish in an agnostic fashion. Our approach thus provides a more robust method of bout classification than was previously proposed with standard Markov Models relying on ad hoc state identification hypothesis. We then unveil temporal persistence in the navigation at low water temperature that was previously overlooked or underestimated. We further show that the model is accurate enough to capture subtle differences in exploratory trajectories between individuals, and has thus potential application in behavioral phenotyping. The code used in this study is made available in a format specifically designed for educational purposes.

Keywords: zebrafish; Markov Chain; Hidden Markov Model; behavior

I. INTRODUCTION

Animal behavior unfolds as a structured sequence of stereotyped motor actions, much like language. Understanding behavior thus requires identifying the vocabulary, *i.e.* to categorize these elementary behavioral units, and to characterize the corresponding grammar, *i.e.* their relative organization [1]. As an illustration, navigation in Zebrafish larvae (see [2–4] for review) consists of a series of discrete swimming events of ~ 100 ms duration, called bouts, separated by $\sim$1-2 second-long dwelling periods. Due to this inherent discretization, the navigation behavior appears particularly well suited to modeling in terms of Markovian dynamical processes. However, to implement this approach effectively, reliable segmentation of consecutive bouts into different categories, or states, is essential.

So far, the categorization of bouts has been carried out independently of the examination of their temporal organization. In [5], unsupervised segmentation was performed through Principal Component Analysis (PCA) of tens of kinematic parameters extracted from the fish's tail and body motion. This approach yielded no less than 13 bout types, a number that the authors found sufficient to encompass the entire behavioral repertoire of the animal, including hunting, escape, social behavior, etc.

In other studies [6–12], the focus was put on the animal strategy of spontaneous exploration in spatially uniform environments or in the presence of sensory gradients (taxis). In this context, a crucial kinematic parameter was the animal orientational dynamics. Bouts were then categorized based on the value of their induced body reorientation, resulting in the labeling of forward and turning bouts (either rightward or leftward). The selection of these various bout types was found to depend on sensory cues, resulting in the animal's capacity to ascend light [6, 9] or temperature [11, 13, 14] gradients.

Although it offered a simple and interpretable description of the animal's explorational dynamics, the 3-state categorization approach in these studies was based on a partition of the bouts according to some threshold, *i.e.* bouts were labeled as turns if the reorientation angle was larger than some fixed value. This approach has two drawbacks. From a statistical point of view, it is prone to biases, in particular when comparing behavioral data obtained in different contexts, such as temperature, luminosity, hunger state, etc, which may systematically impact the way bouts of a given type are executed. From a conceptual point of view, it is unclear why a quantitative observation, such as the reorientation angle, should be unambiguously assigned to a unique behavioral state. Different internal states of the animal, likely related to some distinct neural counterparts, could transiently give rise to similar motor or behavioral correlates.

To understand how much these biases affect the current description of navigation in larval zebrafish, we hereafter re-analyze video recordings of freely swimming animals using more flexible and agnostic methods. We make use of Markovian-based state space models, more precisely the Hidden Markov Model (HMM), as an alternative and agnostic way of parsing exploratory trajectories. HMM have long been successfully applied in a variety of tasks and species, as they provide a robust method to discover underlying structures in temporal data in an unsupervised way [15–18]. Additionally, these models of-

* These two authors contributed equally
† Correspondence: georges.debregeas@sorbonne-universite.fr
‡ Correspondence: simona.cocco@phys.ens.fr

fer a probabilistic framework that can be used to score part of trajectories or even simulate synthetic behavior. To better assess the quality of the analysis, we consider recordings at different water temperatures. Because thermoregulation is critical for survival, and Zebrafish are ectothermic (*a.k.a.* cold-blooded) animals [19], they employ strategies to keep their body temperature within a physiological range (18-33°C). As was shown in previous work, the animal navigates to its optimal temperatures by adjusting its behavior based on the temperature it experiences in its environment.

We systematically compare the 3-state sequences of bouts and their temperature dependence, as derived from two methods: the first one uses threshold-based labeling (as in [11]) followed by Markov Chain modeling (MC); the second one relies on HMM to simultaneously label the bouts and infer their temporal organization. We find that HMM, by inferring a consistent bout labeling, allows one to reveal a more pronounced persistence of the bout type and bout orientation at low temperatures. Yet, this persistence in orientation is compatible with a Markovian description of the dynamics between (hidden) behavioral states, in contrast with results obtained with the ad-hoc thresholding approach.

We further leverage the scoring capability of HMMs to quantitatively assess how the trajectories change from a statistical point of view across time for the same animal, and how these temporal fluctuations compare to the inter-individual variability in the animals' navigation. Remarkably, the models corresponding to distinct animals remain sufficiently different across time to allow for automatic and reliable recognition of the animal identities from the observation of their trajectories.

Last of all, we discuss the implications of these results for our understanding of zebrafish navigation and its underlying neural processes.

II. RESULTS

This section is organized as follows. First, we briefly describe the dataset used in the present work. We then introduce two methods to model the trajectories: naive Markov Chains (MCs) inferred from manually classified data, and Hidden Markov Models (HMMs). The outcomes of the two methods are compared in terms of the markovianity of 3-states bout sequences, and their ability to reproduce the persistent properties of swimming exploration. Last of all, we evaluate the ability of HMM to perform behavioral phenotyping solely based on orientational statistics.

A. Data

The data used in the present paper comes from a previous publication that examined the kinematic of free exploration in zebrafish larvae [11]. The experimental design (Fig.1a) allowed us to record the trajectories of multiple freely swimming larvae aged 5-7 days. A set of kinematic parameters was extracted from the fish trajectories at each bout n, such as the angular change $\delta\theta_n$ in heading direction, as well as the dwelling time and the traveled distance (see Material and Methods sec. IV A).

The experiment was repeated in a range of controlled temperatures, specifically 18°C, 22°C, 26°C, 30°C, and 33°C (Fig.1b). The ambient temperature impacted systematically the statistics of trajectories, leading to qualitatively different behaviors as illustrated in Figure 1b. As the temperature increased, trajectories tended to become more winding and erratic.

B. Modeling with Markov Chains

Observation of the distribution of reorientation angles after each bout in Figure 1d suggests a description of the dynamics in terms of 3 states, labeling each swim bout into forward (F) or turn, either to the left (L) or to the right (R). In practice, this categorization is carried out by thresholding the distribution of re-orientation angles. Denoting the state of swim bout n by b_n we have:

$$b_n = \begin{cases} R, & \text{if } \delta\theta_n < -\delta\theta_0 \\ F, & \text{if } -\delta\theta_0 < \delta\theta_n < +\delta\theta_0 \\ L, & \text{if } \delta\theta_n > +\delta\theta_0 \end{cases} \quad (1)$$

The same threshold $\delta\theta_0 = \pm 10^\circ$ is applied for left and right turns. This choice relies on the hypothesis that zebrafish larvae, as a group, have no preferred direction (*a.k.a.*. non-handedness). As the exact value of $\delta\theta_0$ has minimal qualitative impact on the results of the Markov Chains, we adopt the same value as in [11]; notice that $\delta\theta_0$ is the same for all temperatures to avoid introducing ad hoc temperature-dependent biases. An example of the classification of states along a swimming trajectory is shown in Figure 2c.

Once the bout states are identified, we define a dynamical model for the trajectories $... \to b_{n-1} \to b_n \to b_{n+1} \to ...$ using a Markov Chain (MC). Informally, the sequence in states is described by the probabilistic automaton in Figure 2a. In this model, after each bout n, a new state b_{n+1} is drawn randomly conditioned only to b_n (and not to previous states). The transition probabilities between states, $P(b'|b) = P(b_n = b \to b_{n+1} = b')$, are estimated by counting the occurrences of the transitions $b \to b'$ along the trajectories:

$$P(b'|b) = \frac{\#_{b\to b'}}{\#_{b\to F} + \#_{b\to L} + \#_{b\to R}} \quad (2)$$

with $b, b' \in \{F, L, R\}$.

The top right eigenvector of the 3×3 transition matrix gives access to the stationary probabilities $P(b)$ of the 3 states. These probabilities are in excellent agreement with the frequencies of states (difference < 0.003 for all

bout types and temperatures) estimated through direct counting.

C. Modeling with Hidden Markov Models

We then turn to an agnostic categorization method, where states are inferred rather than *a priori* assigned. To do so, we consider a Hidden Markov Model (HMM) on 3 states, see Figure 2b. Contrary to MC, HMM makes a clear distinction between the observations (here the reorientation angles $\delta\theta_n$ treated as 'symbols') and the states of the system (here b_n, which are not directly accessible from the knowledge of $\delta\theta_n$, in contradistinction with the key assumption underlying MC). The HMM is defined by:

- The transition probabilities $P(b \to b')$ between the hidden states. We enforce the non-handedness by imposing that

$$\begin{aligned} P(F \to L) &= P(F \to R) \\ P(L \to L) &= P(R \to R) \\ P(L \to R) &= P(R \to L) \\ P(L \to F) &= P(R \to F) \end{aligned}$$

  This in turn ensures that steady state bout probability is left-right symmetric ($P(L) = P(R)$).

