PNAS
www.pnas.org

Supplementary Information for

Bayesian inference in ring attractor networks

Anna Kutschireiter, Melanie A. Basnak, Rachel I. Wilson and Jan Drugowitsch

Jan Drugowitsch,
E-mail: jan_drugowitsch@hms.harvard.edu

This PDF file includes:

Supplementary text
Figs. S1 to S5
References for SI reference citations

Contents

1 Circular Kalman filtering **3**
- A Generative model . . . 3
- B Discrete-time Bayesian filtering . . . 3
  - B.1 Angular velocity observations . . . 4
  - B.2 HD observations . . . 4
  - B.3 The circular Kalman filter . . . 5
  - B.4 The quadratic approximation of the circular Kalman filter . . . 5
- C Coordinate transforms [Technical] . . . 5
- D Numerical benchmarks . . . 6
  - D.1 Bootstrap particle filter . . . 6
  - D.2 HD tracking performance measures . . . 7

2 Neural encoding example: encoding of the von Mises distribution with a linear probabilistic population code **8**
- A Tuning with respect to (true) HD ϕ_t . . . 8
- B Tuning with respect to HD estimate μ_t . . . 9

3 Details on Bayesian ring attractor dynamics and parameter tuning **11**
- A Network that exactly implements the circKF . . . 11
- B Network with quadratic nonlinearity . . . 12
- C Continuous vs. discrete networks . . . 13
- D Stochastic correction [Technical] . . . 13

4 Details on *Drosophila*-like network **15**
- A Connectivity motifs in the *Drosophila* HD system connectome . . . 15
- B A multi-network model mimicking the *Drosophila* HD system connectome . . . 15
  - B.1 $AV^{\pm}$ population . . . 16
  - B.2 INH population. . . . 16
  - B.3 Recurrent excitation within HD population . . . 17
  - B.4 Summary of network connectivities . . . 18
- C Drosophila-like network simulations and HD tracking performance . . . 19

5 The impact of neural noise on inference dynamics **20**
- A The qualitative impact of neural noise on inference dynamics . . . 20
- B How neural noise quantitatively impacts the dynamics of μ_t and κ_t . . . 20
  - B.1 The impact of neural noise on x_1 and x_2 . . . 20
  - B.2 The impact of neural noise on μ and κ . . . 21
  - B.3 Neural noise models . . . 21
- C Compensating for noisy neurons when performing inference . . . 22

6 Supplementary Figures **24**

Supporting Information Text

1. Circular Kalman filtering

Here, we present a derivation of the circular Kalman filter (circKF), which we use as an ideal observer model in the main text. The following derivation's main purpose is to provide the reader with some intuition behind the formalism, such that it uses a discrete-time approximation, followed by taking the continuous-time limit. For a mathematically-rigorous, continuous-time derivation of the circKF, please consult (1).

A. Generative model. Assuming time to be discretized in steps of $\mathrm{d}t$, the overall goal is to derive an online estimator for the unobserved true head direction (HD) $\phi_t \in [-\pi, \pi]$ at each point in time t, conditioned on a continuous stream of noisy angular velocity observations $V_t = \{v_0, v_{\mathrm{d}t}, \dots v_t\}$ (in the main text denoted $v_{0:t}$) with $v_\tau \in \mathbb{R}$ and HD observations $Z_t = \{z_0, z_{\mathrm{d}t}, \dots z_t\}$ (in the main text denoted $z_{0:t}$) with $z_\tau \in [-\pi, \pi]$. We assume that these observations are generated from the (true) angular velocity $\dot{\phi}_t = \frac{\phi_t - \phi_{t-\mathrm{d}t}}{\mathrm{d}t}$ and HD ϕ_t, respectively, and are corrupted by zero-mean noise at each point in time:

$$p(v_t|\phi_t, \phi_{t-\mathrm{d}t}) = \mathcal{N}\left(v_t; \frac{\phi_t - \phi_{t-\mathrm{d}t}}{\mathrm{d}t}, \frac{1}{\kappa_v\,\mathrm{d}t}\right), \quad [S1]$$

$$p(z_t|\phi_t) = \mathcal{VM}\left(z_t; \phi_t, \kappa_z\,\mathrm{d}t\right), \quad [S2]$$

where $\mathcal{VM}(\varphi; \mu, \kappa) = \frac{e^{\kappa\cos(\varphi-\mu)}}{2\pi I_0(\kappa)}$ denotes the von Mises distribution of a circular random variable φ with mean μ and precision κ. κ_v and κ_z refer to the precision of the angular velocity and HD observations, respectively. The precision $\kappa_z \mathrm{d}t$ of HD observations scales with $\mathrm{d}t$ to ensure that smaller "time steps" come with less informative HD observations to avoid "oversampling" in the $\mathrm{d}t \to 0$ limit. More technically, we need to ensure that the Fisher information that each HD observation has about the HD scales linearly with $\mathrm{d}t$. As we show in (1, Theorem 2), this Fisher information is given by $I_{z_t}(\phi_t) = \sqrt{2\gamma_z \mathrm{d}t}$ where γ_z is the HD observation Fisher information rate per unit time. For small $\mathrm{d}t \to 0$ we furthermore have $\gamma_z \mathrm{d}t \to (\kappa_z \mathrm{d}t)^2/2$ (see (1)) such that κ_z needs to be adjusted if the simulation time step size Δt changes in order to keep γ_z constant. As our simulations all use the same time step size, we safely ignore this subtlety for the remainder of this text.

We further assume that HD ϕ_t follows a diffusion on the circle, which serves as a dynamic prior over HD in terms of a transition density:

$$p(\phi_t|\phi_{t-\mathrm{d}t}) \sim \mathcal{N}\left(\phi_t; \phi_{t-\mathrm{d}t}, \frac{\mathrm{d}t}{\kappa_\phi}\right) \mod 2\pi, \quad [S3]$$

Here, $\kappa_\phi \geq 0$ is related to the inverse diffusion constant: a large κ_ϕ implies limited diffusion and an almost-stationary stochastic process. In this case, past observations are generally highly informative about the current HD. A small κ_ϕ implies that HD is most likely to change significantly from one time step to the next, indicating that past observations only provide limited information about our current HD.

B. Discrete-time Bayesian filtering. Given the posterior $p(\phi_{t-\mathrm{d}t}|Y_{t-\mathrm{d}t}, Z_{t-\mathrm{d}t})$ at some previous time-step $t - \mathrm{d}t$, we compute the posterior at the current time step t using the conditional dependencies of the model and Bayes' theorem:

$$\begin{aligned} p(\phi_t|V_t, Z_t) &\propto_{\phi_t} p(z_t|\phi_t) p(\phi_t|V_t, Z_{t-\mathrm{d}t}) \\ &= p(z_t|\phi_t) \int \mathrm{d}\phi_{t-\mathrm{d}t}\, p(\phi_t|\phi_{t-\mathrm{d}t}, v_t) p(\phi_{t-\mathrm{d}t}|Z_{t-\mathrm{d}t}, V_{t-\mathrm{d}t}). \end{aligned} \quad [S4]$$

This equation offers a way to *recursively* compute the current posterior density from the previous one, by taking two distinct steps: the so-called prediction and update step. The *prediction step* is a convolution between the previous posterior and the transition density $p(\phi_t|\phi_{t-\mathrm{d}t}, v_t)$, as implemented by the above integral. It tells us how the posterior is expected to evolve in a single time step when only observing angular velocity information, but no HD observations, are present, resulting in the prediction density $p(\phi_t|V_t, Z_{t-\mathrm{d}t})$. Note that the angular velocity observations v_t enter this step through the effective transition probability $p(\phi_t|\phi_{t-\mathrm{d}t}, v_t)$. In the *update step*, we multiply the result of the prediction step with the HD observation likelihood $p(z_t|\phi_t)$. Intuitively, this step can be understood as Bayesian cue integration between the prediction density and the HD observations.

In general, we will not be able to solve Eq. [S4] in closed form* for continuous variables like HD. We thus have to introduce approximations of $p(\phi_t|V_t, Z_t)$ that allow us to consistently perform prediction and update steps. Specifically, as one of the simplest choices for unimodal probability distributions for circular variables, we chose to approximate the posterior by a von Mises distribution,

$$p(\phi_t|V_t, Z_t) \approx \mathcal{VM}(\phi_t; \mu_t, \kappa_t). \quad [S5]$$

By using this approximation, the estimation task reduces to having to find evolution equations, conditioned on angular velocity observations v_t and HD observations z_t, for the two parameters μ_t and κ_t, which are sufficient to fully specify the posterior distribution. In what follows, we will consider the effect of angular velocity observations and HD observations on the two parameters separately.

* In fact, a closed-form solution is almost never achievable for continuous state-spaces. One of the few cases where it is is when prediction and update steps are linear Gaussians, in which case Eq. [S4] yields the Kalman filter.

B.1. Angular velocity observations. In Eq. [S4], angular velocity observations enter through a modified transition density $p(\phi_t|\phi_{t-\mathrm{d}t}, v_t)$, which can be computed using Bayes' theorem:

$$p(\phi_t|v_t, \phi_{t-\mathrm{d}t}) \propto_{\phi_t} p(v_t|\phi_t, \phi_{t-\mathrm{d}t}) p(\phi_t|\phi_{t-\mathrm{d}t}). \tag{S6}$$

The modified transition probability is again a Gaussian, as can be seen from its logarithm being quadratic in ϕ_t,

$$\begin{aligned} -\log p(\phi_t|v_t, \phi_{t-\mathrm{d}t}) &= \frac{\kappa_v \mathrm{d}t}{2}\left(v_t - \frac{\phi_t - \phi_{t-\mathrm{d}t}}{\mathrm{d}t}\right)^2 + \frac{\kappa_\phi}{2\mathrm{d}t}\left(\phi_t - \phi_{t-\mathrm{d}t}\right)^2 + \mathcal{R} \\ &= \frac{1}{2}\frac{\kappa_v + \kappa_\phi}{\mathrm{d}t}\left(\phi_t - \phi_{t-\mathrm{d}t}\right)^2 - \frac{\kappa_v}{\mathrm{d}t}\left(\phi_t - \phi_{t-\mathrm{d}t}\right) v_t \mathrm{d}t + \mathcal{R} \\ &= \frac{1}{2}\frac{\kappa_v + \kappa_\phi}{\mathrm{d}t}\left(\phi_t - \left(\phi_{t-\mathrm{d}t} + \frac{\kappa_v}{\kappa_v + \kappa_\phi} v_t \mathrm{d}t\right)\right)^2 + \mathcal{R}, \end{aligned} \tag{S7}$$

where terms independent of ϕ_t, collectively denoted by $\mathcal{R}$, can be absorbed in the normalization. Hence, the modified transition probability reads:

$$p(\phi_t|v_t, \phi_{t-\mathrm{d}t}) = \mathcal{N}\left(\phi_t; \phi_{t-\mathrm{d}t} + \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t \,\mathrm{d}t, \frac{\mathrm{d}t}{\kappa_\phi + \kappa_v}\right) \mod 2\pi. \tag{S8}$$

Together with the assumption that the posterior of the last time step, $p(\phi_{t-\mathrm{d}t}|V_{t-\mathrm{d}t}, Z_{t-\mathrm{d}t})$, is given by a von Mises distribution with mean $\mu_{t-\mathrm{d}t}$ and precision $\kappa_{t-\mathrm{d}t}$, we can write down the expression for the prediction density $p(\phi_t|V_t, Z_{t-\mathrm{d}t})$ (cf. first line in Eq. [S4]):

$$\begin{aligned} p(\phi_t|V_t, Z_{t-\mathrm{d}t}) &= \int_{-\pi}^{\pi} d\phi_{t-\mathrm{d}t}\, p(\phi_t|v_t, \phi_{t-\mathrm{d}t}) p(\phi_{t-\mathrm{d}t}|Z_{t-\mathrm{d}t}, V_{t-\mathrm{d}t}) \\ &= \int_{-\pi}^{\pi} d\phi_{t-\mathrm{d}t}\, \mathcal{N}\left(\phi_t; \phi_{t-\mathrm{d}t} + \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t \,\mathrm{d}t, \frac{\mathrm{d}t}{\kappa_\phi + \kappa_v}\right) \mathcal{VM}\left(\phi_{t-\mathrm{d}t}; \mu_{t-\mathrm{d}t}, \kappa_{t-\mathrm{d}t}\right). \end{aligned} \tag{S9}$$

Unfortunately, there is no closed-form solution for this integral. To approximate the prediction density $p(\phi_t|V_t, Z_{t-\mathrm{d}t})$ at each moment in time by a von Mises density $\mathcal{VM}(\phi_t; \tilde{\mu}_t, \tilde{\kappa}_t)$, we will use a more sophisticated approximation method, namely a projection filter (2). Such a filter ensures that this approximation is optimal by minimizing the infinitesimal Kullback-Leibler divergence at each moment in time. The technical details can be found in (1), and in this SI we limit ourselves to giving the final result:

$$\mathrm{d}\mu_t = \frac{\kappa_v}{\kappa_v + \kappa_\phi} v_t \,\mathrm{d}t, \tag{S10}$$

$$\mathrm{d}\kappa_t = -\frac{f(\kappa_t)}{2(\kappa_v + \kappa_\phi)} \kappa_t \,\mathrm{d}t. \tag{S11}$$

Here, the decay of the certainty κ_t is governed by the nonlinear function

$$f(\kappa_t) = \frac{A(\kappa_t)}{\kappa_t - A(\kappa_t) - \kappa A(\kappa_t)^2}, \quad \text{with } A(\kappa_t) = \frac{I_1(\kappa_t)}{I_0(\kappa_t)}, \tag{S12}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of the first kind of order 0 and 1. This function takes care of the fact that the true HD ϕ_t follows a diffusion on the circle, which becomes particularly relevant for small values of κ_t. In particular, $f(\kappa_t) \approx 1$ for small κ_t and $f(\kappa_t) \approx 2\kappa_t - 2$ for large κ_t, indicating that the decay is asymptotically quadratic.

B.2. HD observations. Angular-valued HD observations z_t are integrated by multiplying the observation likelihood $p(z_t|\phi_t)$ with the prediction density $p(\phi_t|V_t, Z_{t-\mathrm{d}t})$. If the prediction density is also von Mises (which is the assumption above), this cue integration is closed:

$$\begin{aligned} p(\phi_t|z_t, \mathrm{d}y_t) &= \mathcal{VM}(z_t; \phi_t, \kappa_z \,\mathrm{d}t) \cdot \mathcal{VM}(\phi_t; \tilde{\mu}_t, \tilde{\kappa}_t) \\ &\propto \exp\left(\begin{pmatrix}\cos\phi_t \\ \sin\phi_t\end{pmatrix}^\top \cdot \left(\kappa_z \,\mathrm{d}t \begin{pmatrix}\cos z_t \\ \sin z_t\end{pmatrix} + \tilde{\kappa}_t \begin{pmatrix}\cos\tilde{\mu}_t \\ \sin\tilde{\mu}_t\end{pmatrix}\right)\right) \qquad [S13] \\ &\stackrel{!}{=} \exp\left(\begin{pmatrix}\cos\phi_t \\ \sin\phi_t\end{pmatrix}^\top \cdot \kappa_t \begin{pmatrix}\cos\mu_t \\ \sin\mu_t\end{pmatrix}\right). \qquad [S14] \end{aligned}$$

Thus, the natural parameters of the posterior distribution, $\boldsymbol{x}_t = (x_1, x_2) = (\kappa_t \cos\mu_t, \kappa_t \sin\mu_t)^\top$, can be written as the sum of the natural parameters of the prediction density and the likelihood[†]:

$$\boldsymbol{x}_t = \tilde{\boldsymbol{x}}_t + \kappa_z \begin{pmatrix}\cos z_t \\ \sin z_t\end{pmatrix} \mathrm{d}t \tag{S15}$$

$$d\boldsymbol{x}_t = \boldsymbol{x}_t - \tilde{\boldsymbol{x}}_t = \kappa_z \begin{pmatrix}\cos z_t \\ \sin z_t\end{pmatrix} \mathrm{d}t. \tag{S16}$$

[†] This is not too surprising, as it is well known that in exponential family distributions these update steps boil down to adding up the natural parameters.

