PNAS RESEARCH ARTICLE | NEUROSCIENCE
BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Bayesian inference in ring attractor networks

Anna Kutschireiter[a,1], Melanie A. Basnak[a], Rachel I. Wilson[a], and Jan Drugowitsch[a,1]

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received June 20, 2022; accepted January 12, 2023

Working memories are thought to be held in attractor networks in the brain. These attractors should keep track of the uncertainty associated with each memory, so as to weigh it properly against conflicting new evidence. However, conventional attractors do not represent uncertainty. Here, we show how uncertainty could be incorporated into an attractor, specifically a ring attractor that encodes head direction. First, we introduce a rigorous normative framework (the circular Kalman filter) for benchmarking the performance of a ring attractor under conditions of uncertainty. Next, we show that the recurrent connections within a conventional ring attractor can be retuned to match this benchmark. This allows the amplitude of network activity to grow in response to confirmatory evidence, while shrinking in response to poor-quality or strongly conflicting evidence. This "Bayesian ring attractor" performs near-optimal angular path integration and evidence accumulation. Indeed, we show that a Bayesian ring attractor is consistently more accurate than a conventional ring attractor. Moreover, near-optimal performance can be achieved without exact tuning of the network connections. Finally, we use large-scale connectome data to show that the network can achieve near-optimal performance even after we incorporate biological constraints. Our work demonstrates how attractors can implement a dynamic Bayesian inference algorithm in a biologically plausible manner, and it makes testable predictions with direct relevance to the head direction system as well as any neural system that tracks direction, orientation, or periodic rhythms.

working memory | ring attractor networks | head direction neurons | Bayesian inference | Kalman filter

Significance

Data from human subjects as well as animals show that working memories are associated with a sense of uncertainty. Indeed, a sense of uncertainty is what allows an observer to properly weigh new evidence against their current memory. However, we do not understand how the brain tracks uncertainty. Here, we describe a simple and biologically plausible network model that can track the uncertainty associated with working memory. The representation of uncertainty in this model improves the accuracy of its working memory, as compared to conventional models, because it assigns proper weight to new conflicting evidence. Our model provides an interpretation of observed fluctuations in brain activity, and it makes testable predictions.

Attractor networks are thought to form the basis of working memory (1, 2) as they can exhibit persistent, stable activity patterns (attractor states) even after network inputs have ceased (3). An attractor network can gravitate toward a stable state even if its input is based on partial (unreliable) information; this is why attractors have been suggested as a mechanism for pattern completion (4). However, the characteristic stability of any attractor network also creates a problem: Once the network has settled into its attractor state, it will no longer be possible to see that its inputs might have been unreliable. In this situation, the attractor state will simply represent a point estimate (or "best guess") of the remembered input, without any associated sense of uncertainty. However, real memories often include a sense of uncertainty, (e.g., refs. 5–7), and uncertainty has clear behavioral effects (8–10). This motivates us to ask how an attractor network might conjunctively encode a memory and its associated uncertainty.

A ring attractor is a special case of an attractor that can encode a circular variable (11). For example, there is good evidence that the neural networks that encode head direction (HD) are ring attractors (12–19). In a conventional ring attractor, inputs push a "bump" of activity around the ring, with only short-lived changes in bump amplitude or shape (20, 21); the rapid decay to a stereotyped bump shape is by design, and, as a result, a conventional ring attractor network is unable to track uncertainty. However, it would be useful to modify these conventional ring attractors so that they can encode the uncertainty associated with HD estimates. HD estimates are constructed from two types of observations—angular velocity observations and HD observations (11, 22). Angular velocity observations arise from multiple sources, including efference copies, vestibular or proprioceptive signals, as well as optic flow; these observations indicate the head's rotational movement and, thus, a change in HD (13, 18, 23, 24). These angular velocity observations are integrated over time ("remembered") to update the system's internal estimate of HD, in a process termed angular path integration. Ideally, a ring attractor would track the uncertainty associated with angular path integration errors. Meanwhile, HD observations arise from visual landmarks or other sensory cues that provide information about the head's current orientation (12, 16). These sensory observations can change the system's internal HD estimate, and once that change has occurred, it is generally persistent (remembered). But like any sensory signal, these

Author affiliations: [a]Department of Neurobiology, Harvard Medical School, Boston, MA 02115

Author contributions: A.K., M.A.B., R.I.W., and J.D. designed research; A.K., M.A.B., R.I.W., and J.D. performed research; A.K. and J.D. contributed new reagents/analytic tools; A.K. and J.D. analyzed data; and A.K., M.A.B., R.I.W., and J.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: anna.kutschireiter@gmail.com or jan_drugowitsch@hms.harvard.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2210622120/-/DCSupplemental.

Published February 22, 2023.

sensory observations are noisy; they are not unambiguous evidence of HD. Therefore, the way that a ring attractor responds to each new visual landmark observation should ideally depend on the uncertainty associated with its current HD estimate. This type of uncertainty-weighted cue integration is a hallmark of Bayesian inference (25) and would require a network that is capable of keeping track of its own uncertainty.

In this study, we address three related questions. First, how should an ideal observer integrate uncertain evidence over time to estimate a circular variable? For a linear variable, this is typically done with a Kalman filter; here, we introduce an extension of Kalman filtering for circular statistics; we call this the circular Kalman filter. This algorithm provides a high-level description of how the brain should integrate evidence over time to estimate HD, or indeed any other circular or periodic variable. Second, how could a neural network actually implement the circular Kalman filter? We show how this algorithm could be implemented by a neural network whose basic connectivity pattern resembles that of a conventional ring attractor. With properly tuned network connections, we show that the bump amplitude grows in response to confirmatory evidence, whereas it shrinks in response to strongly conflicting evidence or the absence of evidence. We call this network a Bayesian ring attractor. Third, how does the performance of a Bayesian ring attractor compare to the performance of a conventional ring attractor? In a conventional ring attractor, bump amplitude is pulled rapidly back to a stable baseline value, whereas in a Bayesian ring attractor, bump amplitude is allowed to float up or down as the system's certainty fluctuates. As a result, we show that a Bayesian ring attractor has consistently more accurate internal estimates (or "working memory") of the variable it is designed to encode than a conventional ring attractor.

Together, these results provide a principled theoretical foundation for how ring attractor networks can be tuned to conjointly encode a memory and its associated uncertainty. Although we focus on the brain's HD system as a concrete example, our results are relevant to any other brain system that encodes a circular or periodic variable.