- The emission probabilities, $E(\delta\theta|b)$, relate the observations $\delta\theta$ to the hidden states b. For the forward state, we choose normally distributed reorientation angle emission distributions, centered in zero: $E(\delta\theta|F) = \mathcal{N}(\delta\theta; 0, \sigma)$. For turn states, we use Gamma distributed reorientation angles, with a positive or negative sign according to whether the state is Left or Right: $E(\delta\theta|L) = \Gamma(+\delta\theta; \alpha, \theta)$ and $E(\delta\theta|R) = \Gamma(-\delta\theta; \alpha, \theta)$, constraining $\alpha > 1$. Again, we ensured non-handedness by enforcing the same parameters for the left and right emission distribution. See Material and Methods sec. IV B for details about the validation of these emission distributions.

- A probability distribution for the initial state at the beginning of a trajectory.

We train HMM models for each dataset using the Baum-Welch algorithm, with a custom Julia implementation (available at `https://github.com/ZebrafishHMM2023/ZebrafishHMM2023.jl`).

D. State identification methods have a strong impact on captured behavioral persistence

As the Markov Chain inferred from thresholded data (MC, Fig.2a) and the Hidden Markov Model (HMM, Fig.2b) have the same internal behavioral states, we propose in this section a comparison of these models to investigate the impact of those different labeling methods.

As illustrated with an example trajectory at 22°C in Figure 2c, MC and HMM labeling are quite different. MC labeling can display alternations between Forwards and Turns when the bout reorientations are close to the threshold, while for the same sequence, the HMM tends to consistently label these bouts as Turns. These differences correspond to a reclassification of approximately 60% of Forward bouts into Turning bouts at 22°C (Fig.2d, Supplementary Fig.7a for all temperatures). With the HMM classification, we thus observe longer streaks of consecutive turns in the same direction, with characteristic turn-streak length $L_0^{\mathrm{HMM}} \approx 1.4$ bouts (while $L_0^{\mathrm{MC}} \approx 0.9$ bouts), with $P(L) \propto e^{-L/L_0}$ the probability of observing a streak of L consecutive bouts of the same type (Fig.2e). In contrast, we find no significant difference in characteristic forward-streak length between HMM and MC. Also, as temperature increases, we observe for both models that the characteristic streak length decreases (particularly for forward bouts), which is coherent with our previous understanding of zebrafish navigation, which tends to involve sharper turns (*i.e.*, reorienting themselves using fewer but more pronounced turning bouts) at higher temperatures (see Fig.1b).

Taken together, these results suggest that the Hidden Markov Model might be better at capturing the long-term persistence in reorientation while maintaining a coherent (and perhaps more accurate) bout classification. This is likely due to the model's ability to label bouts of small reorientation angles based on context, leading to a more stable classification where the threshold method would induce oscillations between turn and forward bouts.

To better assess the different impacts between those two labeling methods, we turn our attention to the inferred models.

As expected, we observe significant differences in the steady-state bout-type probability $P(b_n)$ with $b_n \in \{F, L, R\}$ between MC and HMM (Fig.3a). Indeed, HMM finds turning bouts to be significantly more frequent at lower temperatures than MC. While HMM finds very little temperature dependence in bout-type distribution, MC analysis leads to the appealing but potentially erroneous conclusion that the rate of turning bouts increases uniformly with temperature. This temperature-dependent effect is most likely due to the ad hoc hypothesis that the threshold $\delta\theta_0$ is independent of temperature, while the HMM seems to suggest that the width of the $\delta\theta_n$ distribution corresponding to forward bouts increases with temperature (see Supplementary Fig.8b,c).

In order to assess the persistence in bout type, we further compared the transition probability $P(b|b)$ and the unconditional probability $P(b)$ for both forward ($b = F$) and turning bouts ($b = T \in L, R$). Indeed, $P(b|b) = P(b)$

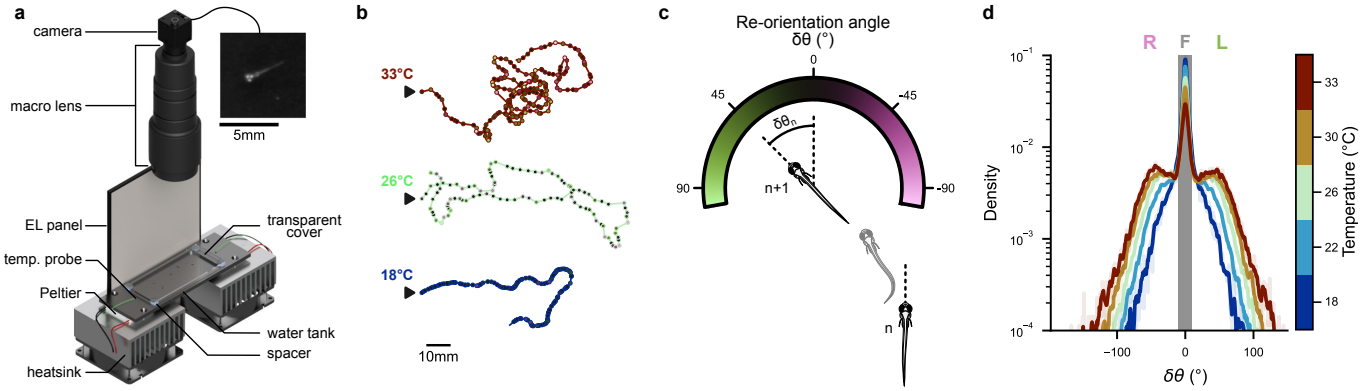

FIG. 1. **Experimental Setup and Behavioral Parameter**: **(a)** Overview of the experimental configuration: Zebrafish larvae navigate freely within a temperature-controlled tank while an imaging system records images at a rate of 25 frames per second. The top-right panel offers a close-up view of a larva in a raw image. Adapted from Le Goc *et al.* [11]. **(b)** Example zebrafish larvae trajectories in 2D space at various temperatures. Each point represents a swim bout, with the color indicating the corresponding re-orientation angle according to panel c. The trajectories' starting points are denoted by black arrows. **(c)** Description of the convention used for the reorientation angle ($\delta\theta_n$) between two consecutive swim bouts (n and $n+1$). **(d)** Distribution of re-orientation angles ($\delta\theta_n$) for each ambient temperature. The grayed-out area corresponds to the re-orientation angles classified as forward bouts by thresholds at $\pm 10°$.

would indicate an absence of persistence or memory in the bout-type selection process, as was previously reported [9] and as is observed with MC ($P_{\mathrm{MC}}(F|F) \approx P_{\mathrm{MC}}(F)$). In contrast, HMM-based analysis suggests significant bout-type persistence at lower temperatures (Fig.3b): we find that $P_{\mathrm{HMM}}(F|F) > P_{\mathrm{HMM}}(F)$ at 18°C and 22°C.

Similarly, memory in bout orientation is better captured by the HMM. Indeed looking at the transition matrix $P(b_n \to b_{n+1}) = P(b_{n+1}|b_n)$ (Fig.3c, Supplementary Fig. 8a) and compared to MC, HMM infers significantly higher $P(L \to L)$, lower $P(L \to F)$, and quasi-unchanged $P(L \to R)$ transition probabilities (and respectively for Right bouts), which enhances the persistence of Left (respectively Right) bouts at the expense of Forward bouts. In other words, the Markovian transitions become more asymmetrical specifically for direction-dependent transitions, leading to a stronger memory of orientation with the HMM than the MC inferred from thresholded data.

Interestingly, these memory effects in the orientation and bout-type vanish at higher temperatures, where the transition matrix becomes uniform (Supplementary Fig.8a), and all bouts become equiprobable ($P(F) \approx P(L) \approx P(R)$, Fig.3a). This suggests more erratic trajectories at higher temperatures, which is indeed in line with our observations (see Fig.1b).

Overall, we found that by using a non-supervised method to simultaneously label the data and infer a Markov Model, we unveiled memory effects in zebrafish reorientation statistics, which had been previously underestimated or overlooked due to ad hoc hypotheses with MC approaches.

E. Markovianity

Previous work on this or similar datasets have used a thresholding method to classify and then model reorientation statistics, but have required the use of 4-state Markov Chains to account for the long-term persistence in the data [9, 11]. Specifically, they used 2 independent Markov Chains, the first controlling forward-turn bout transitions, and the second controlling directional left-right bout transitions (see Supplementary Fig.9a for a diagram of this 4-state model). This was justified by the fact that 3-state models were found to be highly non-Markovian. In particular, a 3-state model cannot account for directional persistence after a forward bout, a mechanism that was nevertheless observed. Indeed, in a transition $T_1 \to F \to T_2$ with $T_1, T_2 \in \{L, R\}$, the memory of orientation T_1 is lost as soon as the animal executes a forward bout F, and thus the selection of T_2 is necessary unbiased (see Materials and Methods sec. IV D.

Given that our 3-state Hidden Markov Model (HMM) re-labels numerous Forward bouts as Turn bouts, we ask whether this improved classification might alleviate this non-Markovianity issue, such that the ad hoc 4-state model might no longer be needed. In this section, we thus propose a new test of Markovian violation specifically designed for our use case, that we apply to both the HMM and MC models.