The updates of the parameters μ_t and κ_t of the von Mises distribution due to the observation z_t are obtained by transforming the update of $\boldsymbol{x}_t$ to polar coordinates:

$$\mathrm{d}\mu_t^{\text{update}} = \mathrm{d}\arctan 2\,(x_2, x_1) = \frac{\kappa_z}{\kappa_t}\sin(z_t - \mu_t)\mathrm{d}t \quad [S17]$$

$$\mathrm{d}\kappa_t^{\text{update}} = \mathrm{d}\sqrt{x_1^2 + x_2^2} = \kappa_z \cos(z_t - \mu_t)\mathrm{d}t. \quad [S18]$$

B.3. The circular Kalman filter. In the continuum limit $\mathrm{d}t \to 0$, we do not distinguish between the parameters of the prediction density, $\tilde{\mu}_t$ and $\tilde{\kappa}_t$, and that of the posterior density, μ_t and κ_t. The circKF equations result from taking the prediction and update steps simultaneously, thereby combining Eq. [S10] with Eq. [S17] for the mean dynamics, and Eq. [S11] with Eq. [S18] for the precision dynamics:

$$\mathrm{d}\mu_t = \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t\,\mathrm{d}t + \frac{\kappa_z}{\kappa_t}\sin(z_t - \mu_t)\mathrm{d}t, \quad [S19]$$

$$\mathrm{d}\kappa_t = -\frac{f(\kappa_t)}{2(\kappa_\phi + \kappa_v)}\kappa_t \mathrm{d}t + \kappa_z \cos(z_t - \mu_t)\mathrm{d}t. \quad [S20]$$

Here, we adhered to expressing these equations in terms of their infinitesimal difference, $\mathrm{d}\mu_t$ and $\mathrm{d}\kappa_t$, instead of a differential equation. This is a standard way to express stochastic differential equations (SDEs), which makes it more straightforward to deal with the non-linear time scaling of the HD observations z_t.

B.4. The quadratic approximation of the circular Kalman filter. If κ_t is sufficiently large, the nonlinearity $f(\kappa_t)$ can be approximated by a linear function, $f(\kappa_t) \approx 2\kappa_t - 2$, such that the decay in Eq. [S20] becomes quadratic:

$$\mathrm{d}\kappa_t \approx -\frac{1}{\kappa_\phi + \kappa_v}\left(\kappa_t^2 - \kappa_t\right)\,\mathrm{d}t + \kappa_z \cos(z_t - \mu_t)\mathrm{d}t. \quad [S21]$$

We use this approximation when implementing the Bayesian ring attractor network.

C. Coordinate transforms [Technical]. The von Mises distribution can be parametrized by its mean and precision parameters, μ and κ, or in terms of its natural parameters, $\boldsymbol{x} = (x_1, x_2)^\top = (\kappa\cos\mu, \kappa\sin\mu)^\top$. These two parametrizations are perfectly equivalent, and can be thought of as the polar and Cartesian coordinates of a vector, respectively. Except when $\kappa = 0$, which we assume to never occur, we can go back and forth between these representations by performing a coordinate transformation.

For the neural network we describe further below, it is easier to decode $\boldsymbol{x}$ than μ and κ from neural population activity. Thus, it is useful to express the circular Kalman filter as SDEs for $\boldsymbol{x}$. Unfortunately, we cannot simply find these SDEs by applying a coordinate transform to Eqs. [S19] and [S20]. Technically speaking, since the angular velocity observations v_t follow a stochastic process, we have to take into account second-order derivatives, which is called Itô's lemma in stochastic calculus (see (3) for an introduction). As we will here show in a slightly technical argument, using stochastic instead of ordinary calculus explains why we need an additional decay term in the network implementation in Sec. 3 that would not arise from a simple coordinate transform. Understanding this argument is not required for understanding our general theory and network implementation, and thus can safely be skipped.

First, we express the generative model in Eqs. [S3] and [S1] in terms of their equivalent Itô stochastic differential equations (SDEs). Defining the infinitesimal increment $\mathrm{d}u_t := v_t\,\mathrm{d}t$, the SDEs read:

$$\mathrm{d}\phi_t = \frac{1}{\sqrt{\kappa_\phi}}\mathrm{d}W_t \quad [S22]$$

$$\mathrm{d}u_t = \mathrm{d}\phi_t + \frac{1}{\sqrt{\kappa_v}}\mathrm{d}V_t, = \frac{1}{\sqrt{\kappa_\phi}}\mathrm{d}W_t + \frac{1}{\sqrt{\kappa_v}}\mathrm{d}V_t, \quad [S23]$$

where $\mathrm{d}W_t \in \mathbb{R} \sim \mathcal{N}(0, \mathrm{d}t)$ and $\mathrm{d}V_t \in \mathbb{R} \sim \mathcal{N}(0, \mathrm{d}t)$ are uncorrelated scalar-valued Brownian motion processes with $\mathrm{d}W_t\,\mathrm{d}V_t = 0$. Since the variance of Brownian motion processes grows linearly in time, we have that $(\mathrm{d}W_t)^2 = \mathrm{d}t$, $(\mathrm{d}V_t)^2 = \mathrm{d}t$, and thus $(\mathrm{d}u_t)^2 = \left(\frac{1}{\kappa_\phi} + \frac{1}{\kappa_v}\right)\mathrm{d}t$. The second equality in Eq. [S23] tells us that whenever angular velocity observations are drawn from the 'true' generative model in Eq. [S1], they automatically inherit the noise of the process that was used to generate ϕ_t.

Itô's lemma tells us how to perform a variable transformation from a stochastic process x_t, which is governed by an Itô SDE, to another stochastic process $y_t = g(x_t)$:

$$\mathrm{d}y_t = dg(x_t) = \left.\frac{\partial g(x)}{\partial x}\right|_{x=x_t}\mathrm{d}x_t + \frac{1}{2}\left.\frac{\partial^2 g(x)}{\partial x^2}\right|_{x=x_t}(\mathrm{d}x_t)^2. \quad [S24]$$

Thus, we can use Itô's lemma to transform the dynamics of μ_t and κ_t in Eqs. [S10] and [S11] to the dynamics of the natural parameters of the von Mises distribution. Note that, since the dynamics of κ_t are independent of the angular velocity

observations, Eq. [S11] is deterministic with $(\mathrm{d}\kappa_t)^2 = 0$:

$$\begin{aligned}\mathrm{d}\boldsymbol{x}_t &= \mathrm{d}\left[\kappa_t \begin{pmatrix}\cos\mu_t\\ \sin\mu_t\end{pmatrix}\right] = \begin{pmatrix}\cos\mu_t\\ \sin\mu_t\end{pmatrix}\mathrm{d}\kappa_t + \kappa_t\begin{pmatrix}-\sin\mu_t\\ \cos\mu_t\end{pmatrix}\mathrm{d}\mu_t + \frac{1}{2}\kappa_t\begin{pmatrix}-\cos\mu_t\\ -\sin\mu_t\end{pmatrix}(\mathrm{d}\mu_t)^2\\ &= -\frac{f(\kappa_t)}{2(\kappa_\phi+\kappa_v)}\kappa_t\begin{pmatrix}\cos\mu_t\\ \sin\mu_t\end{pmatrix}\mathrm{d}t + \frac{\kappa_t\kappa_v}{\kappa_\phi+\kappa_v}\begin{pmatrix}-\sin\mu_t\\ \cos\mu_t\end{pmatrix}\mathrm{d}u_t - \frac{1}{2}\begin{pmatrix}x_1\\ x_2\end{pmatrix}\frac{\kappa_v^2}{(\kappa_v+\kappa_\phi)^2}(\mathrm{d}u_t)^2 \qquad [S25]\\ &= -\frac{1}{2}\frac{f(\kappa_t)}{\kappa_v+\kappa_\phi}\boldsymbol{x}_t\,\mathrm{d}t - \frac{1}{2}\frac{\kappa_v/\kappa_\phi}{\kappa_v+\kappa_\phi}\boldsymbol{x}_t\,\mathrm{d}t + \frac{\kappa_v}{\kappa_v+\kappa_\phi}\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\boldsymbol{x}_t\,\mathrm{d}u_t.\end{aligned}$$

Here, the additional decay term $-\frac{1}{2}\frac{\kappa_v/\kappa_\phi}{\kappa_v+\kappa_\phi}\boldsymbol{x}_t\mathrm{d}t$ arises from the stochastic nature of the increment process u_t.

Since HD observations z_t are added on the level of natural parameters (cf. Eq. [S16]), these can be included in a straightforward manner, yielding the circular Kalman filter in its natural parameter form:

$$\mathrm{d}\boldsymbol{x}_t = -\frac{1}{2}\frac{f(\kappa_t)+\kappa_v/\kappa_\phi}{\kappa_v+\kappa_\phi}\boldsymbol{x}_t\,\mathrm{d}t + \frac{\kappa_v}{\kappa_v+\kappa_\phi}\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\boldsymbol{x}_t\,\mathrm{d}u_t + \kappa_z\begin{pmatrix}\cos z_t\\ \sin z_t\end{pmatrix}\mathrm{d}t. \qquad [S26]$$

D. Numerical benchmarks. As described above, the circKF approximates the posterior at each point in time by a von Mises distribution, and thus is itself an approximate algorithm. To compare its performance, and that of the Bayesian ring attractor to the truly best filtering performance for the assumed generative model, we additionally used a Bootstrap particle filter, which is exact in the limit of an infinite number of particles. Here, we first outline the algorithm itself, and then discuss how we assess filtering performance in general, to compare performance across algorithms.

D.1. Bootstrap particle filter. As a numerical benchmark, we used a Sequential Importance Sampling/Resampling particle filter (4) (SIS-PF; member of the family of Bootstrap particle filters) that we modified to be applicable to angular velocity observations. Here, we briefly outline the numerical implementation of the SIS-PF for our particular filtering problem, and refer the reader to more specialized literature for derivation and convergence results (e.g., in (4, 5)).

The principle behind particle filters is that they provide a weighted empirical estimate of the posterior distribution,

$$p\left(\phi_t|V_t, Z_t\right) \approx \sum_{i=1}^{N} w_t^{(i)}\delta(\phi_t - \varphi_t^{(i)}), \qquad [S27]$$

where we refer to $w_t^{(i)}$ as the importance weight of the i-th particle with position $\varphi_t^{(i)}$. Weighted particle filters are asymptotically exact, i.e. they provide us with the best possible inference performance in the limit of infinitely many particles $N \to \infty$. At each discrete time step, the N particles in the SIS-PF are propagated according to the proposal density π, which we chose to correspond to the modified transition density in Eq. [S8]:

$$\begin{aligned}&\pi\left(\varphi_t^{(j)}|\varphi_{t-\Delta t}^{(j)}, v_t\right)\\ &\quad= \mathcal{N}\left(\varphi_t^{(j)}; \varphi_{t-\Delta t}^{(j)} + \frac{\kappa_v}{\kappa_v+\kappa_\varphi}v_t\,\Delta t, \frac{\Delta t}{\kappa_\varphi+\kappa_v}\right) \mod 2\pi. \qquad [S28]\end{aligned}$$

Subsequently, each particle j is weighted at each time step according to how well the proposed particle distribution fits to the HD observation z_t. This is equivalent to multiplying the previous weight with the observation likelihood (Eq. [S2]):

$$w_t^{(i)} = w_{t-\Delta t}^{(i)} \cdot \mathcal{VM}\left(z_t; \varphi_t^{(i)}, \kappa_z\Delta t\right). \qquad [S29]$$

Lastly, the particles are re-weighted such that the importance weights sum to 1, $\sum_i w_t^{(i)} = 1$:

$$w_t^{(i)} \leftarrow \frac{w_t^{(i)}}{\sum_j w_t^{(j)}} \qquad [S30]$$

In our simulations, we used $N = 10^3$ particles, which is sufficient if HD observations are present.

Mean μ_t and precision $r_t \in [0, 1]$ of the filtering distribution approximated by the SIS-PF can be determined at each time step according to a weighted average on the circle, i.e. the first circular moment:

$$r_t\exp(i\mu_t) = \sum_{j=1}^{N} w_t^{(j)}\exp\left(i\varphi_t^{(j)}\right). \qquad [S31]$$

D.2. HD tracking performance measures. In the main text, we quantified HD tracking performance by estimating the absolute value of the circular average distance between the estimate μ_T at the end of the trial (using the mean of the filter posterior, which is the filter's best guess), and the true HD ϕ_T, averaged across P simulations with different noisy observation sequences, $v_0, \ldots, v_T$ and $z_0, \ldots, z_T$:

$$m_1 = \frac{1}{P} \sum_{k=1}^{P} \exp\left(i\left(\mu_T^{(k)} - \phi_T^{(k)}\right)\right). \quad [S32]$$

Here, m_1 is a complex number, and HD tracking performance corresponds to its absolute value, $|m_1|$ (larger = better / more accurate). Note that this absolute value is one minus the circular variance of the error. As this variance is bounded by zero and one, zero variance implies a performance of $|m_1| = 1$, and maximum variance of one implies a performance of $|m_1| = 0$. To get a sense of how estimates μ_T are distributed around the true HD ϕ_T for a given value of $|m_1|$, we provide representative histograms in Fig. S5.

2. Neural encoding example: encoding of the von Mises distribution with a linear probabilistic population code

In the main text, we assume a bump-like encoding of the HD posterior belief whose bump amplitude is scaled by the encoded certainty κ_t. This implies that the amplitude of the first Fourier component is proportional to the certainty (see main text Eq. (3)). This is trivially fulfilled for the cosine-shaped tuning curves that we used for illustration in the main text (main text Fig. 2). Here, we will demonstrate that this also holds for a more elaborate bump encoding scheme: specifically, we consider the case of a linear probabilistic population code (lPPC) (6–8) with independent Poisson neural noise. The central idea behind such an lPPC is that neuronal activity encodes an exponential family probability distribution, e.g., about HD, such that the natural parameters of this distribution can be retrieved through linear operations, that is, a weighted sum of neural activity.

In what follows, we will first show that an lPPC for a von Mises distribution with independent Poisson neurons gives rise to von Mises shape tuning curves, which are scaled by the encoded certainty (following (6)). Using this result, we will derive the population activity profile as a function of the encoded *estimate* and certainty that results from this encoding scheme, and show that the amplitude of this profile is indeed also proportional to the encoded certainty.

A. Tuning with respect to (true) HD ϕ_t. We assume that tuning curves of the population encoding the posterior $p(\phi_t|V_t, Z_t)$ can be described by a typical shape $\tilde{f}$, which is scaled by the population gain g. That is, the tuning curve of a single neuron i is given by $f_i(\phi_t) = g\,\tilde{f}_i(\phi_t)$. Following (6), we further assume that the neuronal population consists of N independent Poisson neurons, which densely tile the stimulus space of true HDs, ϕ. Thus, we can write down the probability of a population firing pattern $\boldsymbol{r} \in \mathbb{R}_+^N$ as

$$\begin{aligned} p(\boldsymbol{r}|\phi_t, g) &= \prod_i \frac{(g\tilde{f}_i(\phi_t))^{r_i}}{r_i!} \exp\left(-g\tilde{f}_i(\phi_t)\right) \\ &= \exp\left(\sum_i r_i \log(g\,\tilde{f}_i(\phi_t)) - \sum_i \log r_i! - \sum_i g\,\tilde{f}_i(\phi_t)\right) \\ &\propto_{\phi_t} \exp\left(\sum_i r_i \log \tilde{f}_i(\phi_t)\right), \end{aligned} \tag{S33}$$

where we used that $\sum_i g\,\tilde{f}_i(\phi_t)$ is approximately independent of HD ϕ_t due to the dense-tiling assumption.