Results

Circular Kalman Filtering: A Bayesian Algorithm for Tracking a Circular Variable. We begin by asking how an ideal observer should dynamically integrate uncertain evidence to estimate a circular variable, specifically head direction ϕ_t. Additional information being absent, the ideal observer assumes that HD follows a random walk or diffusion on a circle: across small consecutive time steps of size δt, the current HD ϕ_t is assumed to be drawn from a normal distribution, $\phi_t|\phi_{t-\delta t} \sim \mathcal{N}(\phi_{t-\delta t}, \delta t/\kappa_\phi)$ (constrained to a circle) centered on $\phi_{t-\delta t}$ and with variance $\delta t/\kappa_\phi$. This diffusion prior assumes smaller HD changes for a larger precision (i.e., inverse variance), κ_ϕ, and for smaller time steps, δt. Just like the brain's HD system, the ideal observer receives additional HD information through HD observations z_t and angular velocity observations, v_t (Fig. 1*A*). HD observations provide noisy, and thus unreliable, measurements of the current HD drawn from a von Mises distribution, $z_t|\phi_t \sim \mathcal{VM}(\phi_t, \kappa_z\delta t)$ (i.e., the equivalent to a normal distribution on a circle), centered on ϕ_t, and with precision $\kappa_z\delta t$. A higher precision κ_z means that individual HD observations provide more reliable information about the current HD. Angular velocity observations provide noisy measurement of the current HD change $\phi_t - \phi_{t-\delta t}$ drawn from a normal distribution, $v_t|\phi_t, \phi_{t-\delta t} \sim \mathcal{N}\left(\frac{\phi_t-\phi_{t-\delta t}}{\delta t}, \frac{1}{\kappa_v\delta t}\right)$ centered on the current angular velocity and with precision $\kappa_v\delta t$. While higher-precision measurements yield more reliable information, they only do so about the current HD change rather than the HD itself.

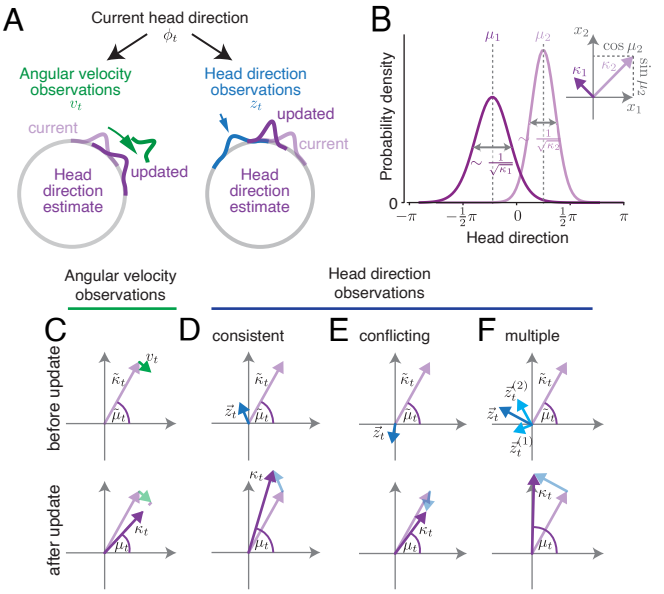

Fig. 1. Tracking HD with the circular Kalman filter. (*A*) Angular velocity observations provide noisy information about the true angular velocity $\dot{\phi}_t$, while HD observations provide noisy information about the true HD ϕ_t. (*B*) At every point in time, the posterior belief $p(\phi_t)$ is approximated by a von Mises distribution, which is fully characterized by its mean μ_t (location of distribution's peak) and its precision/certainty parameter κ_t. Interpreted as the polar coordinates in the 2D plane, these parameters provide a convenient vector representation of the posterior belief (inset). (*C*) An angular velocity observation v_t is a vector tangent to the current HD belief vector. Angular velocity observations continually rotate the current HD estimate; meanwhile, noise accumulation progressively decreases certainty. (*D*) Each HD observation z_t is a vector whose length quantifies the observation's reliability. Adding this vector to the current HD belief vector produces an updated HD belief vector. HD observations compatible with the current HD estimate result in an increased certainty (i.e., a longer belief vector). (*E*) HD observations in conflict with the current belief (e.g., opposite direction of the current estimate) decrease the belief's certainty. (*F*) Multiple HD cues can be integrated simultaneously via vector addition.

The aim of the ideal observer is to use Bayesian inference to maintain a posterior belief over HD, $p(\phi_t|z_{0:t}, v_{0:t})$ given all past observations, $z_{0:t}$ and $v_{0:t}$ (25, 26). Assuming a belief $p(\phi_{t-\delta t}|z_{0:t-\delta t}, v_{0:t-\delta t})$ at time $t-\delta t$, the observer updates this belief upon observing v_t and z_t in two steps. First, it combines its a priori assumption about how HD diffuses across time with the current angular velocity observation v_t to predict ϕ_t at the next time step t, leading to $p(\phi_t|z_{0:t-\delta t}, v_{0:t})$. As both the diffusion prior and angular velocity observations are noisy, this prediction will be less certain than the previous belief it is based on. (*SI Appendix* for formal expression.) Second, the ideal observer uses Bayes' rule to combine this prediction with the current HD observation z_t to form the updated posterior belief $p(\phi_t|z_{0:t}, v_{0:t})$. These two steps are iterated across consecutive time steps to continuously update the HD belief in the light of new observations.

The two steps are also the ones underlying a standard Kalman filter (27, 28). However, while a standard Kalman filter assumes the encoded variable to be linear, we here use a circular variable which requires a different approach. Because filtering on a circle is analytically intractable (29), we choose to approximate the

posterior belief by a von Mises distribution, with mean μ_t and precision κ_t, so that $p(\phi_t|z_{0:t}, v_{0:t}) \approx \mathcal{VM}(\phi_t|\mu_t, \kappa_t)$ (Fig. 1*B*). Then, the mean μ_t, which we will call the HD estimate, determines the peak of the distribution. The precision κ_t quantifies the width of the distribution and, therefore, our certainty in this estimate (larger κ_t = indicating higher certainty). This approximation allows us to update the posterior over a circular variable ϕ_t using a technique called projection filtering (30, 31), resulting in,

$$\mu_t = \mu_{t-\delta t} + \left(\frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t + \frac{\kappa_z}{\kappa_t} \sin(z_t - \mu_t)\right) \delta t, \quad [1]$$

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

Here, $f(\kappa_t)$ is a monotonically increasing nonlinear function that controls the speed of decay in certainty κ_t (*Methods*). Eqs. **1** and **2** describe an algorithm that we call the circular Kalman filter (31) (*Methods*/*SI Appendix* for a continuous-time formulation). This algorithm provides a general solution for estimating the evolution of a circular variable over time from noisy observations.

To understand the circular Kalman filter intuitively, it is helpful to think of the observer's belief as a vector in the 2D plane (Fig. 1*B*), whose direction represents the current estimate μ_t, and whose length represents the associated certainty κ_t. The circular Kalman filter tells us how this vector should change at each time point, based on new observations of angular velocity and HD. Here, we outline the intuition behind the circular Kalman filter, focusing on the HD system as a specific example.

Angular Velocity Observations. We can think of each angular velocity observation as a vector that points at a tangent to the current HD belief vector (Fig. 1*C*) and rotates this belief vector (first term in parenthesis on RHS of Eq. **1**). Angular velocity observations are noisy and, together with the diffusion prior, decrease the belief's certainty (κ_t), meaning that the observer's belief vector becomes shorter (Fig. 1*C*). Thus, when angular velocity observations are the only inputs to the HD network—i.e., when HD observations are absent—the HD belief's certainty κ_t will progressively decay, with a speed of decay that depends on both κ_v and κ_ϕ (first term in parenthesis on RHS of Eq. **2**).