We introduce the *stubbornness factor* f_q to empirically assess the tendency of larvae to maintain their orientation, even after a sequence of q intermediary forward bouts (Fig.4a, Materials and Methods sec. IV C):

$$f_q = \frac{P(T_1 \to F^q \to T_2 | T_1 = T_2)}{P(T_1 \to F^q \to T_2 | T_1 \neq T_2)} \tag{3}$$

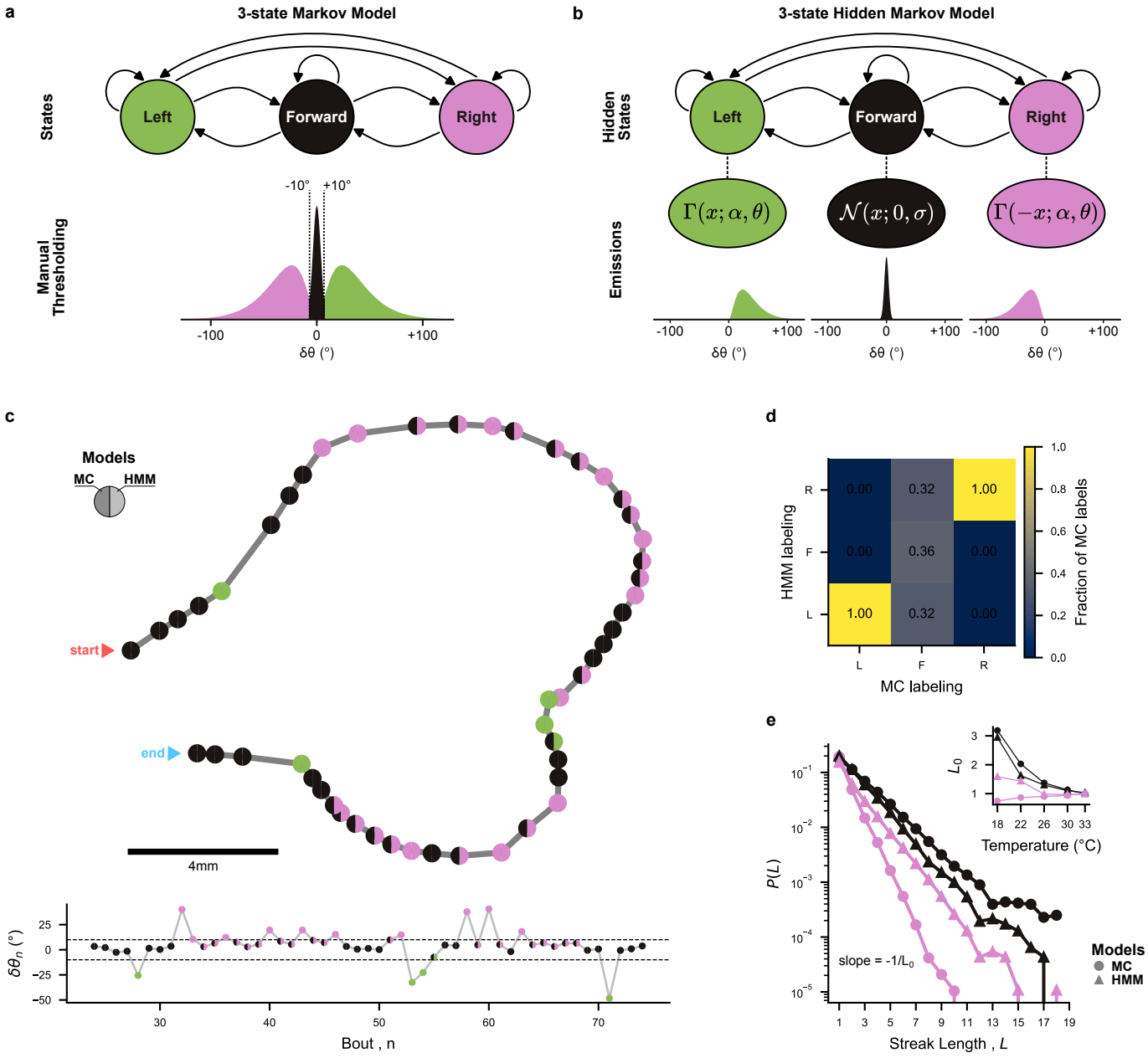

FIG. 2. **3-state Markov Chain and Hidden Markov Model, how behavioral labeling methods affect persistence**: **(a)** Diagram illustrating the 3-state Markov Chain (MC) where behavioral states Forward (F), Left (L), and Right (R) bouts are classified using a hard threshold at $\delta\theta_0 = \pm 10°$. **(b)** Diagram illustrating the 3-state Hidden Markov Model (HMM) with emissions modeled as a normal distribution for Forward bouts and gamma distributions for Turning bouts. **(c)** Differences in labeling between models MC and HMM for an example trajectory at 22°C. Each point represents a swim bout, with left color corresponding to the labeling according to the manual threshold used in MC, and right color corresponding to the labeling according to the HMM using the Viterbi algorithm. *Top:* trajectory in 2D space. *Bottom:* evolution of the reorientation angle $\delta\theta_n$ for this trajectory, with the dashed lines representing the threshold $\delta\theta_0 = \pm 10°$. **(d)** Confusion matrix between MC and HMM labeling, for all trajectories at 22°C (normalized with respect to the MC labeling). **(e)** Probability $P(L)$ of observing a streak of L consecutive forward bouts (black) or L consecutive turning bouts in the same direction (pink), for MC (circles) and HMM (triangles), measured from data at 22°C. *Inset:* Temperature dependence of the exponential decay characteristic length (L_0).

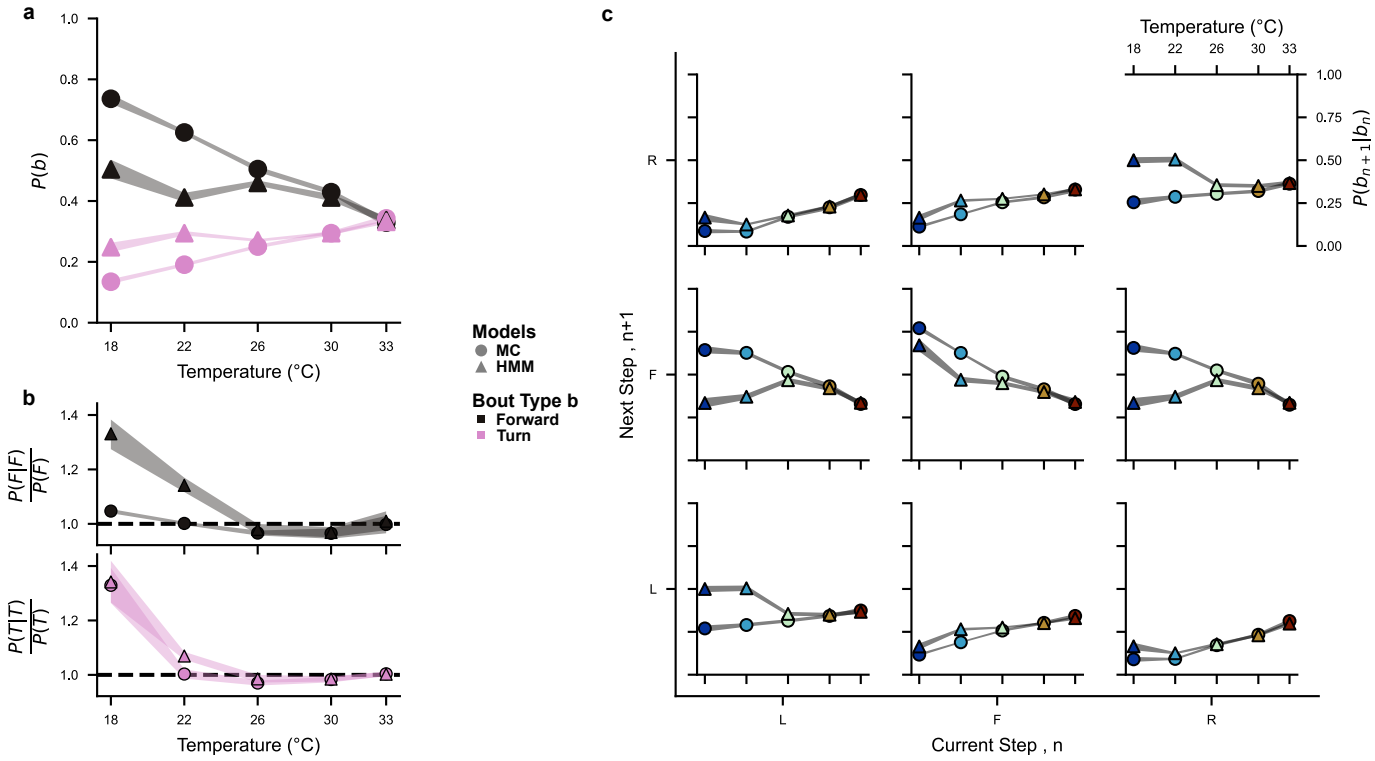

FIG. 3. **Memory effects emerge from better labeling**: **(a)** Temperature dependence of the steady state bout probabilities $P(b)$ for forward bouts ($b = F$, black) and turning bouts ($b \in \{L, R\}$, pink), and for both the Markov Chains inferred from thresholded reorientations (MC, circles) and Hidden Markov Models (HMM, triangles). **(b)** Temperature dependence of the ratio $\frac{P(b|b)}{P(b)}$, for forward bouts ($b = F$, black) and turning bouts ($b \in \{L, R\}$, pink). The dashed line indicates $P(b|b) = P(b)$, where the probability of state b_{n+1} is independent of state b_n (*i.e.* memoryless for $F \to F$ or $T \to T$ transitions). **(c)** Temperature dependence of the transition probabilities $P(b_{n+1}|b_n)$ between forward (F), left (L), and right (R) bouts, for both the Markov Chain (MC, circles) and Hidden Markov Model (HMM, triangles). **(a,b,c)** Throughout this figure, the width of the shaded curves represents the minimum-maximum of 100 cross-validations of both models, where for each cross-validation, model parameters were inferred from 50% of the data.