Assuming that $p(\phi_t|\boldsymbol{r})$ follows an exponential family distribution, such as the von Mises distribution, an lPPC requires that the natural parameters of this distribution can be recovered from the population activity by a linear operation, i.e., a weighted sum. For a general exponential family distribution with d sufficient statistics $\mathbf{T}(\phi_t) \in \mathbb{R}^d$, and natural parameters $\boldsymbol{x}$, we thus can re-parametrize the distribution in terms of the the population activities $\boldsymbol{r}$ (6):

$$\begin{aligned} p(\phi_t|\boldsymbol{r}) &= \frac{1}{Z(\phi_t, \boldsymbol{x})} \exp\left(\mathbf{T}(\phi_t)^T \cdot \boldsymbol{x}\right) \\ &= \frac{1}{Z(\phi_t, \boldsymbol{r})} \exp\left(\mathbf{T}(\phi_t)^T \cdot A\boldsymbol{r}\right), \end{aligned} \tag{S34}$$

where the decoder matrix $A \in \mathbb{R}^{d\times N}$ is defined via $\boldsymbol{x} = A\boldsymbol{r}$. Assuming a uniform prior over HD, that is, $p(\phi_t) \propto 1$, we can relate Eqs. [S33] and [S34] by Bayes' rule, $p(\phi_t|\boldsymbol{r}) \propto p(\boldsymbol{r}|\phi_t, g)$. This results in the following conditions for the tuning curves:

$$p(\phi_t|\boldsymbol{r}) \propto_{\phi_t} p(\boldsymbol{r}|\phi_t), \tag{S35}$$

$$\Rightarrow \log \tilde{\mathbf{f}}(\phi_t) = A^T \cdot \mathbf{T}(\phi_t). \tag{S36}$$

For a von Mises distribution, the natural parameters are given by $\mathbf{T}(\phi_t) = (\cos\phi_t, \sin\phi_t)^T$. Thus, the argument of the exponential in the neurons' tuning curves is a linear combination of sines and cosines. This, in turn, can be written as a single cosine $\propto c\cos(\phi_t - \phi_i)$, where $\phi_i \in [-\pi, \pi]$ denotes the "preferred HD" of neuron i. The tuning curve of a single neuron is thus von-Mises shaped, i.e.,

$$\tilde{f}_i(\phi_t) = \exp\left(\xi\cos(\phi_t - \phi_i)\right), \tag{S37}$$

where ξ is an additional parameter that controls the width of the tuning curves. Furthermore, the decoder matrix is constrained via $(A^T)_i = \xi\,(\cos\phi_i, \sin\phi_i)$.

In order to determine the population gain g, note that we require the natural parameters of the von Mises distribution, $\boldsymbol{x} = \kappa_t\,(\sin\mu_t, \cos\mu_t)$, to be linearly decodable from the population activity via $\boldsymbol{x} = A\boldsymbol{r}$. Since $\boldsymbol{x}$ is proportional in κ_t, this linearity implies that the overall population activity $\boldsymbol{r}$ should also be overall scaled by κ_t. Hence, the tuning curve of a neuron with preferred HD ϕ_i reads:

$$f_i(\phi_t) = g\tilde{f}_i(\phi_t) = \kappa_t \exp\left(\xi\cos(\phi - \phi_i)\right). \tag{S38}$$

To summarize, an lPPC with independent Poisson neurons gives rise to von Mises shaped tuning curves, whose gain is scaled by the encoded certainty κ_t. Importantly, unlike for the encoded von Mises distribution, an increase in certainty κ_t does not cause the resulting activity profile to sharpen.

B. Tuning with respect to HD estimate μ_t. Tuning to true HD ϕ_t can only be measured if we have access to the encoded HD estimate. To instead find the tuning with respect to μ_t and κ_t that parametrize the *distribution* of ϕ_t, we need to average the neuron's tuning for a given μ_t and κ_t over all possible realizations of ϕ_t. This results in the following tuning with respect to μ_t and κ_t:

$$\begin{aligned} f_i(\mu_t, \kappa_t) &= \int_{-\pi}^{\pi} d\phi_t \, f_i(\phi_t) \, \mathcal{VM}\left(\phi_t; \mu_t, \kappa_t\right) \\ &= \frac{\kappa_t}{2\pi I_0(\kappa_t)} \int_{-\pi}^{\pi} d\phi_t \, \exp\left(\xi \cos(\phi_t - \phi_i) + \kappa_t \cos(\phi_t - \mu_t)\right) \\ &= \frac{\kappa_t}{2\pi I_0(\kappa_t)} \int_{-\pi}^{\pi} d\phi_t \, \exp\left(\tilde{\kappa}_{t,i} \cos(\phi_t - \tilde{\mu}_i)\right) \\ &= \kappa_t \frac{I_0(\tilde{\kappa}_{t,i})}{I_0(\kappa_t)}, \end{aligned} \tag{S39}$$

with $\tilde{\kappa}_{t,i} = \sqrt{\xi^2 + \kappa_t^2 + 2\xi\kappa_t \cos(\phi_i - \mu_t)}$. This tuning curve is again bump-shaped, with a peak at the encoded HD estimate μ_t and the bump amplitude modulated by encoded certainty κ_t in a nonlinear manner.

For small values of encoded certainty, the tuning curve approaches a cosine-shaped tuning with a gain that is a nonlinear function of κ_t. To see this, we use the series expansion of the Bessel function for a small argument z,

$$I_0(z) = \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+1)} \left(\frac{z}{2}\right)^{2m} \approx 1 + \frac{1}{4}z^2 + \mathcal{O}(z^4), \tag{S40}$$

and write for the tuning curve in the small-κ_t limit

$$\kappa_t \frac{I_0(\tilde{\kappa}_{t,i})}{I_0(\kappa_t)} \approx \frac{\kappa_t}{I_0(\kappa_t)} \left(1 + \frac{1}{2}\tilde{\kappa}_{t,i}^2\right) = \frac{\kappa_t}{I_0(\kappa_t)} \left(1 + \frac{1}{4}\left(\xi^2 + \kappa_t^2 + \xi\kappa_t \cos(\phi_i - \mu_t)\right)\right). \tag{S41}$$

Thus, the tuning curve of a neuron i for small values of κ_t is cosine-shaped, and modulated by the nonlinear factor $\frac{\xi\kappa_t^2}{4I_0(\kappa_t)}$, which asymptotically approaches $\frac{\xi}{4}\kappa_t^2$ for $\kappa_t \to 0$.

For large values of κ_t, the tuning curve is von-Mises shaped and the gain is asymptotically linear in encoded certainty. To see this, we use the Hankel expansion of the Bessel function $I_0(z)$ in the limit of large arguments z:

$$I_0(z) \approx \frac{e^z}{\sqrt{2\pi z}z} + \mathcal{O}\left(\frac{1}{z^2}\right), \tag{S42}$$

and simplify

$$\kappa_t \frac{I_0(\tilde{\kappa}_{t,i})}{I_0(\kappa_t)} \approx \kappa_t \sqrt{\frac{\kappa_t}{\tilde{\kappa}_{t,i}}} \exp\left(\tilde{\kappa}_{t,i} - \kappa_t\right). \tag{S43}$$

Taylor-expanding the exponent $\tilde{\kappa}_{t,i} - \kappa_t$ for small values of $1/\kappa_t$ yields,

$$\tilde{\kappa}_{t,i} - \kappa_t = \kappa_t \sqrt{1 + \frac{\xi^2}{\kappa_t^2} + \frac{\xi}{\kappa_t}\cos(\phi_i - \mu_t)} - \kappa_t \approx \frac{\xi}{2}\cos(\phi_i - \mu_t) + \frac{\xi^2}{2\kappa_t} + \mathcal{O}\left(\frac{1}{\kappa_t^2}\right). \tag{S44}$$

Further, the pre-factor $\sqrt{\frac{\tilde{\kappa}_{t,i}}{\kappa_t}} \to 1$, and thus the tuning curve in the large-κ_t limit reads:

$$f_i(\mu_t, \kappa_t) \to \kappa_t \exp\left(\frac{\xi}{2}\cos(\phi_i - \mu_t)\right). \tag{S45}$$

The choice of the width parameter ξ determines how large κ_t has to be for the tuning curve to scale linearly with encoded certainty.

In Fig. S1, we demonstrate these limits (assuming $\xi = 1$ without loss of generality), and find numerically that linear scaling of the population activity amplitude holds well even for small κ_t (e.g., $\kappa_t \sim 1$, cf. Fig. S1f). In addition, the width of the profile saturates quickly as we increase κ_t (which indicates the transition from cosine-shaped to von-Mises shaped tuning curve), which makes the shape almost independent of κ_t. Therefore, the population profile is not just a rescaled version of the encoded probability distribution (Fig. S1c), because an increase in certainty does not cause the bump to sharpen indefinitely.

The linear scaling of the amplitude with κ_t, and (almost) constant width, indicate that the parameters of the von Mises distribution, μ_t and κ_t, can be retrieved from the population activity by computing the first Fourier coefficients:

$$\mathcal{F}_1^{\text{even}}\left[f_i(\mu_t, \kappa_t)\right] := \frac{1}{\pi}\int_{-\pi}^{\pi} d\phi_i \, f_i(\mu_t, \kappa_t)\cos(\phi_i) \propto \kappa_t \cos\mu_t = x_{t,1}, \tag{S46}$$

$$\mathcal{F}_1^{\text{odd}}\left[f_i(\mu_t, \kappa_t)\right] := \frac{1}{\pi}\int_{-\pi}^{\pi} d\phi_i \, f_i(\mu_t, \kappa_t)\sin(\phi_i) \propto \kappa_t \sin\mu_t = x_{t,2}. \tag{S47}$$

The certainty κ_t can be retrieved via $\kappa_t = \sqrt{x_{t,1}^2 + x_{t,2}^2}$, and thus is proportional to the amplitude c_1 of the first Fourier component in amplitude-phase form. Likewise, the mean μ_t is the angle of the first Fourier component, i.e. $\mu_t = \arctan 2(x_{t,1}, x_{t,2})$. In other words, the tuning profile can be expanded as

$$f_i(\mu_t, \kappa_t) \sim \kappa_t \cos(\mu_t - \phi_i) + \mathcal{R}, \qquad [S48]$$

where $\mathcal{R}$ collectively denotes the orthogonal other Fourier modes. In Fig. S1g-j, we confirm the proportionality of the amplitudes of the first Fourier coefficient in κ_t numerically.

3. Details on Bayesian ring attractor dynamics and parameter tuning

In the main text we consider a rate-based network model, called the *Bayesian ring attractor*, that implements an approximation to the circKF in the dynamics of its bump position and amplitude. Here, we derive this network in two steps. First, we start with a network that implements the circKF exactly (in the limit of an infinite number of neurons) by implementing the dynamics described by Eqs. [S19] and [S20]. This network won't be a ring attractor, as its activity will decay to zero in the absence of external inputs. After that we will change the network to instead implement the quadratic approximation to the circKF by implementing the dynamics described by Eqs. [S19] and [S21], resulting in the Bayesian ring attractor described in the main text.

Our derivation starts with a general network in the limit of infinitely many neurons, continuously covering the space of preferred HDs. For this network we will analytically derive dynamics of bump position and amplitude. Matching these dynamics to that of the circKF equations then allows us to determine the network parameters required for this implementation. The network we present in the main text is formulated for a finite number of neurons, and here we will further demonstrate that it is straightforward to change between those two representations. In fact, any network coefficients for the infinite-neuron network are chosen such that they also describe those used for the finite-neuron network in the main text.

A. Network that exactly implements the circKF. Let us make an ansatz for a continuous-space, linear network dynamics with an additional non-linear interaction term:

$$\mathrm{d}r_t(\phi) = -\frac{1}{\tau} r_t(\phi)\mathrm{d}t + g\left(r_t(\phi)\right)\cdot r_t(\phi)\mathrm{d}t + \left(W * r_t\right)(\phi)\,\mathrm{d}t + I_t^{\mathrm{ext}}(\phi). \tag{S49}$$

Here, $r_t(\phi)$ denotes the activity of a neuron identified by its preferred HD ϕ at time t, and $I_t^{\mathrm{ext}}(\phi)$ is an external input. Due to the circular symmetry, the recurrent connectivity function $W(\Delta\phi)$ only depends on the relative distance $\Delta\phi$ between two neurons' preferred HD. Further, $\left(W * r_t\right)(\phi) := \frac{1}{\pi}\int d\phi' W(\phi-\phi') r_t(\phi)$ denotes a convolution.

We consider the decomposition of the activity profile $r_t(\phi)$ in terms of its Fourier modes:

$$r_t(\phi) = \frac{1}{2} r_0(t) + \sum_{k=1}^{\infty}\left(r_k^{\mathrm{even}}(t)\cos k\phi + r_k^{\mathrm{odd}}(t)\sin k\phi\right) \tag{S50}$$

$$= \frac{1}{2} r_0(t) + \sum_{k=1}^{\infty} \tilde{r}_k(t)\cos k(\phi - \Psi_k(t)). \tag{S51}$$

Note, that the Fourier coefficients $r_k^{\mathrm{even}}(t)$ and $r_k^{\mathrm{odd}}(t)$ are related to the coefficient's amplitude $\tilde{r}_k(t)$ and phase $\Psi_k(t)$ via a Cartesian to polar coordinate transformation. Taking the derivative on both sides (in the amplitude-phase form) results in:

$$\mathrm{d}r_t(\phi) = \frac{1}{2}\,\mathrm{d}r_0(t) + \sum_{k=1}^{\infty}\Big(\cos k(\phi-\Psi_k(t))\,\mathrm{d}\tilde{r}_k(t) + k\tilde{r}_k(t)\sin k(\phi-\Psi_k(t))\,\mathrm{d}\Psi_k(t)\Big). \tag{S52}$$

Thus, we can determine the dynamics of the Fourier coefficients r_0, $\tilde{r}_k$, and Ψ_k by Fourier-transforming Eq. [S49], and subsequently matching the coefficients in the Fourier modes:

$$\mathrm{d}r_0(t) = \frac{1}{\pi}\int_{-\pi}^{\pi} d\phi\,(\mathrm{d}r_t) = \left(-\frac{1}{\tau} + w_0\right) r_0(t)\,\mathrm{d}t - g(r_t)\tilde{r}_0(t)\,\mathrm{d}t + I_0^{\mathrm{ext}}(t), \tag{S53}$$

$$\begin{aligned}\mathrm{d}\tilde{r}_k(t) &= \frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{d}\phi\cos k(\phi-\Psi_k(t))\,(\mathrm{d}r_t)\\ &= \left(-\frac{1}{\tau} + w_k^{even}\right)\tilde{r}_k(t)\,\mathrm{d}t - g(r_t)\tilde{r}_k(t)\,\mathrm{d}t + I_k(t)\cos(\Phi_k(t)-\Psi_k(t))\end{aligned} \tag{S54}$$

$$\begin{aligned}\mathrm{d}\Psi_k(t) &= \frac{1}{k\tilde{r}_k(t)}\frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{d}\phi\sin k(\phi-\Psi_k(t))\,(\mathrm{d}r_t)\\ &= \frac{w_k^{odd}}{k}\,\mathrm{d}t + \frac{I_k(t)}{k\tilde{r}_k(t)}\sin(\Phi_k(t)-\Psi_k(t)),\end{aligned} \tag{S55}$$

where we used the Fourier decompositions $W(\Delta\phi) = \frac{w_0}{2} + \sum_{k=1}^{\infty}\left(w_k^{\mathrm{even}}\cos(k\Delta\phi) + w_k^{\mathrm{odd}}\sin(k\Delta\phi)\right)$ and $I_t^{\mathrm{ext}}(\phi) = \frac{I_0}{2} + \sum_{k=1}^{\infty} I_k\cos(k(\phi-\Phi_k))$. Note that here, I_k refers to the k-th Fourier amplitude of the input, and not to the modified Bessel function. Furthermore, in the main text we restrict the discussion to w_0, w_1^{even} and w_1^{odd} and denote them $w^{\mathrm{const}} \equiv w_0$, $w^{\mathrm{sym}} \equiv w_1^{\mathrm{even}}$, and $w^{\mathrm{asym}} \equiv w_1^{\mathrm{odd}}$, respectively. Setting $\Psi_1(t) = \mu_t$ and $\tilde{r}_1(t) = \kappa_t$, the dynamics of the first Fourier components in amplitude-phase form read:

$$\mathrm{d}\mu_t = w_1^{odd}\,\mathrm{d}t + I_1(t)\sin(\Phi_1(t)-\mu_t), \tag{S56}$$

$$\mathrm{d}\kappa_t = \left(-\frac{1}{\tau} + w_1^{even}\right)\kappa_t\,\mathrm{d}t - g(r_t)\kappa_t\,\mathrm{d}t + I_1(t)\cos(\Phi_1(t)-\mu_t) \tag{S57}$$

Comparing Eq. [S19] (μ_t from circKF) with Eq. [S56] and Eq. [S20] (κ_t from circKF) with Eq. [S57] allows us to determine conditions for network parameters and external input in Eq. [S49], such that the circKF is exactly implemented in the dynamics of the network's first Fourier mode:

Even recurrent connections	$w_1^{\text{even}} = 1/\tau,$
Odd recurrent connections	$w_1^{\text{odd}} = \dfrac{\kappa_v}{\kappa_\phi + \kappa_v} v_t,$
External input strength	$I_1 = \kappa_z \mathrm{d}t,$
External input phase	$\Phi_1(t) = z_t,$
Nonlinear inhibition	$g(r_t) = \dfrac{f(\kappa_t(r_t))}{2(\kappa_\phi + \kappa_v)}.$

Here, v_t denotes the (observed) angular velocity with reliability κ_v, and z_t the HD observation with reliability κ_z. The nonlinear inhibition needs to be able to compute the amplitude κ_t from the network activity $r_t(\phi)$. Note that this does not impose any conditions on network parameters which do not affect the first Fourier component dynamics, for instance, higher order recurrent interaction strengths w_k with $k \neq 1$. These can in principle be chosen freely.[‡] Note that, in this simple network, angular velocity observations modulate the first odd component of the recurrent connectivity matrix. This is biologically unrealistic, and will be addressed once we move to the multi-population network further below.