HD Observations. We can treat each HD observation as a vector whose length κ_z quantifies the observation's reliability (e.g., the reliability of a visual landmark observation). This HD observation vector is added to the current HD belief vector to obtain the updated HD belief vector. The updated direction of the belief vector depends on the relative lengths of both vectors. A relatively longer HD observation vector, i.e., a more reliable observation relative to the current belief's certainty, results in a stronger impact on the updated HD belief (Fig. 1*D*, second term in parenthesis on RHS of Eqs. **1** and **2**). In line with principles of reliability-weighted Bayesian cue combination (25), HD observations increase the observer's certainty if they are confirmatory (i.e., they indicate that the current estimate is correct or nearly so, Fig. 1*D*). Interestingly, however, if HD observations strongly conflict with the current estimate (e.g., if they point in the opposite direction), they actually decrease certainty (Fig. 1*E*). This notable result is a consequence of the circular nature of the inference task (32). It stands in contrast to the standard (noncircular) Kalman filter, where an analogous observation would always increase the observer's certainty (33) and is thus a key distinction between the standard Kalman filter and the circular Kalman filter.

To summarize, the circular Kalman filter describes how a nearly ideal observer should integrate a stream of unreliable information over time to update a posterior belief of a circular variable. This algorithm serves as a normative standard to judge the performance of any network in the brain that tracks a circular or periodic variable. Specifically, in the HD system, the circular Kalman filter tells us that angular velocity observations should rotate the HD estimate while reducing the certainty in that estimate. Meanwhile, HD observations should update the HD estimate weighted by their reliability, and they should either increase certainty (if compatible with the current estimate) or reduce it (if strongly conflicting with the current estimate). Note that the circular Kalman filter can integrate HD observations from multiple sources by simply adding all their vectors to the current HD belief vector (Fig. 1*F*).

Neural Encoding of a Probability Distribution. Thus far, we have developed a normative algorithmic description of how an observer should integrate evidence over time to track the posterior belief over a circular variable. This algorithm requires the observer to represent their current belief as a probability distribution on a circle. How could a neural network encode this probability distribution? Consider a ring attractor network where adjacent neurons have adjacent tuning preferences so that the population activity pattern is a spatially localized "bump." The bump's center of mass is generally interpreted as a point estimate (or best guess) of the encoded circular variable (12, 34). In the HD system, this would be the best guess of head direction. Meanwhile, we let the bump amplitude encode certainty so that higher amplitude corresponds to higher certainty. Of course, there are other ways to encode certainty—e.g., using bump width rather than bump amplitude. However, there are two good reasons for focusing on bump amplitude. First, as we will see below, this implementation allows the parameters of the encoded probability distribution to be "read out" in a way that supports the vector operations underlying the circKF (Fig. 1 *C*–*F*). Second, recent data from the mouse HD system show that the appearance of a visual cue (which increases certainty) causes bump amplitude to increase; moreover, when the bump amplitude is high, the network is relatively insensitive to the appearance of a visual cue that conflicts with the current HD estimate, again suggesting that bump amplitude is a proxy for certainty (19, 35).

Formally, then, the activity of a neuron i with preferred HD ϕ_i can be written as follows (Fig. 2*A*):

$$r_t^{(i)} = \kappa_t \cos(\phi_i - \mu_t) + \text{ other components}, \quad [3]$$

where μ_t is the encoded HD estimate, κ_t is the associated certainty, and the "other components" might include a constant (representing baseline activity) or minor contributions of higher-order Fourier components. Note that Eq. **3** does not imply that the tuning curve must be cosine-shaped. Rather, it implies that the cosine component of the tuning curve is scaled by certainty. This is satisfied, for example, by any unimodal bump profile whose overall gain is governed by certainty. A particularly interesting case that matches Eq. **3** is a linear probabilistic population code (36, 37) with von Mises-shaped tuning curves and independent Poisson neural noise (*SI Appendix*, Fig. S1).

Importantly, this neural representation would allow downstream neurons to read out the parameters of the probability distribution $p(\phi_t|z_{0:t}, v_{0:t})$ in a straightforward manner. Specifically, downstream neurons could take a weighted sum of the population firing rates (i.e., a linear operation; *Methods*) to

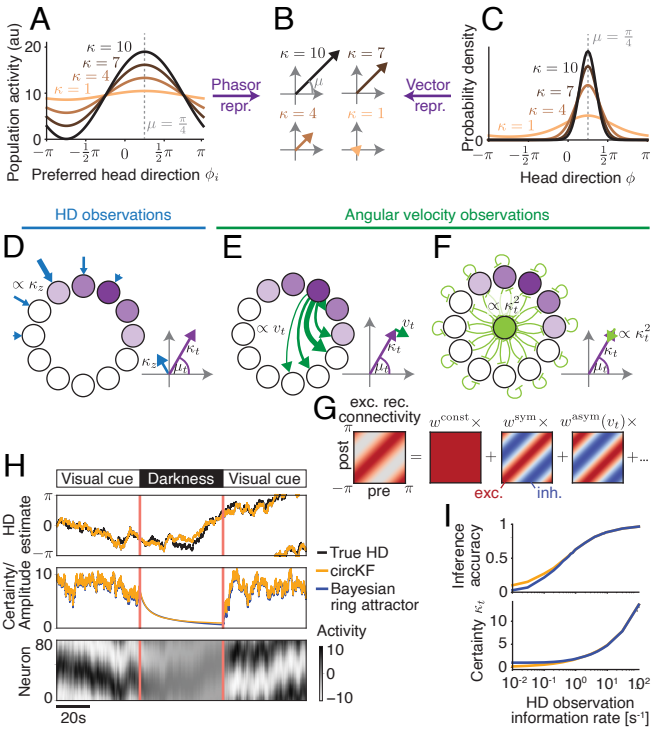

Fig. 2. A recurrent neural network implementation of the circular Kalman filter. The HD belief vector (*B*; Fig. 1*B*) is the "vector representation" of the HD belief (*C*), and the "phasor representation" (obtained from linear decoding) of sinusoidal population activity (*A*; neurons sorted by preferred HD ϕ_i), here shown for HD estimate $\mu = \pi/4$ (shift of activity/density) and different certainties κ (height of activity bump in *A*/sharpness of distribution in *C*). Using this duality between population activity and encoded HD belief, the circular Kalman filter can be implemented by three network motifs (*D*–*F*). (*D*) A cosine-shaped input to the network (strength = observation reliability κ_z) provides HD observation input. (*E*) Rotations of the HD belief vector are mediated by symmetric recurrent connectivities, whose strength is modulated by angular velocity observations. (*F*) Decay in amplitude, which implements decreasing HD certainty, arises from leak and global inhibition. (*G*) Rotation-symmetric recurrent connectivities (here, neurons are sorted according to their preferred HD) can be decomposed into constant, symmetric, asymmetric, and higher-order frequency components (here dots). (*H*) The dynamics of the Bayesian ring attractor implement the dynamics of the ideal observer's belief, as shown in a simulation of a network with 80 neurons. The network received angular velocity observations (always) and HD observations (only in "visual cue" periods). (*I*) The Bayesian ring attractor network tracks the true HD with the same accuracy (*Top*; higher = lower average circular distance to true HD; 1 = perfect, 0 = random; *Methods*) as the circular Kalman filter (circKF, Eqs. **1** and **2**) if HD observations are reliable and, therefore, more informative but with slightly lower accuracy once they become less reliable, and therefore less informative. This drop co-occurs with an overestimate in the belief's certainty κ_t (*Bottom*). HD observation reliability is measured here by the amount of Fisher information per unit time. The accuracies and certainties shown are averages over 5,000 simulation runs (*Methods* for details).