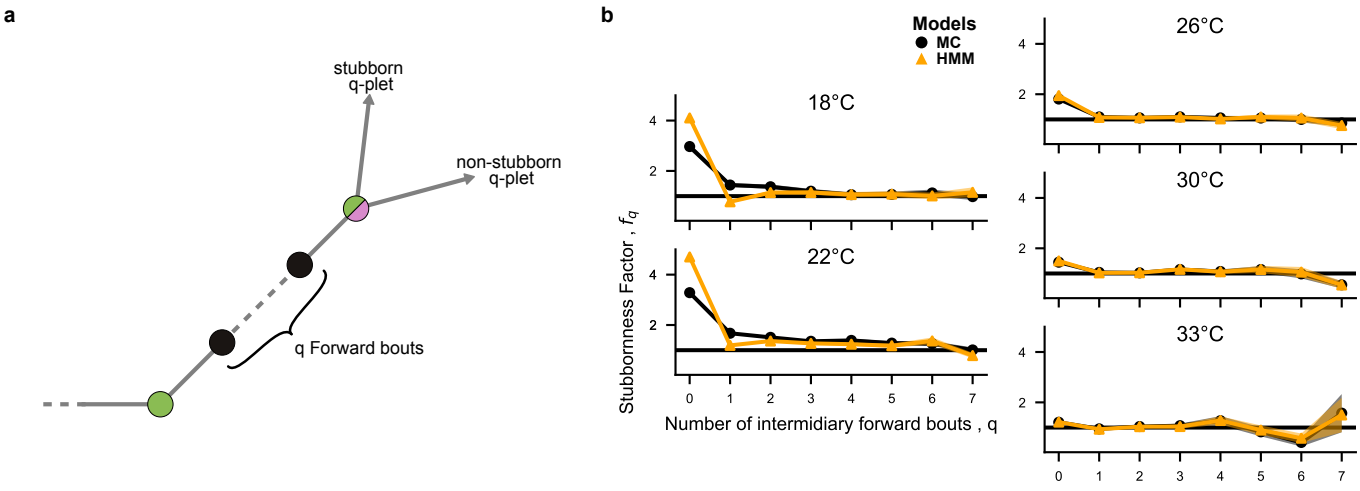

FIG. 4. **Markovianity**: **(a)** Diagram illustrating the definition of the stubbornness. For a q-plet of bouts $T_1 \to F \to \cdots \to F \to T_2$ with q intermediary forward bouts, a stubborn sequence is defined as one where directionality is conserved (*i.e.* $T_1 = T_2$), whilst a non-stubborn sequence will lose the memory of the initial turn (*i.e.* $T_1 \neq T_2$). **(b)** Evolution of the stubbornness factor f_q (see eq.3) with the number of intermediary forward bouts q, comparing the Markov Chain inferred from thresholded trajectories and the Hidden Markov Model (HMM) trained directly from reorientation angles, for each temperature.

with $F^q = \underbrace{F \to F \to \cdots \to F}_{q}$.

As mentioned above, due to the loss of directional memory after a forward bout, a non-handed 3-state Markovian model should have $f_q = 1$ for $q \geq 1$ (Materials and Methods sec. IV D). On the other hand, $f_{q=0}$ is a measurement of directional persistence during uninterrupted sequences of turning bouts.

We find that most of the memory effects captured by the HMM occur at $q = 0$, and that the stubbornness reaches $f_q \approx 1$ for $q \geq 1$, suggesting that the HMM can be considered as quasi-markovian at this temperature. In comparison, and for lower temperatures, the thresholded MC classification displays lower persistence at $q = 0$ but higher stubbornness at $q = 1$ (and less significantly at $q = 2$) (Fig.4b, Supplementary Fig.9b,c). This suggests that the thresholded labeling is indeed less Markovian due to alternations between turning and forward bouts during periods of constant reorientation, as anticipated in the previous section and illustrated on Figure 2c. As this stubbornness is mostly significant at $q = 1$, we expect that most mislabelings are one-off errors.

It is to be noted that the uncertainties presented on Figure 4b and Supplementary Figure 9c are conservative estimates, as there exits a bias inherent to the dataset. Indeed, a very stubborn fish will tend to stay longer within the Region Of Interest (ROI) of the camera, leading to longer trajectories and therefore weighing more on the final result. Hence, it is unclear whether a stubbornness factor $f_q = 1 \pm 0.2$ is truly significant (as suggested by the estimated error bars on Supplementary Fig.9c).

Overall, these results suggest that the non-markovianity of the data labeled via thresholding is mainly caused by the mislabeling of turning bouts as forward bouts during sequences of consecutive turns. The Hidden Markov Model seems to be a clear improvement, producing a labeling that is more Markovian. However, there remain some potential non-markovianities, which we have yet to explain.

F. Behavioral phenotyping from long trajectories

As the HMM captures the properties of trajectories over a population of fish, it is natural to ask whether the approach is accurate enough to characterize the behavior of single animals. We asked two questions:

- How significantly do the statistics describing the behavior of a single freely swimming fish vary over time?
- Are these fluctuations small enough that they allow for unambiguous identification of one fish from another?

To answer these questions, we considered additional experiments in [11], in which individual fish were tracked at 26°C. A total of 18 fish were recorded for over 2 hours. These long trajectories allowed us to assess whether the HMM can capture features that differentiate the variability among different fish from the variability shown by the same fish over time.

We first split the long trajectories of each individual fish into smaller sub-trajectories (chunks) of ≈ 12 minutes each, and trained an HMM on each of these smaller sequences (see diagram in Figure 5a). The parameters of these HMMs exhibit significant variability, compatible with the behavioral diversity of a single fish in time. We then also trained a single HMM on all trajectories of a single fish (the "global" HMM). Figure 5b compares selected parameters of the global HMM for each fish, against the average parameters over several HMMs trained on the chunk trajectories. The vertical error bars correspond to the variability over the different chunks for the same fish. Although a large variability is observed across several chunks for the same fish (evidenced by the large error bars), there is a clear trend between the global HMM and the average behavior of the chunk HMMs. Therefore, although a fish exhibits variability during a long sequence of bouts, the variability between distinct fish is larger.

These results suggest that the HMM models can be used to distinguish different fish from observations of their bout sequences. To test this hypothesis, we split the trajectories of each fish into a training and a withheld test set. After training the HMM on the train set for a particular fish, we computed the likelihood of all fish trajectories in the test set, and compared them. For 14 out of the 18 fish, the trajectory of maximum likelihood corresponds to a bout sequence executed by the fish used to train the HMM (Fig.5c), suggesting that the HMM encodes distinctive behavioral parameters that allow one to successfully discriminate between different fish. Due to the large variability exhibited by a single fish, this discriminative ability is better when large trajectories are available. To confirm this, in Figure 5d we again evaluated the likelihoods of subsampled subsets of the test fish trajectories, and recorded the number of times that the maximum likelihood HMM corresponded to the correct fish. Even when withholding 80% of the trajectories, we can correctly identify 10 out of the 18 fish. These results suggest that individual fish exhibit variable but distinctive behavior.

III. DISCUSSION

With the advancement of tracking methods, ethology has moved to a new era where it is now possible to study in great detail animal behavior in unconstrained naturalistic conditions [20–22]. Such experiments produce vast amounts of high-dimensional data, requiring automated yet robust and interpretable analysis methods. A critical task lies in the identification of behavioral motifs to map the behavior on a low-dimensional state space. However,