To summarize, one potential (out of many possible) network dynamics that implements the circKF in the dynamics of its first Fourier components reads:

$$\mathrm{d}r_t(\phi) = -\frac{1}{\tau}r_t(\phi)\,\mathrm{d}t - \frac{f(\kappa_t(r_t))}{2(\kappa_\phi+\kappa_v)}r_t(\phi)\mathrm{d}t + \frac{1}{\tau}\left(\cos * r_t\right)(\phi)\mathrm{d}t + \frac{\kappa_v}{\kappa_\phi+\kappa_v}v_t\left(\sin * r_t\right)(\phi)\,\mathrm{d}t + I_t^{\text{ext}}(\phi). \tag{S58}$$

Please consult Sec. D for an additional term required to account for r_t being a stochastic process. We have not included this term here, as it only becomes important in the $\mathrm{d}t \to 0$ limit, and does not contribute additional intuition about the network's operation.

B. Network with quadratic nonlinearity. While the network we have derived so far implements the cricKF exactly, its activity decays to zero in the absence of external inputs, such that it is not an attractor network. In this section we will instead use the quadratic approximation to the circKF, which will lead to the Bayesian ring attractor we discuss in the main text. To do so, we use the following nonlinearity for the inhibitory interaction:

$$g(r_t)r_t = w^{\text{quad}}(M * [r_t]_+)(\phi) \circ r_t(\phi), \tag{S59}$$

with rectification nonlinearity $[\cdot]_+$ and a constant function $M = \frac{\pi}{2}$. Here, $\circ$ denotes the Hadamar (piecewise) product. In the main text, we wrote this interaction as $g(r_t)r_t \to w^{\text{quad}}\left(\pi\sum_{i=1}^N [r_t^{(i)}]_+\right)\cdot r_t$, which is equivalent, but less technical.

We assume r_t to be dominated by its first Fourier component, such that the other orders become negligible, i.e. $r_t(\phi) = \kappa_t\cos(\phi-\mu_t) + \mathcal{R}$ with $\mathcal{R}$ small.[§] We find

$$(M * [r_t]_+)(\phi) \approx \frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{d}\phi'\frac{\pi}{2}\left[\kappa_t\cos(\phi'-\mu_t)\right]_+ = \kappa_t. \tag{S60}$$

Fourier-transforming the nonlinearity with respect to the amplitude-phase form yields:

$$\begin{aligned}\frac{1}{\pi}\int_{-\pi}^{\pi}\mathrm{d}\phi'\cos(\phi'-\mu_t)g(r_t)r_t &= \frac{w^{\text{quad}}}{\pi}\int_{-\pi}^{\pi}\mathrm{d}\phi'\,\cos(\phi'-\mu_t)\,(M*[r_t]_+)(\phi')\cdot r_t(\phi') \\ &= \frac{w^{\text{quad}}}{\pi}\kappa_t\int_{-\pi}^{\pi}\mathrm{d}\phi'\,\cos(\phi'-\mu_t)\,r_t(\phi') = w^{\text{quad}}\kappa_t^2.\end{aligned} \tag{S61}$$

Thus, the dynamics of the first Fourier amplitude of a network with this nonlinearity is given by:

[‡] Practically, we chose them such that higher-order Fourier modes and the zero-th mode decay reasonably fast, to produce a unimodal activity bump.

[§] Alternatively, we can consider additionally convolving r_t with a cosine before applying the rectification, effectively filtering out the desired mode.

$$\mathrm{d}\kappa_t = \left(-\frac{1}{\tau} + w_1^{even}\right)\kappa_t\,\mathrm{d}t - w^{\mathrm{quad}}\kappa_t^2\,\mathrm{d}t + I_1(t)\cos(\Phi_1(t) - \mu_t). \quad [S62]$$

The network parameters can be tuned such that the dynamics match that of the quadratic approximation of the circular Kalman filter (Eq. [S19] and [S21]), analogously to the previous section. This yields the following network parameters for a Bayesian ring-attractor network:

$$\begin{aligned}
\text{Even recurrent connections} \quad & w_1^{\mathrm{even}} = 1/\tau + \frac{1}{\kappa_\phi + \kappa_v}, \\
\text{Odd recurrent connections} \quad & w_1^{\mathrm{odd}} = \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t, \\
\text{External input strength} \quad & I_1 = \kappa_z \mathrm{d}t, \\
\text{External input phase} \quad & \Phi_1(t) = z_t, \\
\text{Quadratic inhibition} \quad & w^{\mathrm{quad}} = \frac{1}{\kappa_\phi + \kappa_v},
\end{aligned}$$

C. Continuous vs. discrete networks. The analysis we have presented above is valid for a continuum of neurons, i.e. $N \to \infty$, that span a continuum of preferred HDs. Formally, this implies that the difference in preferred HD between two 'neighboring' neurons converges to zero, $\Delta\phi := \phi_i - \phi_j = \frac{2\pi}{N} \to 0$. In the text and for our simulations, we used a discretized network, where we assumed the preferred HDs of the neurons to be equally spaced, but finite.

It is straightforward to go back and forth between these two representations (cf. (9)): in a discretized network, $\mathbf{r}_t$ denotes a vector of neural activities, indexed by their preferred HD ϕ_i, which becomes a function $r_t(\phi)$ for a continuous network. Likewise, connectivity matrices W become functions with two arguments $W(\phi_i, \phi_j)$, and matrix multiplications become integrals. The circular symmetry of HD implies that the entries of a connectivity matrix only depend on the relative distance between two neurons, and not on absolute position, such that for a connectivity matrix W we can write $W_{ij} = W(\phi_i, \phi_j) = W(\phi_i - \phi_j)$. Thus, we can write matrix multiplications as convolutions (assuming the vectors and matrix are ordered with respect to their preferred HD):

$$(W \cdot \mathbf{r}_t)_i = \sum_{j=1}^{N} W_{ij} r_{t,j} = \frac{N}{2\pi} \sum_{j=1}^{N} W_{ij} r_{t,j} \Delta\phi \quad [S63]$$

$$\overset{N\to\infty, \Delta\phi\to 0}{\to} \frac{N}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\phi' W(\phi, \phi') r_t(\phi') = \frac{N}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\phi' W(\phi - \phi') r_t(\phi') = \frac{N}{2} \left(W * r_t\right)(\phi). \quad [S64]$$

where we defined the convolution as above. Thus, to ensure consistency between the coefficients of the matrices used in the main text and the coefficients of the connectivity functions we used in our analysis in the SI, we scaled the connectivity matrices in the main text by a factor $\frac{2}{N}$.

D. Stochastic correction [Technical]. The derivation in the previous section did not take into account that due to the dependence on the angular velocity observations v_t, the phase $\Psi_k(t)$ is actually an Itô stochastic process, and hence the network activity r_t is, too. Thus, when performing a change of variables, such as the expansion Eq. [S52], we have to use Itô's lemma (Eq. [S24]), and expand up to second order in $\Psi_k(t)$ (we have seen that the dynamics of the amplitude $\tilde{r}_k(t)$ is independent of v_t, and thus only carries first order terms):

$$\mathrm{d}r_t(\phi) = \mathrm{d}\left(\frac{1}{2} r_0(t) + \sum_{k=1}^{\infty} \tilde{r}_k(t) \cos k(\phi - \Psi_k(t))\right) \quad [S65]$$

$$\begin{aligned}
&= \frac{1}{2}\mathrm{d}r_0(t) + \sum_{k=1}^{\infty} \Big(\cos k(\phi - \Psi_k(t))\mathrm{d}\tilde{r}_k(t) + k\tilde{r}_k(t) \sin k(\phi - \Psi_k(t))\mathrm{d}\Psi_k(t) \\
&\qquad - \frac{1}{2} k^2 \tilde{r}_k(t) \cos k(\phi - \Psi_k(t))(\mathrm{d}\Psi_k(t))^2 \Big),
\end{aligned} \quad [S66]$$

This implies that, if we take the effect of stochastic processes into account, comparing the Fourier coefficients in amplitude-phase form will not single out the dynamics of the amplitude $d\tilde{r}_k$, because there are now two terms proportional to $\cos k(\phi - \Psi_k(t))$. Fortunately, the problem can be solved "backwards" using the analogy to coordinate transforms in Section C, thereby restricting ourselves to the first Fourier mode (higher modes are analogous): First, we perform the Fourier transform of the dynamics in Cartesian coordinates, i.e., with respect to $\cos(\phi)$ and $\sin(\phi)$. We then note that changing this into amplitude-phase form is mathematically equivalent to a coordinate transform between the natural parameters of the von Mises distribution and the μ,κ-parametrization. Next, we require that such a coordinate transform ought to result in the dynamics for μ_t and κ_t in Eq. [S56] and [S57]. Using the analogy to Section C, we find that an additional decay term $-\frac{1}{2}\frac{\kappa_v/\kappa_\phi}{\kappa_v+\kappa_\phi} r_t(\phi)\mathrm{d}t$ is needed in the

network dynamics, which implements the Itô correction on the level of the natural parameters (cf. Eq. [S26]). Apart from this additional decay, the conditions on the other network parameters remains unchanged.

This stochastic correction is not strictly needed to gain intuition about the theory, and if anything, the use of continuous-time stochastic calculus seems to make things *less* intuitive. Practically, we used an additional decay term in Eq. [S58] whenever the angular velocity observations were drawn from the true generative model and the time step $\mathrm{d}t$ was small enough to justify the notion of "continuous time", which was the case for all our simulations.

4. Details on *Drosophila*-like network

Relying on large-scale connectomics data of the *Drosophila* HD system (10, 11), we now ask if a Bayesian ring attractor can be implemented in a network that obeys biological network connectivity constraints. Here we show how the motifs of this network – and, by extension, any biological ring attractor network – could potentially implement dynamic Bayesian inference.

A. Connectivity motifs in the *Drosophila* HD system connectome. The ring attractor in the *Drosophila* HD system is composed of three core cell types, called EPG, PEN1 and Δ7 neurons (10–12), cf. Fig. 4A,B. HD is represented as a bump of neural activity in the EPG population (13). These neurons are recurrently connected with excitatory PEN1 neurons. When the fly turns, this differentially activates PEN1 neurons in the right and left brain hemispheres, and because PEN1 neurons have asymmetric (shifted) projections back to EPG neurons, they can rotate the bump of EPG activity in accordance with the fly's rotation (14, 15). This motif effectively establishes the velocity-modulated odd recurrent connectivity required to initiate turns in ring attractor networks (Fig. 4D). Moreover, EPG neurons are recurrently connected with inhibitory Δ7 neurons, which establishes broad inhibition (Fig. 4E). Finally, EPG neurons receive inhibitory inputs from so-called ER neurons, which send HD information to EPG neurons (16–18) (Fig. 4F). In summary, the fly's HD system is equipped with the basic motifs to implement a Bayesian ring attractor.

B. A multi-network model mimicking the *Drosophila* HD system connectome. The main idea of the idealized network in the previous section was to tune the network parameters such that the circKF (or the quadratic approximation of the circKF) was implemented in the coefficients of the first Fourier mode. Here, we will use the connectome of the fruit fly *Drosophila* (10) to build a recurrent neural network, and show that the quadratic approximation of the circKF can be implemented in such an architecture by determining the coefficients analogously. Thereby, we first approximate the connectivity matrices describing this connectome (Fig. 4B) by analytically accessible functions, which nonetheless retain the main features of this connectivity (as outlined, e.g., in (12)), and preserve the motifs that implement the ring-attractor in the *Drosophila* HD system (see review in (19), cf. Fig. 4C). We in turn analytically determine the conditions for the coefficients of the connectivities between (rather than within) the different network populations, such that the dynamics of the first Fourier components match that of the quadratic approximation of the circKF.

Specifically, we consider five neuronal populations: an HD population, r^{HD}, which we designed to track HD estimate and certainty with its bump parameter dynamics, two angular (AV^+ and AV^-) velocity populations, r^{AV^+} and r^{AV^-}, which are tuned to head direction and are differentially modulated by angular velocity input, an inhibitory (INH) population, r^{INH}, and a population I^{ext} that represents external input, that is, the HD observations. As before, the population activities $r(\phi)$ are functions of preferred HDs, ϕ, but we will drop the argument ϕ to keep the notation uncluttered.

We start with the following ansatz for a network dynamics:

$$\begin{aligned}\mathrm{d}r_t^{HD} =& -\frac{1}{\tau_{HD}} r_t^{HD}\,\mathrm{d}t + W_{HD\leftarrow HD} * r_t^{HD}\,\mathrm{d}t + W_{HD\leftarrow AV^+} * r_t^{AV^+} + W_{HD\leftarrow AV^-} * r_t^{AV^-}\,\mathrm{d}t \\ &+ \left(W_{HD\leftarrow INH} * [r_t^{INH}]_+\right) \circ r_t^{HD}\,\mathrm{d}t + I_t^{ext},\end{aligned} \tag{S67}$$

$$\mathrm{d}r_t^{AV^+} = \frac{1}{\tau_{AV^+}}\left(-r_t^{AV^+} + (o^{AV} + v_t) W_{AV^+\leftarrow HD} * r_t^{HD}\right)\mathrm{d}t, \tag{S68}$$

$$\mathrm{d}r_t^{AV^-} = \frac{1}{\tau_{AV^-}}\left(-r_t^{AV^-} + (o^{AV} - v_t) W_{AV^-\leftarrow HD} * r_t^{HD}\right)\mathrm{d}t, \tag{S69}$$

$$\mathrm{d}r_t^{INH} = \frac{1}{\tau_{INH}}\left(-r_t^{INH} + W_{INH\leftarrow HD} * [r_t^{HD}]_+ + W_{INH\leftarrow INH} * r_t^{INH}\right)\mathrm{d}t. \tag{S70}$$

From the connectivity profile ((10), cf. Fig. 4B), we make the following ansatz for the connectivity functions (which results in Fig. 4C):

$$W_{HD\leftarrow HD}(\Delta\phi) = c_0^{HD} + c_1^{HD}\left[\cos\Delta\phi\right], \tag{S71}$$

$$W_{AV^\pm\leftarrow HD}(\Delta\phi) = c^{AV^\pm\leftarrow HD}\delta(\Delta\phi), \tag{S72}$$

$$W_{HD\leftarrow AV^\pm}(\Delta\phi) = c^{HD\leftarrow AV^\pm}\left[\sin\left(\Delta\phi \pm \frac{\pi}{4}\right)\right]_+, \tag{S73}$$

$$W_{INH\leftarrow HD}(\Delta\phi) = \frac{c_0^{INH\leftarrow HD}}{2} + c_1^{INH\leftarrow HD}\cos(\Delta\phi), \tag{S74}$$

$$W_{INH\leftarrow INH} = \frac{c_0^{INH\leftarrow INH}}{2} + c_1^{INH\leftarrow INH}\cos(\Delta\phi), \tag{S75}$$

$$W_{HD\leftarrow INH}(\Delta\phi) = c^{HD\leftarrow INH}\delta(\Delta\phi). \tag{S76}$$

In what follows, we will derive the conditions for the connection strengths in this ansatz that allow an implementation of the quadratic approximation of the circKF in the dynamics of the first Fourier component. Thereby, we make the assumption that the leading order of the HD population activity r_t^{HD} is a cosine, i.e. $r_t^{HD}(\phi) = \frac{r_0^{HD}(t)}{2} + \kappa_t\cos(\phi - \mu_t) + \mathcal{R}$, and that higher-order Fourier modes are negligible. We further assume that the time constants of the $\mathrm{AV}^\pm$ and INH populations, $\tau_{AV^\pm}$ and τ_{INH}, are much smaller than τ_{HD} of the HD population, which allows us to assume that the activity in those populations is stationary.