recover two parameters, $x_1 = \kappa_t \cos(\mu_t)$ and $x_2 = \kappa_t \sin(\mu_t)$. This is notable because x_1 and x_2 represent the von Mises distribution $p(\phi_t|z_{0:t}, v_{0:t})$ in terms of Cartesian vector coordinates in the 2D plane, whereas μ_t and κ_t are its polar coordinates (Fig. 1*B*). Having them accessible as vector coordinates makes it straightforward to implement the vector operations underlying the circKF (Fig. 1 *C*–*F*) in neural population dynamics. For example, as we will see in the next section, the vector sum required to account for HD observations in the circKF (Fig. 1 *D* and *E*) can be implemented by summing neural population activity (36). Overall, the vector representation of the HD posterior belief is related to the phasor representation of neural activity (38), which also translates bump position and amplitude to polar coordinates in the 2D plane (Fig. 2*B*). If the amplitude of the activity bump scales with certainty, the phasor representation of neural activity equals the vector representation of the von Mises distribution (Fig. 2 *B* and *C*).

Neural Network Implementation of the Circular Kalman Filter. Now that we have specified how our model network represents the probability distribution $p(\phi_t|z_{0:t}, v_{0:t})$ over possible head directions, we can proceed to considering the dynamics of this network—specifically, how it responds to incoming information or the lack of information. The circular Kalman filter algorithm describes the vector operations required to dynamically update the probability distribution $p(\phi_t|z_{0:t}, v_{0:t})$ with each new observation of angular velocity or head direction. In the absence of HD observations, the circKF's certainty decays to zero. By Eq. **3**, this implies that neural activity would also decay to zero, such that a network implementing the circKF would not be an attractor network. While we consider such a network in *Methods*, we here focus on the "Bayesian ring attractor" which approximates the circKF in an attractor network, thus establishing a stronger link to previous working memory literature (1, 2). We describe the features of this network with regard to the HD system, but the underlying concepts are general ones which could be applied to any network that encodes a circular or periodic variable. The dynamics of the Bayesian ring attractor network are given by

$$\mathrm{d}\boldsymbol{r}_t = -\frac{1}{\tau}\boldsymbol{r}_t\mathrm{d}t - g(\boldsymbol{r}_t)\cdot\boldsymbol{r}_t\mathrm{d}t + \boldsymbol{W}(v_t)\cdot\boldsymbol{r}_t\mathrm{d}t + \boldsymbol{I}_t^{ext}, \qquad [4]$$

where $\boldsymbol{r}_t$ denotes a population activity vector, with neurons ordered by their preferred HD ϕ_i, τ is the cell-intrinsic leak time constant, $\boldsymbol{W}(v_t)$ is the matrix of excitatory recurrent connectivity that is modulated by angular velocity observations v_t, $\boldsymbol{I}_t^{ext}$ is a vector of HD observations, and $g(\cdot)$ is a nonlinear function that determines global inhibition and that we discuss in more detail further below. Let us now consider each of these terms in detail.

First, HD observations enter the network via the input vector $\boldsymbol{I}_t^{ext}$ in the form of a cosine-shaped spatial pattern whose amplitude scales with reliability κ_z (Fig. 2*D*). This implements the vector addition required for the proper integration of these observations. Specifically, the weight assigned to each HD observation is determined by the amplitude of $\boldsymbol{I}_t^{ext}$, relative to the amplitude of the activity bump in the HD population. Thus, observations are weighted by their reliability, relative to the certainty of the current HD posterior belief, as in the circular Kalman filter (Fig. 1 *D* and *E*). An HD observation that tends to confirm the current HD estimate will increase the amplitude of the bump in HD cells and, thus, the posterior certainty.

Second, the matrix of recurrent connectivity $\boldsymbol{W}(v_t)$ has spatially symmetric and asymmetric components (Fig. 2*G*). The symmetric component consists of local excitatory connections that each neuron makes onto adjacent neurons with similar HD preferences. This holds the bump of activity at its current location in the absence of any other input. The overall strength of the symmetric component (w^{sym}) is a free parameter which we can tune. Meanwhile, the asymmetric component consists of excitatory connections that each neuron makes onto adjacent neurons with shifted HD preferences. This component tends to push the bump of activity around the ring (Fig. 2*E*). Angular velocity observations v_t modulate the overall strength of the asymmetric component ($w^{\mathrm{asym}}(v_t)$), so that positive and negative angular velocity observations push the bump in opposite directions.

Third, the global inhibition term, $-g(\boldsymbol{r}_t)\cdot\boldsymbol{r}_t$ (Fig. 2*F*), causes a temporal decay in HD posterior certainty. Here, the function g's output increases linearly with bump amplitude in the HD population, resulting in an overall quadratic inhibition (*Methods*). Together with the leak, this quadratic inhibition approximates the nonlinear certainty decay $f(\kappa_t)/(2(\kappa_\phi+\kappa_v))$ in the circular Kalman filter, Eq. **2**, that accounts for both the diffusion prior "noise" $1/\kappa_\phi$ and the noise $1/\kappa_v$ induced by angular velocity observations, both of which are assumed known and constant. The approximation becomes precise in the limit of large posterior certainties κ_t.

With the appropriate parameter values, the amplitude of the bump decays slowly as long as new HD observations are unavailable, because global inhibition and leak work together to pull the bump amplitude slowly downward (Fig. 2*H*). This is by design: The circular Kalman filter tells us that certainty decays over time without a continuous stream of new HD observations. This situation differs from conventional ring attractors, whose bump amplitudes are commonly designed to rapidly decay to their stable (attractor) states. In a hypothetical network that perfectly implemented the circular Kalman filter, the bump amplitude would decay to zero. However, in our Bayesian ring attractor, which merely approximates the circular Kalman filter, the bump amplitude decays to a low but nonzero baseline amplitude (κ^*).

As an illustrative example, we simulated a network of 80 HD neurons (*Methods*). We let HD follow a random walk (diffusion on a circle), and we used noisy observations of the time derivative of HD (angular velocity) to modulate the asymmetric component of the connectivity matrix $\boldsymbol{W}(v_t)$. As HD changes, we rotate the cosine-shaped bump in the external input vector $\boldsymbol{I}_t^{ext}$, simulating the effect of a visual cue whose position on the retina depends on HD. This network exhibits a spatially localized bump whose position tracks HD, with an accuracy similar to that of the circular Kalman filter itself (Fig. 2*H*). Meanwhile, the amplitude of the bump accurately tracks the fluctuating HD posterior certainty in the circular Kalman filter, reflecting how noisy angular velocity and HD observations interact to modulate this certainty, Eq. **2**. When the visual cue is removed, the bump amplitude decays toward baseline (Fig. 2*H*). In the limit of infinitely many neurons, this type of network can be tuned to implement the circular Kalman filter exactly for sufficiently high HD certainties. What this simulation shows is that network performance can come close to benchmark performance even with a relatively small number of neurons (*SI Appendix*, Fig. S2).