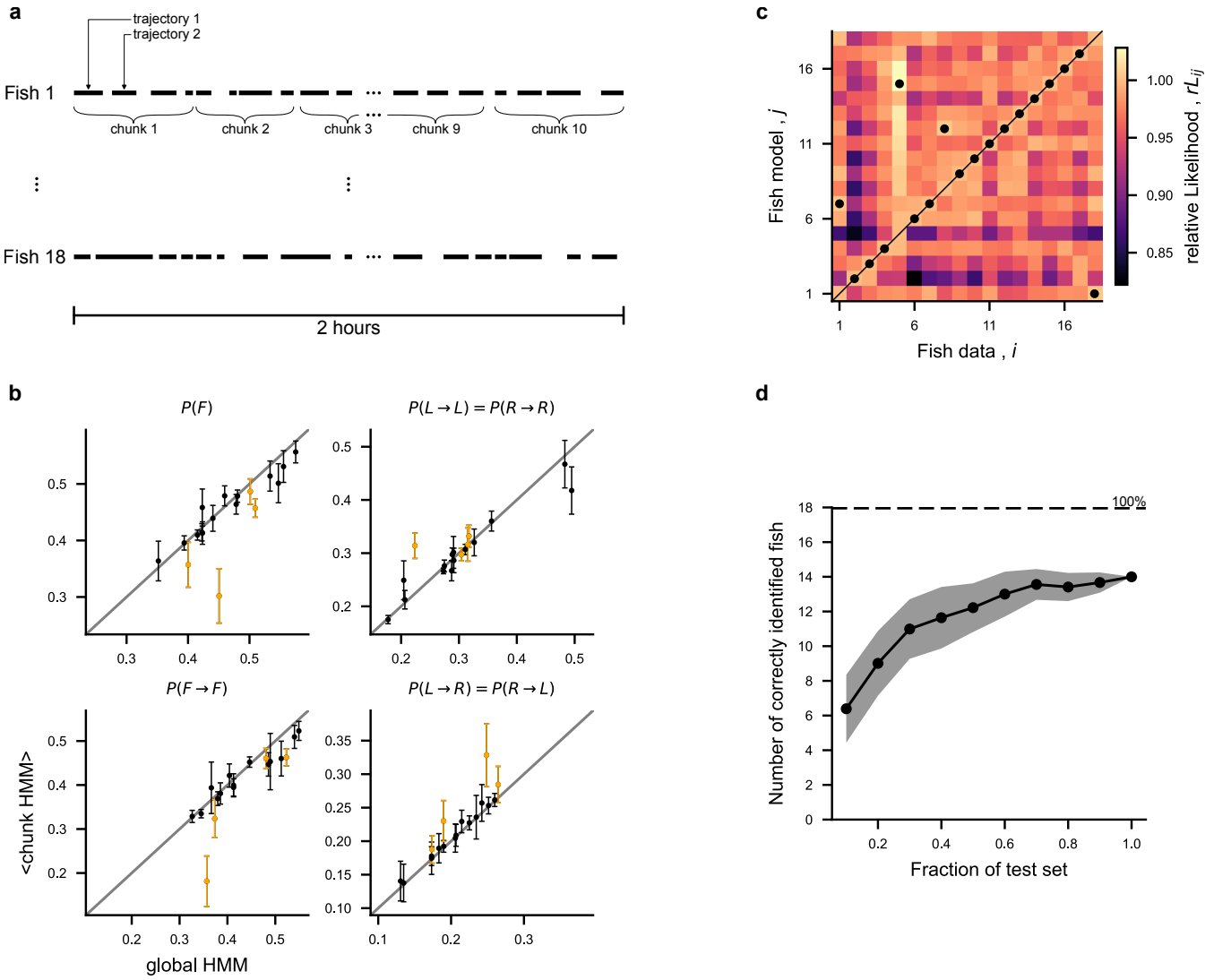

FIG. 5. **Fish identification from long trajectories**: **(a)** Diagram describing the dataset. Trajectories from 18 fish, recorded over 2-hour sessions, were each split into 10 chunks (mean $= 9.5 \pm 0.5$ trajectories per chunk). **(b)** HMM parameters inferred from all trajectories from one fish, compared with the average HMM parameters trained on chunks of that fish's trajectories. Only four HMM parameters are shown for clarity, namely, the steady state probability of forward turns $P(F)$, as well as the transition probabilities for forward-forward ($P(F \to F)$), turn-turn in the same direction ($P(T_1 \to T_2 | T_1 = T_2)$, and turn-turn in the opposite direction $P(T_1 \to T_2 | T_1 \neq T_2)$). Each dot represents a fish, and the error bars correspond to the standard error of the mean. Points labeled in orange correspond to fish misidentified in panel c. **(c)** Confusion matrix between data coming from fish i and HMM trained on data from fish j. The relative likelihood $rL_{i,j} = \frac{L(data_i|model_j)}{L(data_i|model_i)}$ is used to evaluate which fish identity is most likely according to each model (indicated with black dots for clarity). **(d)** Number of correctly identified fish determined from model likelihood when only a fraction f of the test data is used for identification. The shaded area indicates the standard deviation across 100 trials. In each trial, the data trajectories of each fish were randomly split into train and test sets (50%).

no definitive procedure exists for selecting the right number of states or for defining valid labeling criteria. This choice typically depends on available observables and involves a compromise between interpretability and accuracy of representation.

Even for simple behaviors such as the one presented in this article, parsing behavioral data in defined categories can be challenging. In our case, the difficulty arises from (i) the fact that swim bout kinematics are affected by the bath temperature, and (ii) the fact that the distributions of reorientation angles of distinct bout types overlap, in particular at low temperatures. Because they can accommodate such overlaps while taking into account the temporal regularities in the bout sequences, Hidden Markov

Models (HMMs) appear to be ideally suited for such a task.

Given the absence of a definitive ground truth, one might question the central assertion of this article – that Hidden Markov Models outperform standard threshold-based approaches. However, this claim is supported by the fact that the bouts re-labeled by HMM are not randomly placed, but are predominantly forward bouts within (left or right) turn streaks. The use of HMM over MC thus leads to the discovery of enhanced persistence in bout sequences through extended chaining of similar bouts.

The results presented in this article may have interesting implications for the understanding of the neuronal computation regulating navigation in zebrafish larvae. The neuronal circuit responsible for the leftward versus rightward bout selection has been identified in the anterior hindbrain [7]. The activity of this so-called Anterior Rhombencephalic Turning Region (ARTR) exhibits slow alternation between two subpopulations, located in the left and right hemispheres, controlling the orientation of swim bouts[7]. The period of this pseudo-oscillation is consistent with the orientational persistence time observed in the behavioral assay (on the order of 5-20s). Moreover, sensory stimuli such as unilateral visual stimulation and temperature changes can alter the dynamics of the ARTR in a manner that aligns with behavioral observations [11, 23]. The fact that a 3-state Markovian model adequately describes the sequence of bouts suggests that this same neuronal circuit could not only control the orientation of turn bouts but also control the selection of forward versus turning bouts. Recent analyses indicate the potential existence of three metastable states within this circuit, with left active, right active, and both inactive, which could thus correspond to the three bout types [24].

In the last section of this article, we demonstrate that the HMM exhibits sensitivity to natural inter-individual phenotypic variability. Inter- and intra-individual variability are ubiquitous traits of animal behavior [25, 26] and are necessary to ensure a trade-off between flexibility and adaptability to changing environmental demands and robustness in neural development [27]. Our model enables the identification of individual fish solely based on the dynamics of bout sequences. This ability could prove advantageous in the development of algorithms for tracking multiple moving animals. The state-of-the-art existing tools [28, 29] rely on image-based neural networks to identify unmarked individuals using natural variations in their physical and/or behavioral appearance to accomplish fast and reliable multi-individual tracking in a versatile range of different organisms or scenarios. Since our approach is based on gait phenotyping and is independent of image features, it is compatible with low-resolution videos (in which only the animal's position and orientation can be accessed) while still keeping versatility, reliability, and fast execution.

Finally, the improvement made by the following approach over previous studies is twofold. On one hand, not relying on rigid thresholds allows a more efficient description of how behavior changes in response to external perturbations in the environment, and, on the other hand, the approach opens up the possibility of accessing inter- and intra-individual variability.

In addition, to enhance the practical accessibility of Hidden Markov Model (HMM) formalism for analyzing behavioral sequences, we have developed a comprehensive and instructive Python tutorial (`https://github.com/EmeEmu/IBIO-Banyuls2023-Python`). This tutorial can be adapted for specific datasets or used as a resource for broader educational goals.

IV. MATERIALS AND METHODS

A. Dataset

The dataset used in the present study is derived from Le Goc *et al.* [11], and can be accessed directly at `https://doi.org/10.5061/dryad.3r2280ggw`. This dataset comprises spontaneous swimming trajectories of 5 to 7 dpf zebrafish larvae, collected at controlled bath temperatures of 18°C, 22°C, 26°C, 30°C, and 33°C. A camera was used to continuously record the swimming behavior of the fish within an arena of $100\times45\times4.5\text{mm}^3$ for 30 minutes at 25 frames/second. To eliminate border effects, a Region of Interest (ROI) was defined at a distance of 5mm from the arena's walls. Fish that swam outside the defined tracking ROI were considered lost, and a new trajectory was initiated upon their re-entry into the ROI. Therefore, the dataset contains a varying number of fish trajectories, ranging from 532 to 1513 trajectories across the different temperatures (mean = 1148). Individual trajectories were tracked offline using the open-source FastTrack software [30], and were then discretized into sequences of swimming bouts. Hence, each trajectory consists of a sequence of swim bouts, spanning from 9 to 748 bouts per trajectory (mean=60, distributions shown in Supplementary Fig.6a). From this extensive dataset, we exclusively utilized the re-orientation angles, defined as the difference between the heading direction at bout $n+1$ and the heading direction at bout n:

$$\delta\theta_n = \theta_{n+1} - \theta_n \tag{4}$$

(a graphical illustration of this definition can be found in Fig.1c). This parameter encapsulates the angular change between consecutive bouts, providing insight into the fish's ability to modify its orientation during swimming.

B. Emission of reorientation angles in the Hidden Markov Model

To validate the hypothesis that the re-orientation angles can be modeled using normal and gamma distributions, we compared the distribution of the data

with a Gaussian Mixture Model (GMM) and a Gaussian&Gamma Mixture Model:

$$p(\delta\theta) = w_F\mathcal{N}(\delta\theta; 0, \sigma) + w_L\Gamma(\delta\theta; \alpha, \theta) + w_R\Gamma(-\delta\theta; \alpha, \theta)$$

where $w_F+w_L+w_R = 1$, and w_F, w_L, and w_R denote the weights for forward, left, and right states, respectively.

Using Quantile-Quantile (QQ) plots, we show that this last mixture model accurately reproduces the observed distribution of $\delta\theta_n$ in the data, and is much better than a GMM, especially in the tails of the distributions (Supplementary Fig. 7c).

C. Stubbornness factor

The stubbornness factor f_q is a measurement of the animal's preference towards turning in the same direction over changing direction, after q intermediary forward bouts. It is defined as:

$$f_q = \frac{P(T_1 \to F^q \to T_2 | T_1 = T_2)}{P(T_1 \to F^q \to T_2 | T_1 \neq T_2)} \tag{5}$$

with $T_1, T_2 \in \{L, R\}$ and $F^q = \underbrace{F \to F \to \cdots \to F}_{\text{q}}$.