B.1. AV$^{\pm}$ population. As described above, the integration of turning signals in the fruit fly is modulated through differential activation of PEN1 neurons (our AV$^{\pm}$ population) in the right and left brain hemispheres that asymmetrically project back to EPG neurons (our HD population) (14, 15). This motif implements the effective asymmetric angular velocity-dependent recurrent connectivity that is needed to rotate the activity in ring-attractor networks (20, 21). Thus, we will tune the parameters in the HD→AV$^{\pm}$ →HD circuit such that the resulting effective odd recurrent connectivity contribution w_1^{odd} (i.e., that proportional to $\sin(\phi - \mu_t)$ implements the turn in the activity profile due to angular velocity integration, cf. Eq. [S55]).

As a first step, we compute the activities in the AV$^{\pm}$ populations. It is straightforward to check that, if the time constant $\tau_{AV} \ll \tau_{HD}$, the activity in the AV populations can be described by its stationary activity at every point in time:

$$r_t^{AV^{\pm}} = (o_{AV} \pm v_t) W_{AV^{\pm}\leftarrow HD} * r_t^{HD} = c^{AV^{\pm}\leftarrow HD}(o_{AV} \pm v_t)\frac{1}{\pi}\int_{-\pi}^{\pi} \mathrm{d}\phi' \delta(\phi - \phi') r_t^{HD}(\phi') \\ = c^{AV^{\pm}\leftarrow HD}(o_{AV} \pm v_t)\, r_t^{HD}. \quad [S77]$$

Expanding the connectivity function from the HD to the AV$^{\pm}$ populations in a Fourier series yields:

$$W_{HD\leftarrow AV^{\pm}} = c^{HD\leftarrow AV^{\pm}}\left[\pm\sin(\Delta\phi \pm \frac{\pi}{4})\right]_+ = c^{HD\leftarrow AV^{\pm}}\left(\frac{1}{\pi} + \frac{1}{2\sqrt{2}}\cos(\Delta\phi) \pm \frac{1}{2\sqrt{2}}\sin(\Delta\phi)\right) + \mathcal{R}, \quad [S78]$$

allowing us to compute the effective recurrent contributions in the HD population that is mediated via this network motif:

$$W_{HD\leftarrow AV^{+}} * r_t^{AV^{+}} = c^{HD\leftarrow AV^{\pm}} c^{AV^{\pm}\leftarrow HD}(o_{AV} + v_t)\frac{1}{\pi}\int_{-\pi}^{\pi} \mathrm{d}\phi' \left(\frac{1}{\pi} + \frac{1}{2\sqrt{2}}\cos(\phi - \phi') + \frac{1}{2\sqrt{2}}\sin(\phi - \phi') + \mathcal{R}\right) r_t^{HD}(\phi') \\ = c^{HD\leftarrow AV^{\pm}} c^{AV^{\pm}\leftarrow HD}(o_{AV} + v_t)\left(\frac{r_0^{HD}}{\pi} + \frac{\kappa_t}{2\sqrt{2}}\cos(\phi - \mu_t) + \frac{\kappa_t}{2\sqrt{2}}\sin(\phi - \mu_t)\right), \quad [S79]$$

$$W_{HD\leftarrow AV^{-}} * r_t^{AV^{-}} = c^{HD\leftarrow AV^{\pm}} c^{AV^{\pm}\leftarrow HD}(o_{AV} - v_t)\left(\frac{r_0^{HD}}{\pi} + \frac{\kappa_t}{2\sqrt{2}}\cos(\phi - \mu_t) - \frac{\kappa_t}{2\sqrt{2}}\sin(\phi - \mu_t)\right), \quad [S80]$$

and thus

$$W_{HD\leftarrow AV^{+}} * r_t^{AV^{+}} + W_{HD\leftarrow AV^{-}} * r_t^{AV^{-}} \\ = c^{HD\leftarrow AV^{\pm}} c^{AV^{\pm}\leftarrow HD}\left(2\frac{o_{AV}}{\pi} r_0^{HD} + \kappa_t \frac{o_{AV}}{\sqrt{2}}\cos(\phi - \mu_t) + \kappa_t v_t \frac{1}{\sqrt{2}}\sin(\phi - \mu_t)\right). \quad [S81]$$

Thus, this motif implements an effective odd recurrent connectivity with $w_1^{odd} = \frac{c^{HD\leftarrow AV^{\pm}} c^{AV^{\pm}\leftarrow HD}}{\sqrt{2}} v_t$. We require that the effective odd recurrent connectivity is the same as in the Bayesian ring attractor, that is,

$$w_1^{\mathrm{odd}} = \frac{c^{HD\leftarrow AV^{\pm}} c^{AV^{\pm}\leftarrow HD}}{\sqrt{2}} v_t \stackrel{!}{=} \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t, \quad [S82]$$

and thus the condition for the coefficients reads:

$$c^{HD\leftarrow AV^{\pm}} = \frac{\sqrt{2}}{c^{AV^{\pm}\leftarrow HD}} \frac{\kappa_v}{\kappa_\phi + \kappa_v}. \quad [S83]$$

Interestingly, due to the offset o_{AV} we also obtain a recurrent contribution to the activity baseline $r_0(t)$, and a contribution to the *even* first order recurrent connectivity,

$$w_1^{\mathrm{even,\ AV}} = c^{HD\leftarrow AV^{\pm}} c^{AV^{\pm}\leftarrow HD} \frac{o_{AV}}{\sqrt{2}} \quad [S84]$$

$$= \frac{\kappa_v}{\kappa_\phi + \kappa_v} o_{AV}.. \quad [S85]$$

We will return to this when computing the recurrent connectivities within the HD populations below.

B.2. INH population.. In our network, the recurrent interaction with the INH population implements the quadratic inhibition. In the same way we tracked the effective odd recurrent through the AV$^{\pm}$ recurrent loop, we will here determine the effective quadratic interaction strength w^{quad} as a function of the network parameters, and then tune it in order to implement the quadratic approximation of the circKF.

To determine the activity in the INH population, we first expand $\left[r_t^{HD}\right]_+$ in its Fourier series:

$$\begin{aligned}\left[r_t^{HD}\right]_+ &\approx \left[\frac{r_0^{HD}}{2} + \kappa_t \cos(\phi - \mu_t)\right]_+ \\ &\approx \frac{r_0^{HD}\phi_c}{2\pi} + \frac{\kappa_t}{\pi}\sin\phi_c + \left(\frac{\kappa_t}{\pi}\phi_c + \frac{r_0^{HD}}{2\pi}\sin\phi_c\right)\cos(\phi - \mu_t) + \mathcal{R} \\ &\approx \frac{r_0^{HD}}{4} + \frac{\kappa_t}{\pi} + \left(\frac{\kappa_t}{2} + \frac{r_0^{HD}}{\pi}\right)\cos(\phi - \mu_t) + \mathcal{R}, \end{aligned} \tag{S86}$$

with cutoff angle $\phi_c = \arccos\left(-\frac{\kappa_t}{2r_1^{HD}}\right) \approx \frac{\pi}{2} + \frac{r_0^{HD}}{2\kappa_t}$ for $\kappa_t \gg r_0^{HD}/2$. With the dynamics of the INH population in Eqs. [S70], and the connectivity functions in [S74] and [S75], we can write down the dynamics of the first two Fourier coefficients in the INH population:

$$\tau_{INH}\,\mathrm{d}r_0^{INH} = \left(-r_0^{INH} + \left(\frac{r_0^{HD}}{2} + \frac{2}{\pi}\kappa_t\right)c_0^{INH\leftarrow HD} + c_0^{INH\leftarrow INH} r_0^{INH}\right)\mathrm{d}t, \tag{S87}$$

$$\tau_{INH}\,\mathrm{d}r_1^{INH} = \left(-r_1^{INH} + \left(\frac{1}{2}\kappa_t + \frac{1}{\pi}r_0^{HD}\right)c_1^{INH\leftarrow HD} + c_1^{INH\leftarrow INH} r_1^{INH}\right)\mathrm{d}t. \tag{S88}$$

Assuming again that the dynamics in the INH population is much faster than in the HD population, $\tau_{INH} \ll \tau_{HD}$, we can write down the stationary activities of the activity profile in the INH population:

$$r_0^{INH} = \frac{\frac{r_0^{HD}}{2} + \frac{2}{\pi}\kappa_t}{1 - c_0^{INH\leftarrow INH}} c_0^{INH\leftarrow HD}, \tag{S89}$$

$$r_1^{INH} = \frac{\frac{1}{2}\kappa_t + \frac{1}{\pi}r_0^{HD}}{1 - c_1^{INH\leftarrow INH}} c_1^{INH\leftarrow HD}. \tag{S90}$$

Plugging this into Eq. [S67], we obtain the change in the amplitude of the first Fourier mode through the interaction with the INH population:

$$\begin{aligned}\left(W_{HD\leftarrow INH} * [r_t^{INH}]_+\right)\cdot r_t^{HD} &= c^{HD\leftarrow INH}\left(\frac{r_0^{INH}}{2} + r_1^{INH}\cos(\phi - \mu_t)\right)\cdot r_t^{HD}(\phi) \\ &= c^{HD\leftarrow INH}\left(\frac{r_0^{INH}}{2}\kappa_t + r_1^{INH}\frac{r_0^{HD}}{2}\right)\cos(\phi - \mu_t) + \mathcal{R} \\ &= c^{HD\leftarrow INH}\Bigg[\frac{c_0^{INH\leftarrow HD}}{\pi(1 - c_0^{INH\leftarrow INH})}\kappa_t^2 + \left(\frac{c_0^{INH\leftarrow HD}}{1 - c_0^{INH\leftarrow INH}} + \frac{c_1^{INH\leftarrow HD}}{1 - c_1^{INH\leftarrow INH}}\right)\frac{r_0^{HD}}{4}\kappa_t \\ &\quad + \frac{c_1^{INH\leftarrow HD}}{\pi(1 - c_1^{INH\leftarrow INH})}\frac{(r_0^{HD})^2}{2}\Bigg]\cos(\phi - \mu_t). \end{aligned} \tag{S91}$$

The first term on the right hand side has our desired quadratic interaction. It matches that of the quadratic approximation of the circKF $w^{quad} = 1/(\kappa_\phi + \kappa_v)$, if the following condition is fulfilled:

$$c^{HD\leftarrow INH} = -\frac{1}{\kappa_\phi + \kappa_v}\frac{\pi(1 - c_0^{INH})}{c_0^{INH\leftarrow HD}}. \tag{S92}$$

The other terms in Eq. [S91] are "nuisance" terms, which, if too large, may significantly interfere with the inference dynamics. However, if r_0^{HD} is small compared to κ_t, which we confirmed in simulations to be generally the case, the effect of the nuisance terms is negligible. This can further be stabilized by choosing $|c_1^{INH\leftarrow HD}| \ll |1 - c_1^{INH\leftarrow INH}|$. Interestingly, this implies that certainty κ_t mainly governs the activity in the *zero-th* order of the INH activity (Eq. [S89]).

B.3. Recurrent excitation within HD population. In the same spirit as above, here we compute the effective even recurrent connectivity of the network in order to match it with recurrent interaction w_1^{even} in the network implementation of the circKF. Starting from the Fourier expansion of the recurrent connectivity,

$$W_{HD\leftarrow HD} = c_0^{HD} + c_1^{HD}\left[\cos(\Delta\phi)\right]_+ \approx c_0^{HD} + \frac{c_1^{HD}}{\pi} + \frac{c_1^{HD}}{2}\cos(\Delta\phi) + \mathcal{R}, \tag{S93}$$

we determine the change in activity due to the recurrent interaction within the HD population:

$$W_{HD\leftarrow HD} * r_t^{HD} = \left(c_0^{HD} + \frac{c_1^{HD}}{\pi}\right)r_0^{HD} + \frac{c_1^{HD}}{2}\kappa_t\cos(\phi - \mu_t) + \mathcal{R}. \tag{S94}$$

Recall that the interaction with the AV$^{\pm}$ populations also induced an effective *even* recurrent connectivity (Eq. [S85]), such that the overall even recurrent connectivity in the network is given by,

$$w_1^{\text{even}} = w_1^{\text{even, HD}} + w_1^{\text{even, AV}} = \frac{c_1^{HD}}{2} + \frac{\kappa_v}{\kappa_\phi + \kappa_v} o_{AV} \overset{!}{=} \frac{1}{\tau} + \frac{1}{\kappa_\phi + \kappa_v}. \quad [S95]$$

This defines the following condition for the recurrent interaction within the HD population:

$$c_1^{HD} = 2\left(\frac{1}{\tau} + \frac{1}{\kappa_\phi + \kappa_v} - \frac{\kappa_v}{\kappa_\phi + \kappa_v} o_{AV}\right). \quad [S96]$$

The zero-order contribution in Eq. [S94] multiplying c_1^{HD} is significant, and exceeds the first-order interaction in magnitude, which makes the network unstable. We thus require a negative constant recurrent connectivity to balance this zero-order contribution, chosen such that this contributions in the dynamics of r_0^{HD} decays over time:

$$2\left(c_0^{HD} + \frac{c_1^{HD}}{\pi}\right) \overset{!}{<} \frac{1}{\tau}, \quad [S97]$$

and thus we arrive at our final condition:

$$c_0^{HD} < \frac{1}{2\tau} - \frac{c_1^{HD}}{\pi}. \quad [S98]$$

B.4. Summary of network connectivities. To summarize, we analytically determined that the following connectivity matrices in the network dynamics in Eq. [S67]-[S70] implement the quadratic approximation of the circKF in the HD population. As a reminder, these network dynamics are:

$$\begin{aligned}
\mathrm{d}r_t^{HD} =& -\frac{1}{\tau_{HD}} r_t^{HD}\,\mathrm{d}t + W_{HD\leftarrow HD} * r_t^{HD}\,\mathrm{d}t + W_{HD\leftarrow AV^+} * r_t^{AV^+} + W_{HD\leftarrow AV^-} * r_t^{AV^-}\,\mathrm{d}t \\
&+ \left(W_{HD\leftarrow INH} * [r_t^{INH}]_+\right) \circ r_t^{HD}\,\mathrm{d}t + I_t^{ext}, \\
\mathrm{d}r_t^{AV^+} =& \frac{1}{\tau_{AV^+}}\left(-r_t^{AV^+} + (o^{AV} + v_t) W_{AV^+\leftarrow HD} * r_t^{HD}\right)\mathrm{d}t \\
\mathrm{d}r_t^{AV^-} =& \frac{1}{\tau_{AV^-}}\left(-r_t^{AV^-} + (o^{AV} - v_t) W_{AV^-\leftarrow HD} * r_t^{HD}\right)\mathrm{d}t \\
\mathrm{d}r_t^{INH} =& \frac{1}{\tau_{INH}}\left(-r_t^{INH} + W_{INH\leftarrow HD} * [r_t^{HD}]_+ + W_{INH\leftarrow INH} * r_t^{INH}\right)\mathrm{d}t.
\end{aligned}$$