Interestingly, when we vary the reliability of HD observations, we can observe two operating regimes in the network. When HD observations have high reliability, bump amplitude is high and accurately tracks HD certainty (κ_t). Thus, in this regime, the network performs proper Bayesian inference (Fig. 2*I*). Conversely, when HD observations have low reliability, bump amplitude is low but constant, because it is essentially pegged to its baseline value (the network's attractor state). In this regime, bump amplitude exaggerates the HD posterior certainty, and the network looks more like a conventional ring attractor. We will analyze these two regimes further in the next section.

Bayesian vs. Conventional Ring Attractors. Conventional ring attractors (1, 12, 39) are commonly designed to operate close to their attractor states, so that bump amplitude is nearly constant. This is not true of the Bayesian ring attractor described above, where bump amplitude varies by design. The motivation for this design choice was the idea that, if bump amplitude varies with certainty, the network's HD estimate would better match the true HD, because evidence integration would be closer to Bayes-optimal. Here, we show that this idea is correct.

Specifically, we measure the average accuracy of the network's HD encoding for different HD observation reliabilities for both the Bayesian ring attractor and a conventional ring attractor. We vary the HD observation information rate from highly unreliable, leading to almost random HD estimates (circKF inference accuracy close to zero in Fig. 3*B*), to highly reliable, leading to almost perfect HD estimates (circKF inference accuracy close to one), respectively. To model a conventional ring attractor, we use the same equations as we used for the Bayesian ring attractor, but we adjust the network connection strengths so that the bump amplitude decays to its stable baseline value very quickly (Fig. 3*A*). Specifically, we strengthen both local recurrent excitatory connections (w_{sym}) and global inhibition ($g(\boldsymbol{r}_t)$) while maintaining their balance, because their overall strengths are what controls the speed (β) of the bump's return to its baseline amplitude (κ^*) in the regime near κ^*, assuming no change in the cell-intrinsic leak time constant τ (*Methods*). With stronger overall connections, the bump amplitude decays to its stable baseline value more quickly. We then adjust the strength of global inhibition without changing the local excitation strength to maximize the accuracy of the network's HD encoding; note that this changes κ^* but not β. This yields a conventional ring attractor where the bump amplitude is almost always fixed at a stable value (κ^*), with κ^* optimized for maximal encoding accuracy. Even after this optimization of the conventional ring attractor, it does not rival the accuracy of the Bayesian ring attractor. The Bayesian attractor performs consistently better, regardless of the amount of information available to the network, i.e., the level of certainty in the new HD observations (Fig. 3*B*).

This performance difference arises because the conventional ring attractor does not keep track of the HD posterior's certainty. Ideally, the weight assigned to each HD observation depends on the current posterior certainty, as well as the reliability of the observation itself (Fig. 3*C*). A conventional ring attractor will assign more reliable observations a higher weight but does not take into account the posterior certainty. By contrast, the Bayesian ring attractor takes all these factors into account (Fig. 3*C*). The Bayesian ring attractor's performance drops to that of the conventional ring attractor only once HD observations become highly unreliable. In that regime, the Bayesian ring attractor operates close to its attractor state and thus stops accurately tracking HD certainty, making the attractor–network approximation to the circKF most apparent. Effectively, it becomes a conventional attractor network.

To obtain more insight into the effect of bump decay speed (β) on network performance, we can also simulate many versions of our network with different values of β, which we generate by varying the overall strength of balanced local recurrent excitatory (w_{sym}) and global inhibitory connections ($g(\boldsymbol{r}_t)$). We in turn vary the overall strength of global inhibition in order to find the best baseline bump amplitude (κ^*) for each value of β. The network with the best performance overall had a slow bump decay speed (low β), as expected (Fig. 3 *D* and *E*). While it featured similar performance to the Bayesian ring attractor, it had slightly different β and κ^* parameters. This is because the Bayesian ring attractor was analytically derived to well approximate the circKF for sufficiently high certainties, whereas the "best network" was numerically optimized to perform well on average. As the bump decay speed β increased further, performance dropped. However, this could be partially mitigated by increasing

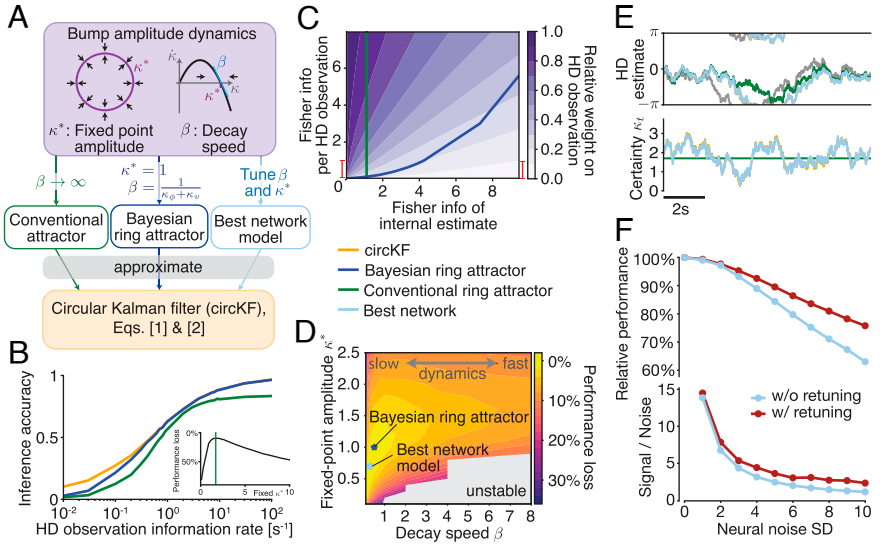

Fig. 3. Ring attractors with slow dynamics approximate Bayesian inference. (*A*) The ring attractor network in Eq. **4** can be characterized by fixed point amplitude κ^* and decay speed β, which depend on the network connectivities. Thus, the network can operate in different regimes: a regime, where the bump amplitude is nearly constant ("conventional attractor"), a regime where amplitude dynamics are tuned to implement a Bayesian ring attractor, or a regime with optimal performance ("best network," determined numerically). (*B*) HD estimation performance as measured by inference accuracy as a function of the HD observation information rate (as in Fig. 2*I*). κ^* for the "conventional" attractor was chosen to numerically maximize average accuracy, weighted by a prior across HD information rates (*Inset*/*Methods*). (*C*) The weight with which a single observation contributes to the HD posterior belief varies with informativeness of both the HD observations (Fisher information for 10-ms observation) and the current HD posterior (weight 1 = HD observation replaces HD estimate; 0 = HD observation leaves HD estimate unchanged). The update weight of the Bayesian attractor is close to optimal, visually indistinguishable from the circKF; not shown here, but *SI Appendix*, Fig. S3. Fisher information per 10-ms observation is directly related to the Fisher information rate, and the vertical red bar shows the equivalent range of information rate shown in panel *B*. (*D*) Overall inference performance loss (compared to a particle filter; performance measured by average inference accuracy, as in *B*, 0%: same average inference accuracy as a particle filter, 100%: random estimates), averaged across all levels of observation reliability (*Methods*) as a function of the bump amplitude parameters κ^* and β (only for $\kappa^* > 0$ and $\beta > 0$ as infinite network weights arise otherwise). (*E*) Simulated example trajectories of HD estimate/bump positions of HD estimate/bump positions (*Top*) and certainties/bump amplitudes (*Bottom*). The Bayesian ring attractor (not shown) is visually indistinguishable from the circKF and best network. (*F*) Relative performance (*Top*; 100% = inference accuracy without neural noise; performance measured as in panel *D*) and signal-to-noise ratio (*Bottom*; average κ_t divided by κ_t SD due to neural noise) drop with increasing neural noise (noise SD for additive noise in the network of 64 neurons). Retuning β and κ^* to maximize performance (purple vs. light blue = optimal parameters for noise-free network, panel *D*) reduces the drop in inference accuracy and S/N.

baseline bump amplitude (κ^*) to prevent overweighting of new observations.