It can be computed from a sequence of classified bouts b_n by first identifying and counting the q-plets $T_1 \to F^q \to T_2$ where $T_1 = T_2$ and where $T_1 \neq T_2$:

$$\begin{cases} N_= & = \#(T_1 \to F^q \to T_2, T_1 = T_2) \\ N_\neq & = \#(T_1 \to F^q \to T_2, T_1 \neq T_2) \end{cases} \tag{6}$$

and then computing their ratio:

$$f_q = \frac{N_=}{N_\neq} \tag{7}$$

In practice, this ratio has a physical interpretation only for long sequences of bouts where $N_= >> 1$ and $N_\neq >> 1$. As the trajectories in our dataset can be quite short (Supp Fig. 6a), we compute f_q from all trajectories at a specific temperature, increasing the chance of observing a high number of stubborn ($N_=$) and non-stubborn ($N_\neq$) trajectories.

By considering that the probability of a given q-plet is stubborn follows a binomial distribution ($\mathbb{E}(N_=) = pN$ and $\mathbb{E}(N_\neq) = (1-p)N$ with $N = N_= + N_\neq$), we can evaluate the uncertainty in stubbornness as:

$$\Delta f_q = f_q \frac{1}{N_= + N_\neq}\left(\sqrt{\frac{N_=}{N_\neq}} + \sqrt{\frac{N_\neq}{N_=}}\right) \tag{8}$$

D. Stubbornness factor and 3-state Markov Chain

The stubbornness factor can be defined directly from the transition matrix.

For $q = 0$, calculations are simple:

$$f_{q=0} = \frac{P(L \to L) + P(R \to R)}{P(L \to R) + P(R \to L)} \tag{9}$$

For $q \geq 1$, the stubbornness factor is defined from the transition matrix as:

$$\begin{aligned} S_{L,q} &= P(L \to F^q \to L) \\ &= P(L)P(L \to F)P^q(F \to F)P(F \to L) \\ W_{L,q} &= P(L \to F^q \to R) \\ &= P(L)P(L \to F)P^q(F \to F)P(F \to R) \\ f_q &= \frac{S_{L,q} + S_{R,q}}{W_{L,q} + W_{R,q}} \end{aligned}$$

with $S_{L,q}$ the probability of a trajectory which starts and ends with a left bout, $W_{L,q}$ the probability of a trajectory which starts with a left bout and ends with a right bout, and $S_{R,q}$ $W_{R,q}$ their symmetrical opposites.

For a 3-state model, the forward-forward bout probability cancels out, giving:

$$f_q = \frac{P(L)P(L \to F)P(F \to L) + P(R)P(R \to F)P(F \to R)}{P(L)P(L \to F)P(F \to R) + P(R)P(R \to F)P(F \to L)}$$

and with our non-handedness hypothesis: $P(L) = P(R)$, $P(L \to F) = P(R \to F)$, and $P(F \to L) = P(F \to R)$, yielding:

$$f_q = 1 \quad \forall q > 0 \tag{10}$$

Acknowledgment. We acknowledge the following funding:

Author	Funder
M. DK.	École Doctorale Frontière de l'Innovation en Recherche et Education - Programme Bettencourt
J. FdCD.	Université PSL, AI Junior Fellow program
M. C.	European Union, Horizon 2020 Programme (H2020 MSCA ITN Project SmartNets GA-860949)
V. B.	European Research Council (ERC) under the European Union's Horizon 2020 research innovation program grant agreement number 715980
R. M., G. D., S. C.	Locomat ANR-21-CE16-0037

Data and Code availability. All the data and code used in the present article are available under GNU General Public License version 3 at `https://github.com/ZebrafishHMM2023/ZebrafishHMM2023_CodeAndData`.

The custom Julia implementation of Hidden Markov Model used is available under MIT License at `https://github.com/ZebrafishHMM2023/ZebrafishHMM2023.jl`).

The tutorial on using Hidden Markov Models for behavioral sequence analysis is available under GNU General Public License version 3 at `https://github.com/EmeEmu/IBIO-Banyuls2023-Python`. Originally, it was created for the i-Bio Summer School "Advanced Computational Analysis for Behavioral and Neurophysiological Recordings" held in Banyuls-sur-Mer in the summer of 2023.

[1] N. Tinbergen, *The study of instinct* (Pygmalion Press, an imprint of Plunkett Lake Press, 2020).

[2] M. B. Orger and G. G. de Polavieja, Zebrafish behavior: opportunities and challenges, Annual review of neuroscience **40**, 125 (2017).

[3] J. R. Meyers, Zebrafish: development of a vertebrate model organism, Current Protocols Essential Laboratory Techniques **16**, e19 (2018).

[4] J. H. Bollmann, The zebrafish visual system: from circuits to behavior, Annual review of vision science **5**, 269 (2019).

[5] J. C. Marques, S. Lackner, R. Félix, and M. B. Orger, Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering, Current Biology **28**, 181 (2018).

[6] X. Chen and F. Engert, Navigational strategies underlying phototaxis in larval zebrafish, Frontiers in Systems Neuroscience **8**, 10.3389/fnsys.2014.00039 (2014).

[7] T. W. Dunn, Y. Mu, S. Narayan, O. Randlett, E. A. Naumann, C.-T. Yang, A. F. Schier, J. Freeman, F. Engert, and M. B. Ahrens, Brain-wide mapping of neural activity controlling zebrafish exploratory locomotion, Elife **5**, e12741 (2016).

[8] E. J. Horstick, Y. Bayleyen, J. L. Sinclair, and H. A. Burgess, Search strategy is regulated by somatostatin signaling and deep brain photoreceptors in zebrafish, BMC biology **15**, 1 (2017).

[9] S. Karpenko, S. Wolf, J. Lafaye, G. Le Goc, T. Panier, V. Bormuth, R. Candelier, and G. Debrégeas, From behavior to circuit modeling of light-seeking navigation in zebrafish larvae, eLife **9**, e52882 (2020), publisher: eLife Sciences Publications, Ltd.

[10] E. J. Horstick, Y. Bayleyen, and H. A. Burgess, Molecular and cellular determinants of motor asymmetry in zebrafish, Nature Communications **11**, 1170 (2020).

[11] G. Le Goc, J. Lafaye, S. Karpenko, V. Bormuth, R. Candelier, and G. Debrégeas, Thermal modulation of Zebrafish exploratory statistics reveals constraints on individual behavioral variability, BMC Biology **19**, 208 (2021).

[12] D. L. Barabási, G. F. Schuhknecht, and F. Engert, Functional neuronal circuits emerge in the absence of developmental activity, Nature Communications **15**, 364 (2024).

[13] M. Haesemeyer, D. N. Robson, J. M. Li, A. F. Schier, and F. Engert, A brain-wide circuit model of heat-evoked swimming behavior in larval zebrafish, Neuron **98**, 817 (2018).

[14] L. S. E. Haesemeyer, Robson, A brain-wide circuit model of heat-evoked swimming behavior in larval zebrafish, Neuron **4**, 10.1016/j.neuron.2018.04.013 (2018).

[15] A. B. Wiltschko, M. J. Johnson, G. Iurilli, R. E. Peterson, J. M. Katon, S. L. Pashkovski, V. E. Abraira, R. P. Adams, and S. R. Datta, Mapping sub-second structure in mouse behavior, Neuron **88**, 1121 (2015).

[16] J. M. Mueller, P. Ravbar, J. H. Simpson, and J. M. Carlson, Drosophila melanogaster grooming possesses syntax with distinct rules at different temporal scales, PLoS computational biology **15**, e1007105 (2019).

[17] T. Gallagher, T. Bjorness, R. Greene, Y.-J. You, and L. Avery, The geometry of locomotive behavioral states in c. elegans, PloS one **8**, e59865 (2013).

[18] L. Tao, S. Ozarkar, J. M. Beck, and V. Bhandawat, Statistical structure of locomotion and its modulation by odors, Elife **8**, e41235 (2019).

[19] M. Haesemeyer, Thermoregulation in fish, Molecular and Cellular Endocrinology **518**, 110986 (2020).

[20] A. E. Brown and B. De Bivort, Ethology as a physical science, Nature Physics **14**, 653 (2018).

[21] T. D. Pereira, J. W. Shaevitz, and M. Murthy, Quantifying behavior to understand the brain, Nature neuroscience **23**, 1537 (2020).

[22] A. Kennedy, The what, how, and why of naturalistic behavior, Current opinion in neurobiology **74**, 102549 (2022).

[23] S. Wolf, A. M. Dubreuil, T. Bertoni, U. L. Böhm, V. Bormuth, R. Candelier, S. Karpenko, D. G. Hildebrand, I. H. Bianco, R. Monasson, *et al.*, Sensorimotor computation underlying phototaxis in zebrafish, Nature communications **8**, 651 (2017).

[24] S. Wolf, G. Le Goc, G. Debrégeas, S. Cocco, and R. Monasson, Emergence of time persistence in a data-driven neural network model, eLife **12**, e79541 (2023).

[25] K. Honegger and B. de Bivort, Stochasticity, individuality and behavior, Current Biology **28**, R8 (2018).

[26] A. K. Shaw, Causes and consequences of individual variation in animal movement, Movement ecology **8**, 12 (2020).