Recurrent excitation within HD population:

$$\begin{aligned}
(W_{HD\leftarrow HD})_{ij} &= \frac{2}{N_{HD}}\left(c_0^{HD} + c_1^{HD}\left[\cos\left(\phi_i^{HD} - \phi_j^{HD}\right)\right]_+\right), \\
\text{with}\quad c_1^{HD} &= 2\left(\frac{1}{\kappa_\phi + \kappa_v} + \frac{1}{\tau_{HD}} - o^{AV}\frac{\kappa_v}{\kappa_\phi + \kappa_v}\right),\quad c_0^{HD} < \frac{1}{2\tau} - \frac{c_1^{HD}}{\pi}.
\end{aligned} \quad [S99]$$

Recurrent excitation between HD and AV+ and AV- populations:

$$(W_{AV^\pm\leftarrow HD})_{ij} = c^{AV^\pm\leftarrow HD}\delta_{ij}, \quad [S100]$$

$$\begin{aligned}
(W_{HD\leftarrow AV^\pm})_{ij} &= \frac{2}{N_{AV^\pm}} c^{HD\leftarrow AV^\pm}\left[\sin\left(\phi_i^{HD} - \phi_j^{AV^\pm} \pm \frac{\pi}{4}\right)\right]_+, \\
\text{with}\quad c^{HD\leftarrow AV^\pm} &= \frac{\sqrt{2}}{c^{AV^\pm\leftarrow HD}}\frac{\kappa_v}{\kappa_\phi + \kappa_v}.
\end{aligned} \quad [S101]$$

Recurrent inhibition between HD and INH populations:

$$(W_{INH\leftarrow HD})_{ij} = \frac{2}{N_{HD}}\left(\frac{c_0^{INH\leftarrow HD}}{2} + c_1^{INH\leftarrow HD}\cos\left(\phi_i^{INH} - \phi_j^{HD}\right)\right), \quad \text{[S102]}$$

$$(W_{INH\leftarrow INH})_{ij} = \frac{2}{N_{INH}}\left(\frac{c_0^{INH}}{2} + c_1^{INH}\cos\left(\phi_i^{INH} - \phi_j^{HD}\right)\right), \quad \text{[S103]}$$

$$\text{with} \quad |c_1^{INH\leftarrow HD}| << |1 - c_1^{INH}|$$

$$(W_{HD\leftarrow INH})_{ij} = c^{HD\leftarrow INH}\delta_{ij},$$

$$\text{with} \quad c^{HD\leftarrow INH} = -\frac{1}{\kappa_\phi + \kappa_v}\frac{\pi(1 - c_0^{INH})}{c_0^{INH\leftarrow HD}}. \quad \text{[S104]}$$

Activities of the EXT population were assumed to give rise to a bump-shaped inhibitory input opposite of the HD observation, loosely related to how ring neurons mediate such input to the EPG neurons (17, 18). We thus modeled this bump-shaped input to the HD population directly without explicitly representing a dynamics of the EXT population.

External input:

$$I_{i,t}^{ext} = -2\kappa_z \mathrm{d}t\left[\cos(\phi_i^{HD} - z_t + \pi)\right]_+. \quad \text{[S105]}$$

The network dynamics still has a considerable number of degrees of freedom. That is, the baseline o^{AV}, network connectivity strengths $c^{AV^\pm\leftarrow HD}$, $c_0^{INH\leftarrow HD}$, $c_1^{INH\leftarrow HD}$, c_0^{INH}, c_1^{INH}, and time scales τ_{HD}, τ_{AV^+}, τ_{AV^-} and τ_{INH} can essentially be chosen freely. If the number of neurons N differs between populations, the δ_{ij}'s can be replaced by a normalized, Gaussian-shaped kernel with a finite width. For our analytical results to hold, we require $\tau_{HD} \gg \tau_{AV^+}, \tau_{AV^-}, \tau_{INH}$. We further constrained the network by choosing $c_0^{INH\leftarrow HD} > 0$, $c_1^{INH\leftarrow HD} \leq 0$ and $|c_0^{INH\leftarrow HD}| > |c_1^{INH\leftarrow HD}|$, which leads to the broad excitatory input into the INH population, and the formation of an 'antibump', similarly to the one observed in $\Delta 7$ neurons (12).

C. Drosophila-like network simulations and HD tracking performance. To demonstrate that the multi-population network can indeed implement the quadratic approximation to the circKF, we measured its HD tracking performance and compared it to the circKF and the Bayesian ring attractor.

We used the following parameters in the associated network simulations (Fig. 4G,H): $\kappa_v = 5$, $T = 20$, $\Delta t = 0.001$, results are averages over $P = 5'000$ simulations. Network architecture followed the full network in Eqs. [S67]-[S70], with baseline $o^{AV} = 0$, time scales $\tau_{HD} = 0.1$, $\tau_{AV^+} = \tau_{AV^-} = 0.01$, $\tau_{INH} = 0.001$, connection strengths $c_0^{HD} = -0.2$, $c_1^{HD} = 0$, $c^{AV^\pm\leftarrow HD} = 1$, $c_0^{INH\leftarrow HD} = 0.5$, $c_1^{INH\leftarrow HD} = -0.5$, $c_0^{INH} = 0.1$, $c_1^{INH} = 0$. Further, in the discretized dynamics we chose $N_{HD} = 100$, $N_{AV^+} = 50$, $N_{AV^-} = 50$, $N_{INH} = 100$, and $N_{EXT} = 100$.

As shown in Fig. 4G,H, the network simulations confirmed that this network indeed achieves a HD tracking performance indistinguishable to that of our idealized Bayesian ring attractor network. Thus, even when we add the constraints dictated by the actual connectivity patterns of neural networks in the brain, the resulting network is still able to implement dynamic Bayesian inference.

5. The impact of neural noise on inference dynamics

So far we have assumed the the only sources of noise were noisy inputs from angular velocity and HD observations. Here we ask how the inference dynamics are impacted if the neurons that constitute the ring attractor are also noisy. We will do so in three steps. First, we will make a qualitative observation of how such neural noise is expected to impact the dynamics of μ_t and κ_t. Second, we will derive expressions for the impact of such noise on μ_t and κ_t for different noise models. Third, we will ask how we can ensure that neural noise has a minimal impact on the performed inference. For all steps we return to our single-population ring attractor whose dynamics are described by Eq. [S49], and assume that neural noise impacts the activity of neuron j by

$$\mathrm{d}r_{t,j} = h\left(r_{t,j}\right)\mathrm{d}W_{t,j}, \tag{S106}$$

where $h(\cdot)$ is a function of neural activity, and the $\mathrm{d}W_{t,j}$'s are Brownian motion processes that are uncorrelated across neurons. Different noise models correspond to different assumptions about the form of $h(\cdot)$. As for large population sizes N, individual neural noise can be averaged out and will have limited impact (22). Therefore, we assume N to be sufficiently small for neural noise to matter, but to be sufficiently large such that we can well-approximate various sums by their integral limit.

A. The qualitative impact of neural noise on inference dynamics. With neural noise, the population dynamics equation Eq. [S49] becomes

$$\mathrm{d}r_t(\phi) = \cdots + I_t^{ext}(\phi) + \eta_t(\phi), \tag{S107}$$

where $I_t^{ext}(\phi)$ is our model's (stochastic) external input, and the newly added $\eta_t(\phi)$ captures the activity perturbations induced by neural noise. This shows that we can interpret neural noise as yet another stochastic input to the network. This implies that this noise impacts the dynamics for η_t and κ_t (previously Eqs. [S56] & [S57]) through

$$\mathrm{d}\mu_t = \cdots + I_1(t)\sin\left(\Phi_1(t) - \mu_t\right) + \eta_1(t)\sin\left(\xi_1(t) - \mu_t\right), \tag{S108}$$

$$\mathrm{d}\kappa_t = \cdots + I_1(t)\cos\left(\Phi_1(t) - \mu_t\right) + \eta_1(t)\cos\left(\xi_1(t) - \mu_t\right), \tag{S109}$$

where $I_1(t)$ and $\Phi_1(t)$ are amplitude and phase of the first Fourier component of $I_t^{ext}(\phi)$, and $\eta_1(t)$ and $\xi_1(t)$ are the analogous quantities for the neural noise $\eta_t(\phi)$. As this noise is uniform on the circle, its phase is also uniform on the circle, and its amplitude is roughly constant (for some fixed N). This implies that both $\eta_1(t)\sin\left(\xi_1(t) - \mu_t\right)$ and $\eta_1\cos\left(\xi_1(t) - \mu_t\right)$ will have the same variance. Crucially, the HD estimate μ_t is by Eq. [S56] formed by integrating all of its terms, such that the added noise term results in a diffusion of this estimate (22). The certainty κ_t, in contrast, by Eq. [S57] performs a leaky integration of its term, such that it low-pass filters the noise — it somewhat perturbs κ_t, but its contribution will be bounded.

B. How neural noise quantitatively impacts the dynamics of μ_t and κ_t. To get a better quantitative understanding of the impact of neural noise, we here derive expressions for its impact on μ_t and κ_t for different noise models. First, we will assess the impact of the generic noise model, Eq. [S106] on the posterior parameters, x_1 and x_2, in their Cartesian form. Second, we will translate this impact to polar coordinates, μ and κ. Third, we will consider three different noise models to see how those impact the dynamics of μ and κ. To simplify notation we assume some fixed time t, and leave the $\cdot_t$ subscript implicit.

B.1. The impact of neural noise on x_1 and x_2. For finite N, x_1 and x_2 are computed as

$$x_1 = \frac{2}{N}\sum_{j=1}^{N}\cos\left(\phi_j\right)r_j, \qquad x_2 = \frac{2}{N}\sum_{j=1}^{N}\sin\left(\phi_j\right)r_j, \tag{S110}$$

where ϕ_j is the preferred HD of neuron j, and where the $2/N$ pre-factor ensures appropriate normalization. The generic neural noise model, Eq. [S106], thus leads to

$$\mathrm{d}x_1 = \frac{2}{N}\sum_{j=1}^{N}\cos\left(\phi_j\right)h\left(r_j\right)\mathrm{d}W_j, \qquad \mathrm{d}x_2 = \frac{2}{N}\sum_{j=1}^{N}\sin\left(\phi_j\right)h\left(r_j\right)\mathrm{d}W_j, \tag{S111}$$

independent of the current population activity $\boldsymbol{r}$ (except through $h(r_j)$). It can be shown that $\langle \mathrm{d}x_i\rangle = 0$ for $i \in \{1, 2\}$, and that

$$\mathrm{cov}\left(\mathrm{d}\boldsymbol{x}\right) = \frac{4}{N^2}\begin{pmatrix} \boldsymbol{c}^{2^T}\boldsymbol{h}^2 & \boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s} \\ \boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s} & \boldsymbol{s}^{2^T}\boldsymbol{h}^2 \end{pmatrix}\mathrm{d}t, \tag{S112}$$

where we have defined the N-element vectors $\boldsymbol{c}$, $\boldsymbol{s}$, and $\boldsymbol{h}$ with elements $c_j = \cos\left(\phi_j\right)$, $s_j = \sin\left(\phi_j\right)$, and $h_j = h\left(r_j\right)$, where the $\cdot^2$'s are element-wise, and where $\mathrm{diag}\left(\boldsymbol{h}^2\right)$ denotes a diagonal matrix with diagonal $\boldsymbol{h}^2$. Thus, the noise-induced evolution of $\boldsymbol{x}$ is described by the two-dimensional process

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{G}\mathrm{d}\boldsymbol{W}, \tag{S113}$$

with $\boldsymbol{G}$ given by

$$\boldsymbol{G} = \frac{2}{N\sqrt{\boldsymbol{c}^{2^T}\boldsymbol{h}^2}}\begin{pmatrix} \boldsymbol{c}^{2^T}\boldsymbol{h}^2 & 0 \\ \boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s} & \sqrt{\boldsymbol{s}^{2^T}\boldsymbol{h}^2\boldsymbol{c}^{2^T}\boldsymbol{h}^2 - \left(\boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s}\right)^2} \end{pmatrix}, \tag{S114}$$

such that $\mathrm{cov}\left(\mathrm{d}\boldsymbol{x}\right) = \boldsymbol{G}\boldsymbol{G}^T\mathrm{d}t$. Overall, this shows that neural noise will not cause a drift of $\boldsymbol{x}$ but will introduce (potentially) correlated noise in both x_1 and x_2.

B.2. The impact of neural noise on μ and κ. To translate the impact of neural noise from natural parameters $\boldsymbol{x}$ to parameters (μ, κ), let us consider μ and κ in turn.

The impact of noise on μ. We have $\mu = \text{atan2}(x_2, x_1)$, whose gradient and Hessian with respect to $\boldsymbol{x}$ are

$$\nabla_{\boldsymbol{x}}\mu = \frac{1}{\kappa^2}\begin{pmatrix}-x_2\\x_1\end{pmatrix}, \qquad \boldsymbol{H}_{\boldsymbol{x}}\mu = \frac{1}{\kappa^4}\begin{pmatrix}2x_1x_2 & x_2^2-x_1^2\\ x_2^2-x_1^2 & -2x_1x_2\end{pmatrix}, \tag{S115}$$

where we have used $\kappa = \sqrt{x_1^2+x_2^2}$. Applying Itô's Lemma to this mapping results in

$$\begin{aligned}\mathrm{d}\mu &= \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{G}^T\boldsymbol{H}_{\boldsymbol{x}}\mu\boldsymbol{G}\right)\mathrm{d}t + \left(\nabla_{\boldsymbol{x}}\mu\right)^T\boldsymbol{G}\mathrm{d}\boldsymbol{W}\\ &= \frac{4}{\kappa^4N^2}\left(\left(\boldsymbol{c}^{2^T}\boldsymbol{h}^2 - \boldsymbol{s}^{2^T}\boldsymbol{h}^2\right)x_1x_2 + \boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s}\left(x_2^2-x_1^2\right)\right)\mathrm{d}t\\ &\quad + \frac{2}{\kappa^2N\sqrt{\boldsymbol{c}^{2^T}\boldsymbol{h}^2}}\left(\left(\boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s}x_1 - \boldsymbol{c}^{2^T}\boldsymbol{h}^2x_2\right)\mathrm{d}W_1 + \sqrt{\boldsymbol{s}^{2^T}\boldsymbol{h}^2\boldsymbol{c}^{2^T}\boldsymbol{h}^2 - \left(\boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s}\right)^2}x_1\mathrm{d}W_2\right),\end{aligned} \tag{S116}$$

containing both a drift (second-to-last line) and a diffusion term (last line).

The impact of noise on κ. We have $\kappa = \sqrt{x_1^2+x_2^2}$ whose gradient and Hessian with respect to $\boldsymbol{x}$ are

$$\nabla_{\boldsymbol{x}}\kappa = \frac{1}{\kappa}\begin{pmatrix}x_1\\x_2\end{pmatrix}, \qquad \boldsymbol{H}_{\boldsymbol{x}}\kappa = \frac{1}{\kappa^3}\begin{pmatrix}x_2^2 & -x_1x_2\\ -x_1x_2 & x_1^2\end{pmatrix}. \tag{S117}$$

Applying Itô's Lemma to this mapping results in

$$\begin{aligned}\mathrm{d}\kappa &= \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{G}^T\boldsymbol{H}_{\boldsymbol{x}}\kappa\boldsymbol{G}\right)\mathrm{d}t + \left(\nabla_{\boldsymbol{x}}\kappa\right)^T\boldsymbol{G}\mathrm{d}\boldsymbol{W}\\ &= \frac{2}{\kappa^3N^2}\left(\boldsymbol{s}^{2^T}\boldsymbol{h}^2x_1^2 - 2\boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s}x_1x_2 + \boldsymbol{c}^{2^T}\boldsymbol{h}^2x_2^2\right)\mathrm{d}t\\ &\quad + \frac{2}{\kappa N\sqrt{\boldsymbol{c}^{2^T}\boldsymbol{h}^2}}\left(\left(\boldsymbol{c}^{2^T}\boldsymbol{h}^2x_1 + \boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s}x_2\right)\mathrm{d}W_1 + \sqrt{\boldsymbol{s}^{2^T}\boldsymbol{h}^2\boldsymbol{c}^{2^T}\boldsymbol{h}^2 - \left(\boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s}\right)^2}x_2\mathrm{d}W_2\right),\end{aligned} \tag{S118}$$

again containing both a drift and a diffusion term.