We have seen that a slow bump decay (low β), i.e., the ability to deviate from the attractor state, is essential for uncertainty-related evidence weighting. That said, lower values of β are not always better. In the limit of very slow decay ($\beta \to 0$), bump amplitude would grow so large that new HD observations have little influence rendering the network nearly "blind" to visual landmarks. Conversely, in the limit of fast dynamics ($\beta \to \infty$), the network is highly responsive to new observations; however, it also has almost no ability to weight those new observations relative to other observations in the recent past. In essence, β controls the speed of temporal discounting in evidence integration. Ideally, the bump decay speed β should be matched to the expected speed at which stored evidence becomes outdated and thus loses its value, as controlled by κ_ϕ and κ_v.

To summarize, we can frame the distinction between a conventional ring attractor and a Bayesian ring attractor as a difference in the speed of the bump's decay to its stable state. In a conventional ring attractor, the bump decays quickly to its stable state, whereas in a Bayesian ring attractor, it decays slowly. Slow decay maximizes the accuracy of HD encoding because it allows the network to track its own internal certainty. Nonetheless, reasonable performance can be achieved even if the bump's decay is fast because a conventional ring attractor can still assign more informative observations a higher weight; it simply fails to assign the current HD estimate its proper weight.

The Impact of Neural Noise on Inference Accuracy. So far, we have assumed that the only sources of noise in our network are noisy angular velocity and HD observations. However, biological networks consist of neurons that are themselves noisy, resulting in another source of noise (40). What is the impact of that noise on inference accuracy?

If the network contains a sufficient number of similarly tuned neurons, their noise can be easily averaged out (37, 41), so that neural noise does not have a noticeable impact on the accuracy of inference. That said, for smaller networks, like those of insects (42), neural noise might significantly decrease inference accuracy. Indeed, simulating a network of 64 noisy neurons shows that, once the noise becomes sufficiently large, inference accuracy drops to below 70% of its noise-free value (Fig. 3*F*).

To better understand how neural noise perturbs inference, we derived its impact on the dynamics of the HD estimate μ_t and its certainty κ_t (*SI Appendix*). The derivations revealed that, irrespective of the form of the neural noise (additive, multiplicative, etc.), this noise has two effects. First, it causes an unbiased random diffusion of μ_t and, thus, an increasingly imprecise memory of the HD estimate. Second, it causes a positive drift and random diffusion of κ_t, which, if the drift is not accounted for, results in an overestimation of one's certainty and thus overconfidence in the HD estimate. These results mirror previous work that has shown a diffusion of μ_t in ring attractors close to the attractor state (41). We here show that such a diffusion persists even if the network operates far away from the attractor state, as is the case in our Bayesian ring attractor. While it is impossible to completely suppress the impact of neural noise, the derivations revealed that we can lessen its impact by retuning the network's connectivity strengths. Indeed, doing so reduced the drop in inference accuracy by about 35% when compared

to the network tuned to optimize noise-free performance (Fig. 3 *D* and *F*) and also boosted the network's signal-to-noise ratio (Fig. 3*F*). Lastly, our derivations show that the impact of noise vanishes once the network's population size becomes sufficiently large, in line with previous results (41). For example, increasing the network size four-fold would halve the effective noise's SD (assuming additive noise, *SI Appendix*). Overall, we have shown that neural noise causes a drop in performance that can in part be mitigated by retuning the network's connectivity strengths or by increasing its population size.

Tuning a Biological Ring Attractor for Bayesian Performance. Thus far, we have focused on model ring attractors with connection weights built from spatial cosine functions (Fig. 2*G*) because this makes the mathematical treatment of these networks more tractable. However, this raises the question of whether a biological neural network can actually implement an approximation of the circular Kalman filter, even without these idealized connection weights. The most well-studied biological ring attractor network is the HD system of the fruit fly *Drosophila melanogaster* (Fig. 4*A*) (17), and the detailed connections in this network have recently been mapped using large-scale electron microscopy connectomics (42). We therefore asked whether the motifs from this connectomic dataset—and, by extension, motifs that could be found in any biological ring attractor network—could potentially implement dynamic Bayesian inference.

To address this issue, we modeled the key cell types in this network (42–44) (HD cells, angular velocity cells, and global inhibition cells), using connectome data to establish the patterns of connectivity between each cell type (Fig. 4 *B*–*F* and *SI Appendix*, *Text*). We then analytically tuned the relative connection strengths between different cell types such that the dynamics of the bump parameters in the HD population implement an approximation of the circular Kalman filter. We also added a nonlinear element in the global inhibition layer as this is required to approximate the circular Kalman filter. We found that this network achieves a HD encoding accuracy which is indistinguishable from that of our idealized Bayesian ring attractor network (Fig. 4 *G* and *H*). Thus, even when we use connectome data to incorporate biological constraints on the network, the network is still able to implement dynamic Bayesian inference.