[27] P. R. Hiesinger and B. A. Hassan, The evolution of variability and robustness in neural development, Trends in Neurosciences **41**, 577 (2018).

[28] A. Pérez-Escudero, J. Vicente-Page, R. C. Hinz, S. Arganda, and G. G. De Polavieja, idtracker: tracking individuals in a group by automatic identification of unmarked animals, Nature methods **11**, 743 (2014).

[29] T. Walter and I. D. Couzin, Trex, a fast multi-animal tracking system with markerless identification, and 2d estimation of posture and visual fields, eLife **10**, e64000 (2021).

[30] B. Gallois and R. Candelier, Fasttrack: an open-source software for tracking varying numbers of deformable objects, PLoS computational biology **17**, e1008697 (2021).

SUPPLEMENTARIES

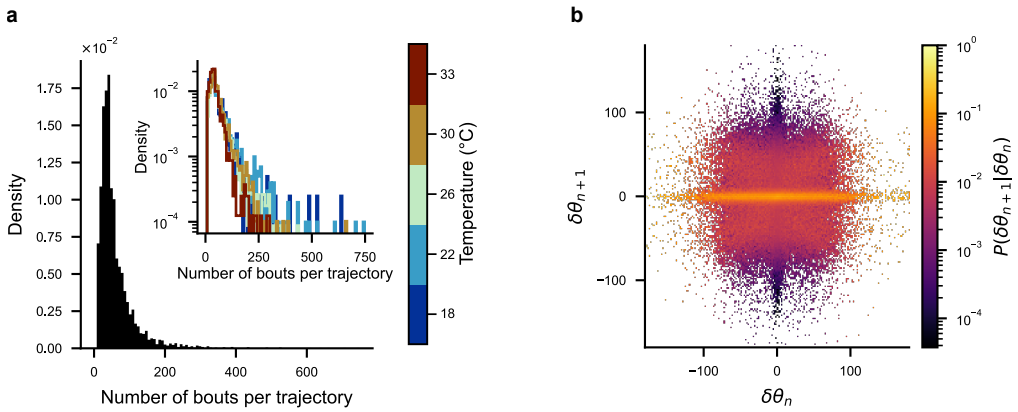

FIG. 6. **Supplementary panels to Fig.1**: **(a)** Distributions of the number of bouts per trajectory in the entire dataset (black), and for each recorded temperature (inset, colored). **(b)** Observed transition probabilities between reorientation angles for the entire dataset.

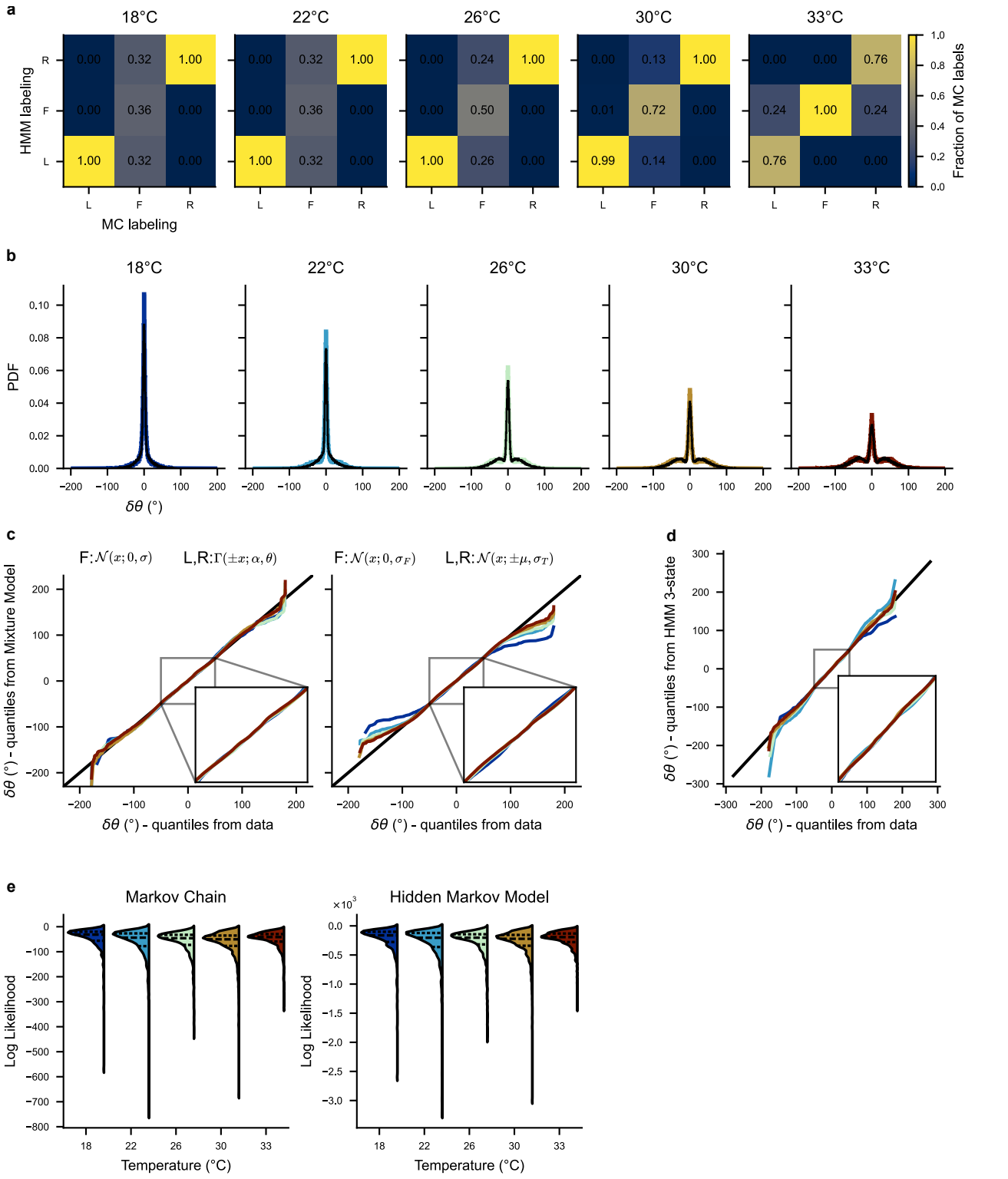

FIG. 7. **Supplementary panels to Fig.2**: **(a)** Confusion matrices between labeling of MC and HMM for all temperatures (normalized with respect to the MC labeling). **(b)** Comparison between the distributions of reorientation angles observed from the data (colored) and the distributions of reorientation angles generated by the 3-state Hidden Markov Model (HMM; black), for each temperature. **(c)** Quantile-Quantile plot between distributions of reorientation angles observed from the data and Mixture Models, at each temperature. *Left:* Mixture Model defined from a central Normal distribution (forward bouts) and two Gamma distributions (left and right turning bouts), corresponding to the model of HMM emissions. *Right:* Gaussian Mixture Model. *Insets:* Zoom on ±50°. **(d)** Quantile-Quantile plot between the distributions of reorientation angles observed from the data and the distributions of reorientation angles generated by HMM. *Insets:* Zoom on ±50°. **(e)** Distribution of log Likelihoods (LLHs) for both the Markov chains inferred from thresholded data (left) and HMM (right). For each model and each temperature, LLHs were computed for 100 models inferred from 50% of the trajectories (randomly constructed training set) and on the remaining 50% of the trajectories (testing set). Dashed lines show the quartiles of each distribution.