B.3. Neural noise models. To get a better understanding of the resulting μ and κ dynamics, we will now consider different noise models. In particular, we will consider additive, Poisson-like multiplicative, and Weber-like multiplicative noise. The difference between Poisson-like and Weber-like multiplicative noise is that, for Poisson-like noise, the noise *variance* scales with neural activity, whereas, for Weber-like noise, it is the noise *standard deviation* that scales with neural activity. While we make no assumptions about the shape of population activity for the additive noise case, we will assume sinusoidal activity for multiplicative noise

$$r_j \approx \kappa\cos\left(\mu-\phi_j\right) + b = \kappa\cos(\mu)\cos\left(\phi_j\right) + \kappa\sin(\mu)\sin\left(\phi_j\right) + b = x_1c_j + x_2s_j + b, \tag{S119}$$

where b denotes the baseline activity. This assumption is required to find analytical results, and is warranted by noting that our single-population networks were designed to filter out higher-order Fourier components, such that their contribution should be minimal.

Additive noise. For additive neural noise we assume $h(r_j) = h_j = \sigma_{nn}$, independent of neural activity. This leads to

$$\boldsymbol{c}^{2^T}\boldsymbol{h}^2 = \boldsymbol{s}^{2^T}\boldsymbol{h}^2 = \frac{N}{2}\sigma_{nn}^2, \qquad \boldsymbol{c}^T\mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s} = 0, \tag{S120}$$

where we have taken the large-N integral limit for the involved sums. Substituting these expressions into Eqs. [S116] & [S118] results in

$$\mathrm{d}\mu = \frac{\sqrt{2}\sigma_{nn}}{\kappa^2\sqrt{N}}\left(-x_2\mathrm{d}W_1 + x_1\mathrm{d}W_2\right), \tag{S121}$$

$$\mathrm{d}\kappa = \frac{\sigma_{nn}^2}{\kappa N}\mathrm{d}t + \frac{\sqrt{2}\sigma_{nn}}{\kappa\sqrt{N}}\left(x_1\mathrm{d}W_1 + x_2\mathrm{d}W_2\right), \tag{S122}$$

with moments

$$\langle\mathrm{d}\mu\rangle = 0, \qquad \langle\mathrm{d}\kappa\rangle = \frac{\sigma_{nn}^2}{\kappa N}\mathrm{d}t, \tag{S123}$$

$$\mathrm{var}\left(\mathrm{d}\mu\right) = \frac{2\sigma_{nn}^2}{\kappa^2N}\mathrm{d}t, \qquad \mathrm{var}\left(\mathrm{d}\kappa\right) = \frac{2\sigma_{nn}^2}{N}\mathrm{d}t, \tag{S124}$$

$$\mathrm{cov}\left(\mathrm{d}\mu, \mathrm{d}\kappa\right) = 0. \tag{S125}$$

This shows that additive neural noise causes μ to only diffuse without introducing additional drift, and κ to both drift and diffuse. The drift of κ is obvious in hindsight, as it corresponds to the on average increasing radius of a two-dimensional random walk.

Poisson-like multiplicative noise. For Poisson-like multiplicative noise we assume $h(r_j) = h_j = \alpha\sqrt{r_j}$ such that, by Eq. [S106], the noise variance, $\mathrm{var}\left(\mathrm{d}r_j\right) = \alpha^2 r_j \mathrm{d}t$ is linear in the neuron's activity r_j. Assuming population activity to be described by Eq. [S119] results in

$$\boldsymbol{c}^{2^T}\boldsymbol{h}^2 = \boldsymbol{s}^{2^T}\boldsymbol{h}^2 = \frac{\alpha^2 N b}{2}, \qquad \boldsymbol{c}^T \mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s} = 0. \tag{S126}$$

where we have again taken the large-N integral limit for the involved sums. Substituting these expressions into Eqs. [S116] & [S118] results in

$$\mathrm{d}\mu = \frac{\sqrt{2b}\alpha}{\kappa^2\sqrt{N}}\left(-x_2\mathrm{d}W_1 + x_1\mathrm{d}W_2\right), \tag{S127}$$

$$\mathrm{d}\kappa = \frac{\alpha^2 b}{\kappa N}\mathrm{d}t + \frac{\sqrt{2b}\alpha}{K\sqrt{N}}\left(x_1\mathrm{d}W_1 + x_2\mathrm{d}W_2\right), \tag{S128}$$

with moments

$$\langle\mathrm{d}\mu\rangle = 0, \qquad \langle\mathrm{d}\kappa\rangle = \frac{\alpha^2 b}{\kappa N}\mathrm{d}t, \tag{S129}$$

$$\mathrm{var}\left(\mathrm{d}\mu\right) = \frac{2\alpha^2 b}{\kappa^2 N}\mathrm{d}t, \qquad \mathrm{var}\left(\mathrm{d}\kappa\right) = \frac{2\alpha^2 b}{N}\mathrm{d}t, \tag{S130}$$

$$\mathrm{cov}\left(\mathrm{d}\mu, \mathrm{d}\kappa\right) = 0. \tag{S131}$$

The moments are the same as for the additive noise model with a baseline activity-dependent noise variance $\sigma_{nn}^2 = \alpha^2 b$.

Weber-like multiplicative noise. For Weber-like multiplicative noise we assume $h(r_j) = h_j = \alpha r_j$ such that, by Eq. [S106], the noise standard deviation, $\sqrt{\mathrm{var}\left(\mathrm{d}r_j\right)} = \alpha r_j\sqrt{\mathrm{d}t}$ is linear in the neuron's activity r_j. Assuming again that population activity is described by Eq. [S119] results in

$$\boldsymbol{c}^{2^T}\boldsymbol{h}^2 = \frac{N\alpha^2}{2}\left(\frac{1}{4}\left(x_1^2 - x_2^2\right) + \frac{1}{2}\kappa^2 + b^2\right), \quad \boldsymbol{s}^{2^T}\boldsymbol{h}^2 = \frac{N\alpha^2}{2}\left(\frac{1}{4}\left(x_2^2 - x_1^2\right) + \frac{1}{2}\kappa^2 + b^2\right), \quad \boldsymbol{c}^T \mathrm{diag}\left(\boldsymbol{h}^2\right)\boldsymbol{s} = \frac{N\alpha^2}{4}x_1 x_2. \tag{S132}$$

Substituting these expressions into Eqs. [S116] & [S118] results in

$$\mathrm{d}\mu = \frac{\sqrt{2}\alpha\sqrt{\frac{1}{4}\kappa^2 + b^2}}{\kappa^2\sqrt{N}\sqrt{\frac{1}{4}\left(x_1^2 - x_2^2\right) + \frac{1}{2}\kappa^2 + b^2}}\left(-\sqrt{\frac{1}{4}\kappa^2 + b^2}x_2\mathrm{d}W_1 + \sqrt{\frac{3}{4}\kappa^2 + b^2}x_1\mathrm{d}W_2\right) \tag{S133}$$

$$\mathrm{d}\kappa = \frac{\alpha^2}{\kappa N}\left(\frac{1}{4}\kappa^2 + b^2\right)\mathrm{d}t + \frac{\sqrt{2}\alpha\sqrt{\frac{3}{4}\kappa^2 + b^2}}{\kappa\sqrt{N}\sqrt{\frac{1}{4}\left(x_1^2 - x_2^2\right) + \frac{1}{2}\kappa^2 + b^2}}\left(\sqrt{\frac{3}{4}\kappa^2 + b^2}x_1\mathrm{d}W_1 + \sqrt{\frac{1}{4}\kappa^2 + b^2}x_2\mathrm{d}W_2\right). \tag{S134}$$

with moments

$$\langle\mathrm{d}\mu\rangle = 0, \qquad \langle\mathrm{d}\kappa\rangle = \frac{\alpha^2}{\kappa N}\left(\frac{1}{4}\kappa^2 + b^2\right)\mathrm{d}t, \tag{S135}$$

$$\mathrm{var}\left(\mathrm{d}\mu\right) = \frac{2\alpha^2}{\kappa^2 N}\left(\frac{1}{4}\kappa^2 + b^2\right)\mathrm{d}t, \qquad \mathrm{var}\left(\mathrm{d}\kappa\right) = \frac{2\alpha^2}{N}\left(\frac{3}{4}\kappa^2 + b^2\right)\mathrm{d}t, \tag{S136}$$

$$\mathrm{cov}\left(\mathrm{d}\mu, \mathrm{d}\kappa\right) = 0. \tag{S137}$$

In summary, neither noise model results in a drift in μ, but all cause its diffusion with a diffusion variance that depends on the chosen noise model. As this diffusion holds irrespective of whether the system is at its attractor states, these results generalizes previous results for diffusion close to the attractor state (22). Furthermore, all noise models result in a positive drift in κ away from the origin, as well as a noise model-dependent diffusion variance. In all cases, both drift and diffusion magnitude for both μ and κ drop with N, and so become negligible once the population becomes significantly large, again generalizing the results in (22) to dynamics away from the attractor state.

C. Compensating for noisy neurons when performing inference. As we have seen, neural noise affects both the dynamics of μ and κ. For all noise models, it adds a zero-mean diffusion to μ, and a positive drift and diffusion to κ. The additional perturbations are all of order $1/N$ and so become negligible once the neural population becomes sufficiently large. For small population sizes, however, it might introduce perturbations that significantly impact inference accuracy in the network filter, or, in other words, to significantly deviate from the circular Kalman filter. Here we discuss how to qualitatively counter-act these perturbations to keep their impact to a minimum.

Let us first focus on μ. Without neural noise, the circular Kalman filter already assumes μ a-priori to follow a zero-mean diffusion on the circle, Eq. [S3], and additional diffusion due to noisy angular velocity observations, Eq. [S1]. Both reduce certainty in the HD estimate, which the filter accounts for by a drop in κ, as implemented by a leak term in Eq. [S20]. The additional zero-mean diffusion introduced by neural noise further reduces the HD estimate's certainty and thus needs to be

accounted for by an additional leak of κ whose strength depends on the noise model. Thus, the impact of neural noise on μ can be adequately accounted for by an additional leak of κ.

The impact of neural noise on κ requires a similar counter-measure. Without neural noise, the leak in the dynamics of κ, Eq. [S20], results in a leaky accumulation of all remaining terms. This also applies to diffusion introduced by neural noise: it will be integrated with leak, resulting its impact to be bounded. The stronger the leak, the weaker its impact. The drift introduced by neural noise has a different effect: if not accounted for, it would cause the inference of κ to be biased. In particular, as the drift is positive for all noise models, it would result in an overestimation of κ and so in overconfidence of the network filter. Fortunately, we can account for this drift with an additional leak term of the same size as the drift. Thus, the impact of neural noise on κ results in bounded additional diffusion of κ, and a drift that can be accounted for by an additional leak of κ.

To summarize, neural noise results in an additional, unavoidable diffusion of μ, and a drift and diffusion of κ, both of which can be accounted for by an additional leak of κ. The exact expression for the required leak depends on the chosen noise model, and for neither model precisely matches our Bayesian ring attractor's exact architecture and parametrization. Therefore, we used numerical optimization to find the parameters that maximize HD tracking performance rather than relying on the above analytical expressions. As we show in the main text and Fig. S4, in light of neural noise, such a network with re-tuned parameters outperforms one that is only optimally tuned for the noise-free case, as expected from the above analysis.

6. Supplementary Figures

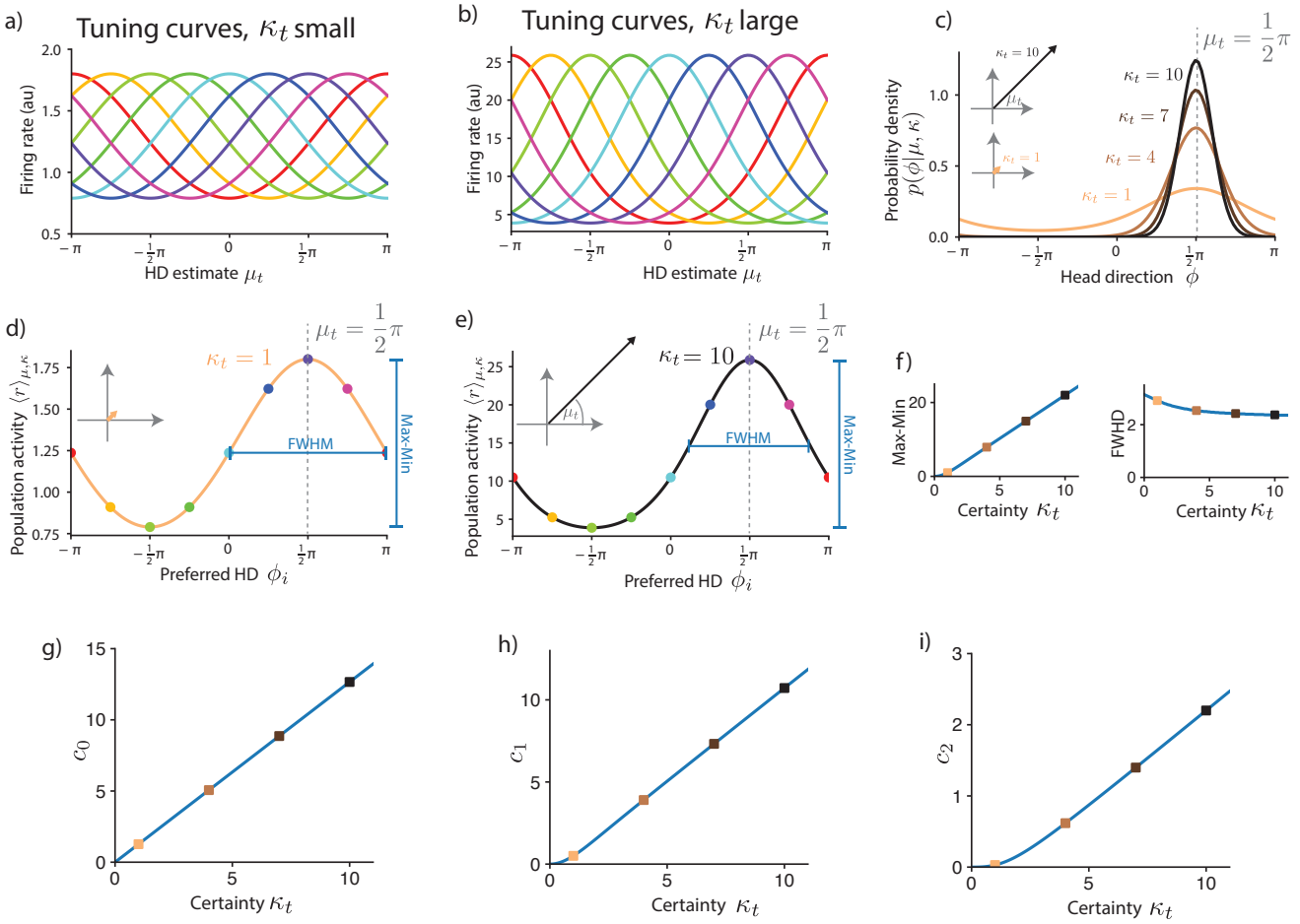

Fig. S1. Encoding the HD with linear probabilistic population codes. a) Tuning curves with respect to encoded HD estimate for small values of encoded certainty κ_t are cosine-shaped. Here, we show tuning curves of 8 example neurons with $\kappa_t = 1$ (colors indicate preferred HD ϕ_i). **b)** Tuning curves with respect to HD estimate for large values of encoded certainty κ_t are von-Mises shaped (same 8 example neurons as in a, but for $\kappa_t = 10$). **c)** Von Mises probability densities for different values of encoded certainty κ_t and fixed mean $\mu_t = \frac{\pi}{2}$. Note that the density sharpens around the mean with increasing certainty. Inset shows vector representation of a von Mises distribution with mean $\mu_t = \frac{\pi}{2}$, and, respectively, $\kappa_t = 10$ and $\kappa_t = 1$. **d)** Population activity profile (average neural firing rate conditioned on HD estimate μ_t and certainty κ_t) encoding the von Mises densities with mean $\mu_t = \frac{\pi}{2}$ and certainty $\kappa_t = 1$. Neurons are sorted by preferred HD ϕ_t. Colored dots correspond to activity of neurons with tuning curves as in a). The phasor representation of the neural activity (inset) matches the vector representation of the encoded von Mises distribution in c. **e)** Population activity profile encoding the von Mises densities with mean $\mu_t = \frac{\pi}{2}$ and certainty $\kappa_t = 10$. **f)** Left: The amplitude (Max-Min) of the activity profile scales (approximately) linearly with certainty κ_t, except for very small values of κ_t. Right: The population activity bump's width (full width at half maximum, FWHM) is mostly unaffected by uncertainty κ_t, and saturates at a finite value for large κ_t, unlike the von Mises distribution it encodes (e.g., b), whose FWHM approaches zero for large values of κ_t. **g-j)** The Fourier component amplitudes of the population activity profile are mostly linear in encoded certainty κ_t, indicating that (i) the whole profile is scaled by κ_t, and that (ii) only focusing on the first Fourier component in our analysis is justified. For the tuning curves, we used $\xi = 1$ without loss of generality.