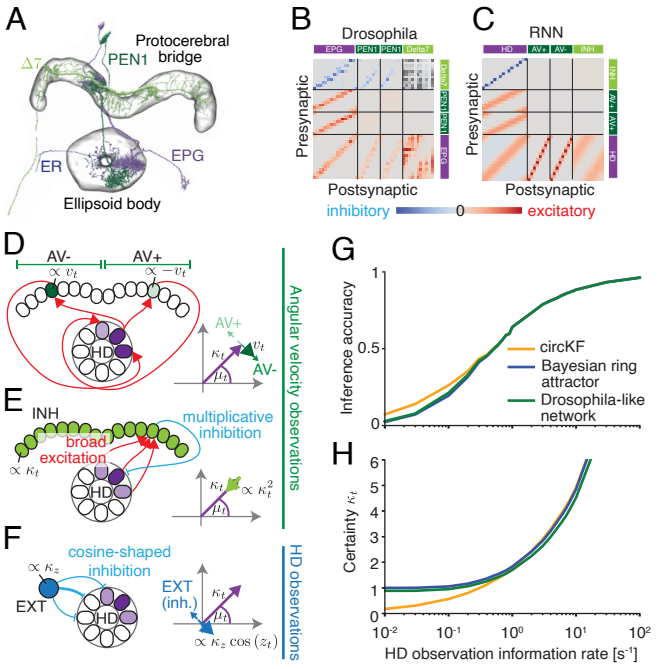

Fig. 4. A Drosophila-like network implementing the circular Kalman filter. (*A*) Cell types in the *Drosophila* brain that could contribute to implementing the circular Kalman filter. (*B*) Connectivity between EPG, Δ7, and PEN1 neurons, as recovered from the hemibrain:v1.2.1 database (43). Neurons were sorted by their spatial position as a proxy for their preferred HD. The total number of synaptic connections between each cell pair was taken to indicate the functional connection strength between these cells. The polarity of Δ7 → Δ7 connections is unknown, and therefore, these connections are omitted. (*C*) The connectivity profile of a recurrent neural network (RNN) (Fig. 2 *D*–*F*) that implements an approximate circKF is strikingly similar to the connectivity of neurons in the *Drosophila* HD system. To avoid confusion with actual neurons, we refer to the neuronal populations in this idealized RNN as head direction (HD), angular velocity (AV+ and AV-, in reference to the two hemispheres), inhibitory (INH), and external input (EXT) populations. (*D*) Differential activation of AV populations (left/right: high/low) across hemispheres as well as shifted feedback connectivity from AV to HD populations effectively implements the asymmetric (or shifted) connectivity needed to turn the bump position (here, clockwise shift for anticlockwise turn). (*E*) Broad excitation of the INH population by the HD population, together with a one-to-one multiplicative interaction between INH and HD population, implements the quadratic decay of the bump amplitude needed for the reduction in certainty arising from probabilistic path integration. (*F*) External input is mediated by inhibiting HD neurons with the preferred direction opposite to the location of the HD observation, effectively implementing a vector sum of belief with HD observation. (*G* and *H*) The inference accuracy of the *Drosophila*-like network is indistinguishable from that of the Bayesian ring attractor. Inference accuracy, certainty, and HD observation information rate are measured as for Fig. 2*I*.

Discussion

Uncertainty can affect navigation strategy (45, 46), spatial cue integration (47, 48), and spatial memory (49). This provides a motivation for understanding how uncertainty is represented in the neural networks that encode and store spatial variables for navigation. There is good reason to think that these networks are built around attractors. Thus, it is crucial to understand how attractors in general—and ring attractors in particular—might track uncertainty in spatial variables like head direction.

In this study, we have shown that a ring attractor can track uncertainty by operating in a dynamic regime away from its stable baseline states (its attractor states). In this regime, bump amplitude can vary because local excitatory and global inhibitory connections in the ring attractor are relatively weak. By contrast, stronger overall connections produce a more conventional ring attractor that operates closer to its attractor states. Because the "Bayesian" ring attractor has a variable bump amplitude, bump amplitude grows when recent HD observations have been more reliable; in this situation, the network automatically ascribes more weight to its current estimate, relative to new evidence. Importantly, we have shown that nearly optimal evidence weighting does not require exact tuning of the network connections. Indeed, even when we used connectome data to implement a network with realistic biological connectivity constraints, the network could still support near-optimal evidence weighting.

A key element of our approach is that bump amplitude is used to represent the internal certainty of the system's Bayesian HD posterior belief. In our framework, internal certainty determines the weight ascribed to new evidence, relative to past evidence. As such, the representation of internal certainty plays a crucial role in maximizing the accuracy of our Bayesian ring attractor. This stands in contrast to recent network models of the HD system that do not encode internal certainty, even though they weigh HD observations in proportion to their reliabilities (50, 51). Notably, our network also automatically adjusts its cue integration weights to perform close-to-optimal Bayesian inference for HD observations of varying reliability. Recent work (52) described how observations of a circular variable

(such as HD) could be integrated across brief time periods, provided that these observations all have the same reliability; however, this work considered neither the problem of integrating observations of differing reliability nor the role of angular velocity observations. Moreover, the network described in that study operated below the fixed point of the bump amplitude, and therefore, it can only correctly weight incoming observations over a short period before reaching the fixed point.

Another important element of our approach was that we benchmarked our network model against a rigorous normative standard, the circular Kalman filter, which was derived analytically in ref. 31 and described here in terms of intuitive vector operations. Being able to rely on the circular Kalman filter was important because it allowed us to analytically derive the proper parameter values of our network model, so that the network's estimate matched the estimate of an ideal observer. A remarkable property of the circular Kalman filter is that new HD observations will actually decrease certainty if they conflict strongly with the current estimate. This is not a property of a standard (noncircular) Kalman filter or a neural network designed to emulate it (33). The power of conflicting evidence to decrease certainty is particular to the circular domain. Our Bayesian ring attractor network automatically reproduces this important aspect of the circular Kalman filter. Of course, the circular Kalman filter has applications beyond neural network benchmarking, as the accurate estimation of orientation or any other periodic variable has broad applications in the field of engineering.

When adequately tuned, our network can implement a persistent working memory of a circular variable, as, for example, the orientation of a visual stimulus in a visual working memory task. Recent models for such tasks attribute memory recall errors to the stochastic emission of a limited number of spikes (7, 53). As in our network, neural noise can be averaged out once the network has a sufficiently large number of neurons; for this reason, memory errors can be attributed only to the noise of individual neurons in small networks with few neurons. Significant memory errors in larger Bayesian ring attractors thus have to result from other sources of noise, such as imperfect connectivity weights, or correlated input noise from, e.g., shared inputs, that fundamentally limits the amount of information that these inputs provide to the network (54).

In the brain's HD system, the internal estimate of HD is based on not only HD observations (visual landmarks, etc.) but also angular velocity observations. The process of integrating these angular velocity observations over time is called angular path integration. Angular path integration is inherently noisy, and therefore, uncertainty will grow progressively when HD observations are lacking. Our Bayesian ring attractor network is notable in explicitly treating angular path integration as a problem of probabilistic inference. Each angular velocity observation has limited reliability, and this causes the bump amplitude to decay in our network as long as HD observations are absent, in a manner that well approximates the certainty decay of an ideal observer. In this respect, our network differs from previous investigations of ring attractors having variable bump amplitude (55).

Our work makes several testable predictions. First, we predict that the HD system should contain the connectivity motifs required for a Bayesian ring attractor. Our analysis of Drosophila brain connectome data supports this idea; we expect similar network motifs to be present in the HD networks of other animals, such as that of mice (15, 19), monkeys (56), humans (57), or bats (58). In the future, it will be interesting to determine whether synaptic inhibition in these networks is nonlinear, as predicted by our models.

Second, we predict that bump amplitude in the HD system should vary dynamically, with higher amplitudes in the presence of reliable external HD cues, such as salient visual landmarks. In particular, when bump amplitude is high, the bump's position should be less sensitive to the appearance of new external HD cues. Notably, an experimental study from the mouse HD system provides some initial support for these predictions (19). This study found that the amplitude of population activity in HD neurons (what we call bump amplitude) increases in the presence of a reliable visual HD cue. Bump amplitude also varied spontaneously when all visual cues were absent (in darkness); intriguingly, when the bump amplitude was higher in darkness, the bump position was slower to change in response to the appearance of a visual cue, suggesting a lower sensitivity to the cue. In the future, more experiments will be needed to clarify the relationship between bump amplitude, certainty, and cue integration. In particular, it is puzzling that multiple studies (16, 18, 19, 24, 59, 60) have found that bump amplitude increases with angular velocity, as higher angular velocities should not increase certainty.