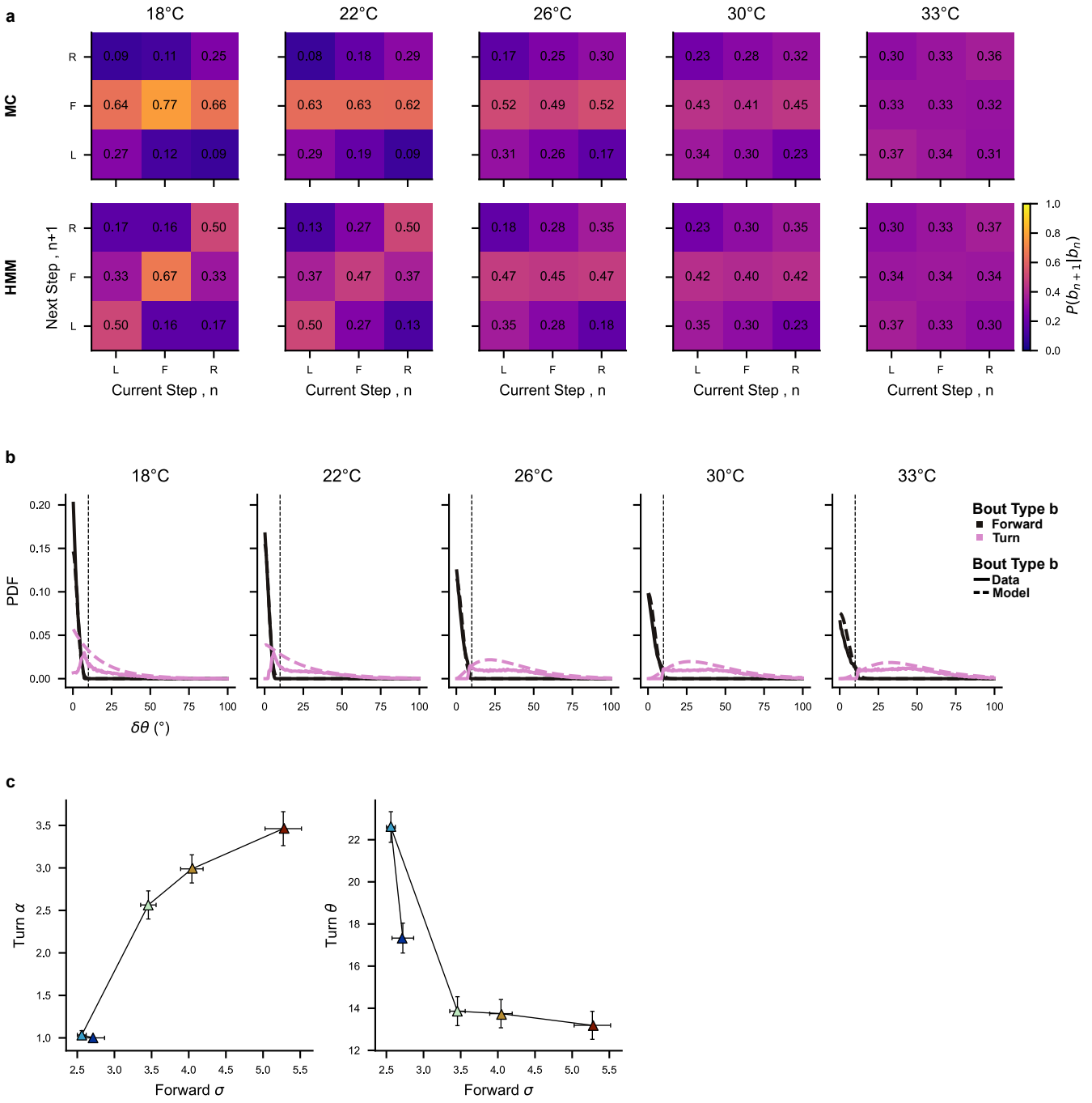

FIG. 8. **Supplementary panels to Fig.3**: **(a)** Transition matrices between forward (F), left (L) and right (R) states, for both the Markov chains inferred from thresholded data (MC) and Hidden Markov Model (HMM), and for each temperature. **(b)** Distributions of absolute reorientation angles labeled as forward bouts (solid black) and turning bouts (left or right; solid pink by the Hidden Markov Model (HMM). Dashed lines show the HMM emission distribution for forward and turning bouts (black and pink respectively). The threshold $\delta\theta_0 = 10^\circ$ used in the Markov Chain model is shown for reference as a vertical black line. **(c)** Parameters of the HMM emission distribution, with σ the standard deviation of the central Normal distribution (forward bouts), α and θ the shape and scale of the Gamma distribution (turning bouts). Each dot corresponds to one temperature, and error bars were computed from the minimum-maximum of 100 cross-validations (trained on randomly selected 50% of the datasets).

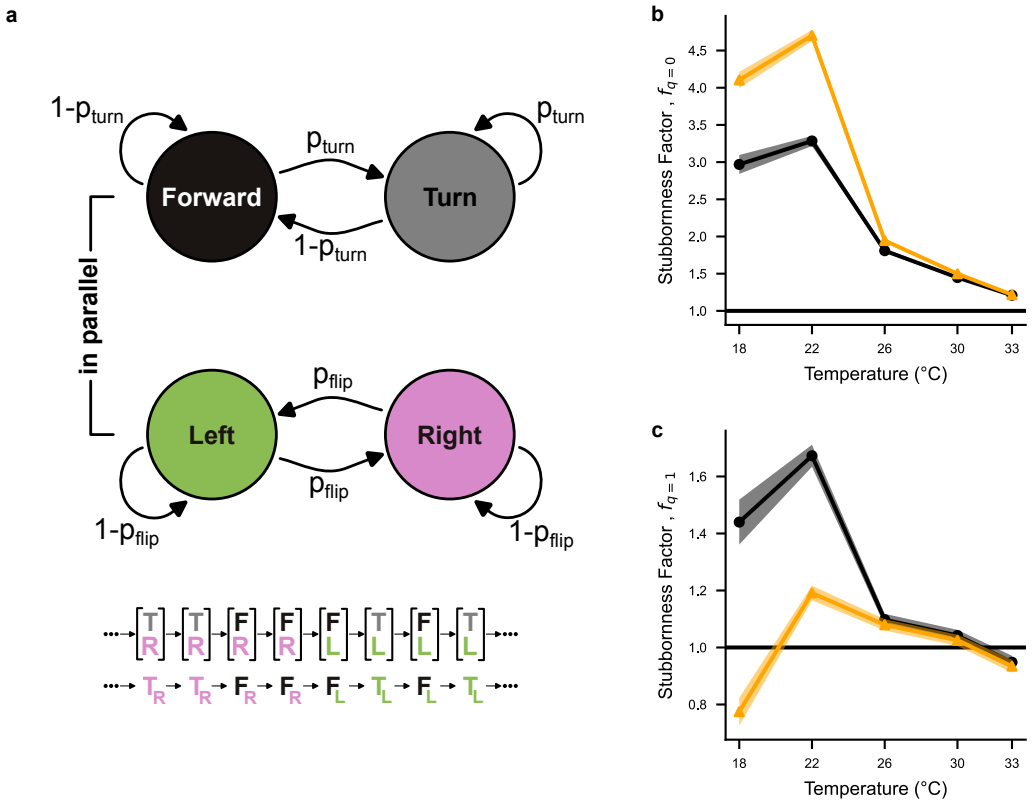

FIG. 9. **Supplementary panels to Fig.4**: **(a)** Diagram of the 4-state Markov chain used in previous publications [9, 11]. Two Markov Chains run in parallel, with the first chain controlling bout type (forward or turn) and the second controlling direction (left or right). With this model, the system can be in one of four states: $[T, L], [T, R], [F, L], [F, R]$, thus left and right states represent internal directional states (not only observed behavioral orientations). **(b)** Temperature dependence of the stubbornness factor at $q = 0$ intermediary Forward bouts ($f_{q=0} = \frac{P(L\to L)+P(R\to R)}{P(L\to R)+P(R\to L}$). We interpret this factor as a measurement of directional persistence during sequences of turning bouts. **(c)** Temperature dependence of the stubbornness factor at $q = 1$ intermediary Forward bouts ($f_{q=1} = \frac{P(L\to F\to L)+P(R\to F\to R)}{P(L\to F\to R)+P(R\to F\to L}$). We interpret this factor as a measurement of directional memory after one forward bout, which for a 3-state model is a second order non-markovianity. **(b,c)** Throughout this figure, the width of the shaded curves represent the estimated error in stubbornness factor (see Materials and Methods IV C).

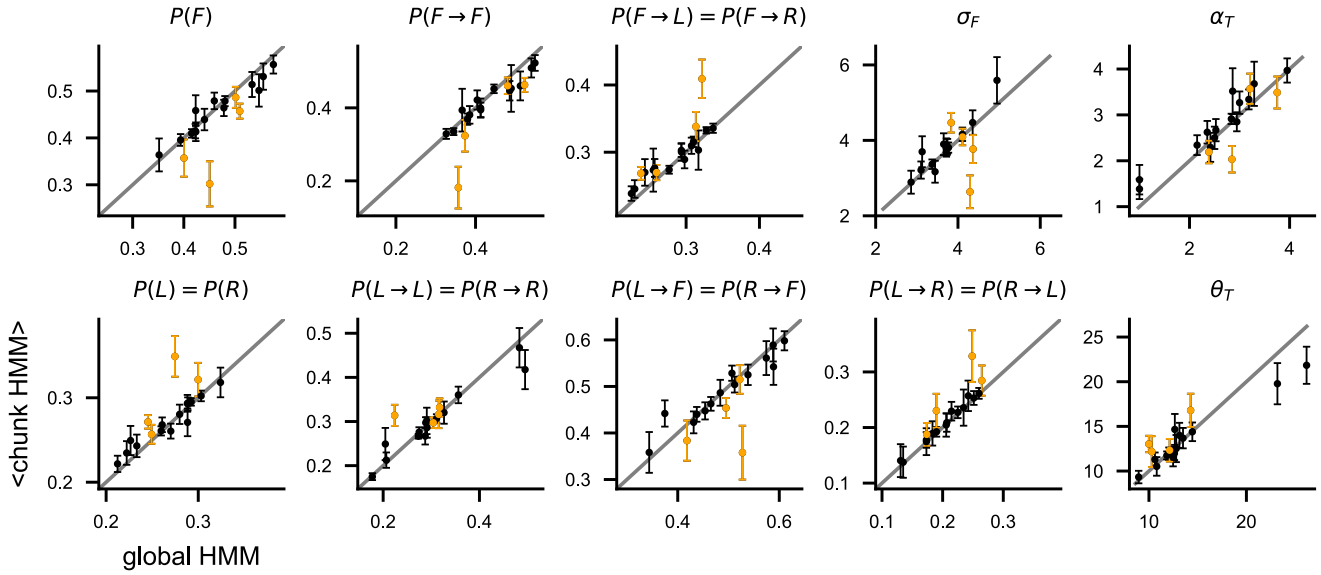

FIG. 10. **Supplementary panels to Fig.5** Hidden Markov Model parameters inferred from all trajectories from an individual fish, compared with the average parameters inferred from chunks of that fish's trajectories. All HMM parameters are shown. Each dot represents a fish, with error bars corresponding to standard error of the mean. Points labeled in orange correspond to fish misidentified in Fig.5c.