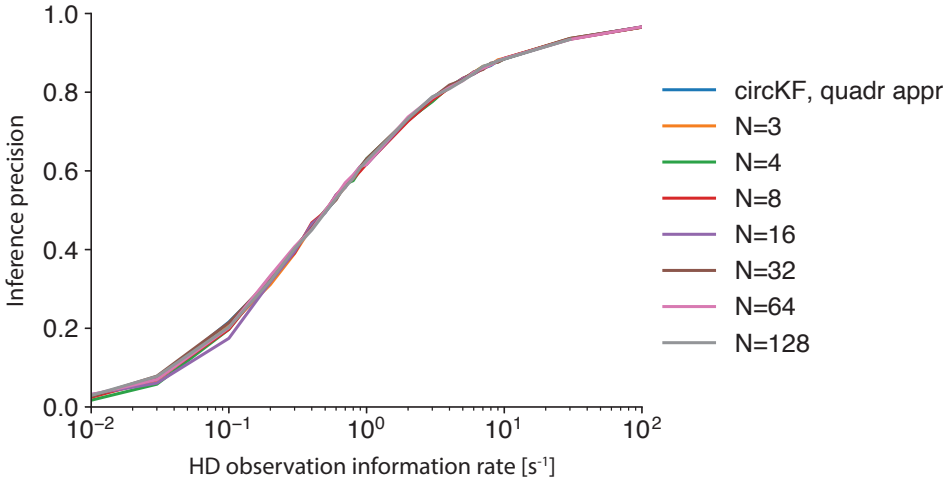

Fig. S2. Network inference performance is mostly independent of the number of neurons N in the Bayesian ring attractor network. Here, for each value of the observation reliability κ_z and number of neurons in the network N we compute the circular average distance of the network's HD estimate μ_T from the true HD ϕ_T at the end of a simulation of length $T = 20$ from $P = 10'000$ simulated trajectories. The blue line (hidden below other lines) shows the performance of the quadratic approximation to the circular Kalman filter that the networks aim to implement. The network parameters of the single-population network in Eq. [S49] were those of the Bayesian ring attractor, i.e. $w_1^{\text{even}} = \frac{1}{\tau} + \frac{1}{\kappa_\phi + \kappa_v}$, $w_1^{\text{odd}} = \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t$, and $w^{\text{quad}} = \frac{1}{\kappa_\phi + \kappa_v}$. Other simulation parameters were: $\kappa_\phi = 1$, $\kappa_v = 1$, and $\Delta t = 0.01$.

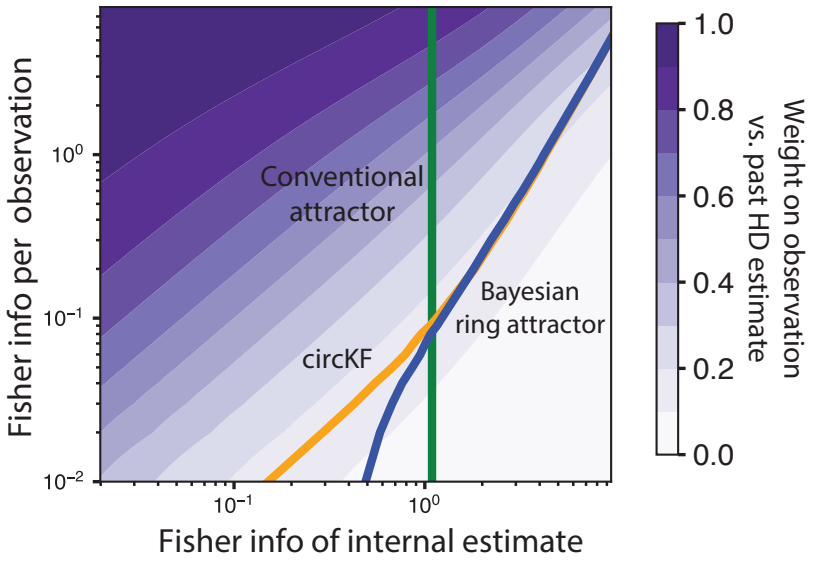

Fig. S3. The weight with which a single observation contributes to the HD estimate varies with informativeness of both the HD observations and the current HD estimate. Same plot as main text Fig. 4, only on a log-log scale. Here, we additionally plot the resulting updates for the circKF, to demonstrate that the Bayesian ring attractor (blue curve) only deviates from the circKF (yellow curve) for very uninformative observations.

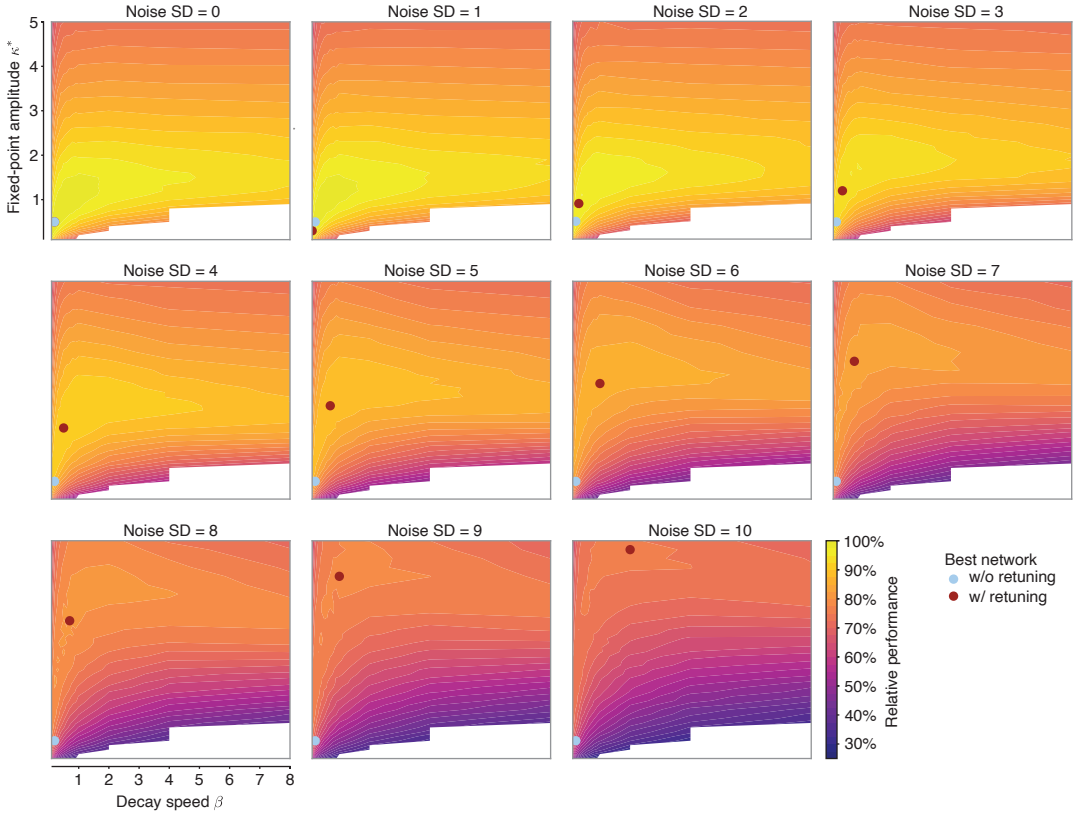

Fig. S4. Neural noise changes the optimal fixed point amplitude and decay speed. We simulated a network of $N = 64$ neurons with different levels of additive Gaussian noise with variance $\sigma_{nn}^2 \delta t$ to each neuron within each time step δt, for different fixed point amplitudes κ^* and decay speeds β. As in main text Fig. 3D, the performance of each network was assessed by its average inference accuracy over different HD observation information rates, weighted by a prior over these information rates (see Methods for simulation details and parameters). Each panel shows this performance, relative to the best performance of a noise-free network, for a grid over values of κ^* and β. As can be seen, the optimal κ^* and β that maximizes relative performance changes with σ_{nn} (purple dot), and differs from the best κ^* and β for the noise-free network (light blue dot). In particular, larger noise requires re-tuning the network to use a larger κ^* and β.

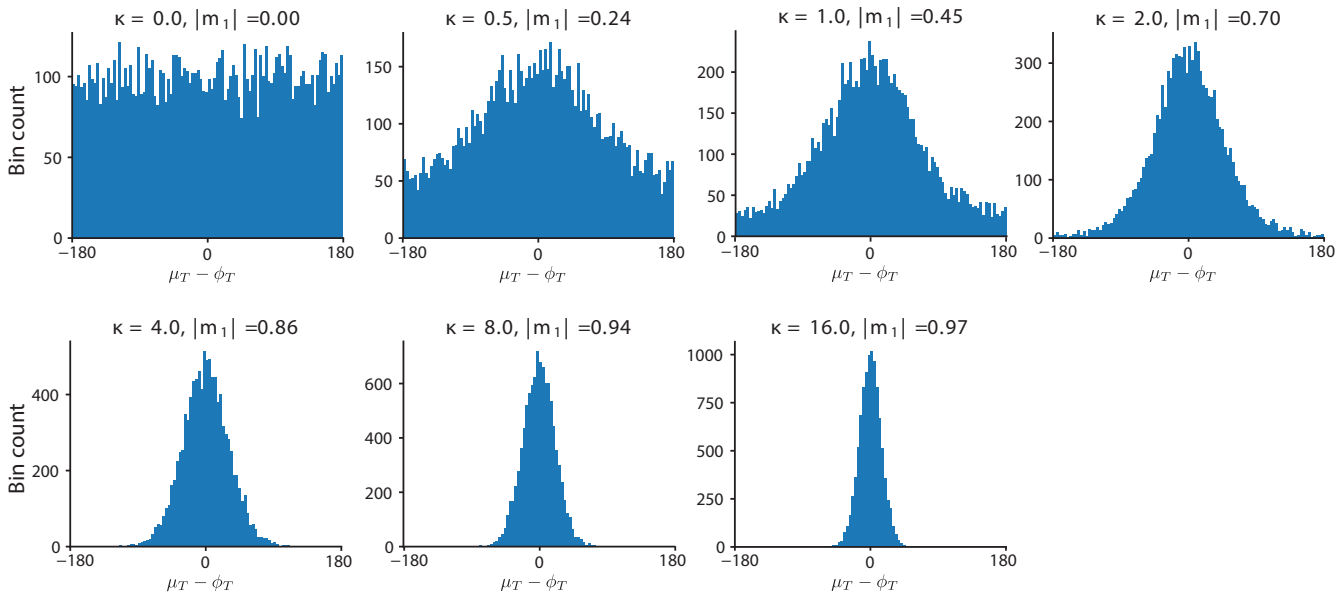

Fig. S5. Visualizing the HD tracking perfomance measure. To provide a better intuition for the used HD tracking performance measure we here show how a specific distribution of HD tracking errors (horizontal axis, in degrees) relates to this performance measure. In particular, we drew 10000 samples from a von Mises distribution $\mu_T - \phi_T \sim \mathcal{VM}(0, \kappa)$, where each drawn sample simulates one single deviation of the estimated HD (i.e., the mean of the filter posterior, μ_T) from the actual, true HD, ϕ_T. The different panels show the histogram of simulated errors for different κ's (see panel headings). Our filtering performance measure, that is, the absolute value of the first circular average of the samples, can be computed for the von Mises distribution via $|m_1| = \frac{I_1(\kappa)}{I_0(\kappa)}$ (23). We confirmed numerically that this analytical expression matches the circular average empirically determined from these simulated errors. Simulating HD tracking errors by draws from a von Mises distribution was here only performed for convenience. The HD tracking errors arising in simulations of the filtering algorithms do not necessarily follow such a distribution.

References

1. A Kutschireiter, L Rast, J Drugowitsch, Projection Filtering with Observed State Increments with Applications in Continuous-Time Circular Filtering. *IEEE Transactions on Signal Process.* **70**, 686–700 (2022) Conference Name: IEEE Transactions on Signal Processing.
2. D Brigo, B Hanzon, F Le Gland, Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli* **5**, 495–534 (1999).
3. CW Gardiner, *Stochastic methods: a handbook for the natural and social sciences*, Springer series in synergetics. (Springer, Berlin Heidelberg) No. 13, 4th ed edition, (2009).
4. A Doucet, S Godsill, C Andrieu, On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* p. 12 (2010).
5. A Kutschireiter, SC Surace, JP Pfister, The Hitchhiker's guide to nonlinear filtering. *J. Math. Psychol.* **94**, 102307 (2020).
6. WJ Ma, JM Beck, PE Latham, A Pouget, Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–8 (2006).
7. JM Beck, PE Latham, A Pouget, Marginalization in Neural Circuits with Divisive Normalization. *J. Neurosci.* **31**, 15310–15319 (2011).
8. A Pouget, JM Beck, WJ Ma, PE Latham, Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–8 (2013).
9. P Dayan, LF Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems*, Computational neuroscience. (Massachusetts Institute of Technology Press, Cambridge, Mass), (2001).
10. LK Scheffer, et al., A connectome and analysis of the adult Drosophila central brain. *eLife* **9**, e57443 (2020) Publisher: eLife Sciences Publications, Ltd.
11. BK Hulse, et al., A connectome of the Drosophila central complex reveals network motifs suitable for flexible navigation and context-dependent action selection. *eLife* **10**, e66039 (2021) Publisher: eLife Sciences Publications, Ltd.
12. DB Turner-Evans, et al., The Neuroanatomical Ultrastructure and Function of a Biological Ring Attractor. *Neuron* **108**, 145–163.e10 (2020).
13. JD Seelig, V Jayaraman, Neural dynamics for landmark orientation and angular path integration. *Nature* **521**, 186–191 (2015).
14. D Turner-Evans, et al., Angular velocity integration in a fly heading circuit. *eLife* **6**, e23496 (2017).
15. J Green, et al., A neural circuit architecture for angular integration in Drosophila. *Nature* **546**, 101–106 (2017) Publisher: Nature Publishing Group.
16. JJ Omoto, et al., Visual Input to the Drosophila Central Complex by Developmentally and Functionally Distinct Neuronal Populations. *Curr. Biol.* **27**, 1098–1110 (2017).
17. YE Fisher, J Lu, I D'Alessandro, RI Wilson, Sensorimotor experience remaps visual input to a heading-direction network. *Nature* **576**, 121–125 (2019).
18. SS Kim, AM Hermundstad, S Romani, LF Abbott, V Jayaraman, Generation of stable heading representations in diverse visual scenes. *Nature* pp. 1–6 (2019) Publisher: Springer US.
19. BK Hulse, V Jayaraman, Mechanisms Underlying the Neural Computation of Head Direction. *Annu. Rev. Neurosci.* **43**, 31–54 (2020).
20. W Skaggs, J Knierim, H Kudrimoti, B McNaughton, A model of the neural basis of the rats sense of direction in *Advances in neural information processing systems*, eds. G Tesauro, D Touretzky, T Leen. (MIT Press), Vol. 7, (1994).
21. K Zhang, Representation of Spatial Orientation by the Intrinsic Dynamics of the Head-Direction Cell Ensemble: A Theory. *The J. Neurosci.* **16**, 2112–2126 (1996).
22. Y Burak, IR Fiete, Fundamental limits on persistent activity in networks of noisy neurons. *Proc. Natl. Acad. Sci.* **109**, 17645–17650 (2012).
23. KV Mardia, PE Jupp, *Directional Statistics*. (John Wiley & Sons), (2000) Pages: 3.