In the future, more investigation will be needed to understand evidence accumulation on longer timescales. The circular Kalman filter is a recursive estimator: At each time step, it considers only the observer's internal estimate from the previous time step as well as the current observation of new evidence. However, when the environment changes, it would be useful to use a longer history of past observations (and past internal estimates) to readjust the weight assigned to the changing sources of evidence. Available data suggest that Hebbian plasticity can progressively strengthen the influence of the external sensory cues that are most reliably correlated with HD (17, 49, 61, 62). The interaction of Hebbian plasticity with attractor dynamics could provide a mechanism for extending statistical inference to longer timescales (13, 63–69).

In summary, our work shows how ring attractors could implement dynamic Bayesian inference in the HD system. Our results have significance beyond the encoding of head direction—e.g., they are potentially relevant for the grid cell ensemble, which appears to be organized around ring attractors even though it encodes linear rather than circular variables. Moreover, our models could apply equally to any brain system that needs to compute an internal estimate of a circular or periodic variable, such as visual object orientation (6, 70) or circadian time. More generally, our results demonstrate how canonical network motifs, like those common in ring attractor networks, can work together to perform close-to-optimal Bayesian inference, a problem with fundamental significance for neural computation.

Materials and Methods

Ideal Observer Model: The Circular Kalman Filter. Our ideal observer model–the circular Kalman filter (circKF) (31)–performs dynamic Bayesian inference for circular variables. It computes the posterior belief of an unobserved (true) HD $\phi_t \in [-\pi, \pi]$ at each point in time t, conditioned on a continuous stream of noisy angular velocity observations $v_{0:t} = \{v_0, v_{\mathrm{d}t}, \dots v_t\}$ with $v_\tau \in \mathbb{R}$, and HD observations $z_{0:t} = \{z_0, z_{\mathrm{d}t}, \dots z_t\}$ with $z_\tau \in [-\pi, \pi]$. In contrast to the discrete-time description in Results, we here provide a continuous-time formulation of the filter. Specifically, we assume that these observations are generated from the true angular velocity $\dot{\phi}_t$ and HD ϕ_t, corrupted by zero-mean noise at each point in time, via

$$v_t|\dot{\phi}_t \sim \boldsymbol{\mathcal{N}}\left(\dot{\phi}_t, \frac{1}{\kappa_v\,\mathrm{d}t}\right), \quad [5]$$

$$z_t|\phi_t \sim \boldsymbol{\mathcal{VM}}\left(\phi_t, \kappa_z\,\mathrm{d}t\right). \quad [6]$$

Here, $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian with mean μ and variance σ^2, $\mathcal{VM}(\mu, \kappa)$ denotes a von Mises distribution of a circular random variable with mean μ and precision κ, and κ_v and κ_z quantify reliabilities of angular velocity and HD observations, respectively. Note that as $dt \to 0$, the precision values of both angular velocity and HD observations approach 0, in line with the intuition that reducing a time-step size dt results in more observations per unit time, which should be accounted for by less precision per observation to avoid "oversampling" (*SI Appendix* for a subtlety for how κ_z scales with time).

To support integrating information over time, the model assumes that current HD ϕ_t depends on past HD ϕ_{t-dt}. Specifically, in the absence of further evidence, the model assumes that HD diffuses on a circle,

$$\phi_t|\phi_{t-dt} \sim \mathcal{N}\left(\phi_{t-dt}, \frac{dt}{\kappa_\phi}\right) \mod 2\pi, \quad [7]$$

with a diffusion coefficient that decreases with κ_ϕ.

The circKF in Eqs. **1** and **2** assumes that the HD posterior belief can be approximated by a von Mises distribution with time-dependent mean μ_t and certainty κ_t, i.e. $p(\phi_t|v_{0:t}, z_{0:t}) \approx \mathcal{VM}(\phi_t; \mu_t, \kappa_t)$. Such an approximation is justified if the posterior is sufficiently unimodal and can, for instance, be compared to a similar approximation employed by extended Kalman filters for noncircular variables.

An alternative parametrization of the von Mises distribution to its mean μ_t and precision κ_t is its natural parameters, $\boldsymbol{x}_t = (\kappa_t \cos\mu_t, \kappa_t \sin\mu_t)^T$. Geometrically, the natural parameters can be interpreted as the Cartesian coordinates of a "HD belief vector" and (μ_t, κ_t) as its polar coordinates (Fig. 1*B*). As we show in SI, the natural parameter parametrization makes including HD observations in the circKF straightforward. In fact, it becomes a vector addition. In contrast, including angular velocity observations is mathematically intractable, such that the circKF relies on an approximation method called projection filtering (30) to find closed-form dynamic expressions for posterior mean and certainty (see ref. 31 for technical details and *SI Appendix* for a more accessible description of the circKF).

Taken together, the circKF for the model specified by Eqs. **5**–**7** reads

$$d\mu_t = \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t\, dt + \frac{\kappa_z}{\kappa_t} \sin(z_t - \mu_t) dt, \quad [8]$$

$$d\kappa_t = -\frac{f(\kappa_t)}{2(\kappa_\phi + \kappa_v)} \kappa_t dt + \kappa_z \cos(z_t - \mu_t) dt, \quad [9]$$

which is the continuous-time equivalent to Eqs. **1** and **2** in Results and where $f(\kappa)$ is a monotonically increasing nonlinear function,

$$f(\kappa) = \frac{A(\kappa)}{\kappa - A(\kappa) - \kappa A(\kappa)^2}, \quad \text{with } A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}, \quad [10]$$

and $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of the first kind of order 0 and 1, respectively.

For a sufficiently large κ (i.e., high certainty), the nonlinearity $f(\kappa)$ approaches the linear function, $f(\kappa) \to 2\kappa - 2$. In our quadratic approximation, we thus replace the nonlinearity with a quadratic decay:

$$d\kappa_t = -\frac{1}{\kappa_\phi + \kappa_v}\left(\kappa_t^2 - \kappa_t\right) dt + \kappa_z \sin(z_t - \mu_t) dt, \quad [11]$$

which well approximates the circKF in the high certainty regime.

Network Model. We derived a rate-based network model that implements (approximations of) the circKF, by encoding the von Mises posterior parameters in activity $\boldsymbol{r}_t \in \mathbb{R}^N$ of a neural population with N neurons. Thereby, we focused on the simplest kind of network model that supports such an approximation, which is given by Eq. **4**. In that equation, τ is the cell-intrinsic leak time constant, $g : \mathbb{R}^N \to \mathbb{R}_+$ is a scalar nonlinearity, and the elements of $\boldsymbol{r}_t$ are assumed to be ordered by the respective neuron's preferred HD, $\phi_1, \ldots, \phi_N$ (Eq. **3**). We decomposed the recurrent connectivity matrix into $\boldsymbol{W}(v_t) = w^{\text{const}} \frac{1}{N} \mathbf{1}\mathbf{1}^T + w^{\text{sym}} \boldsymbol{W}^{\cos} + w^{\text{asym}}(v_t) \boldsymbol{W}^{\sin}$, where $\mathbf{1}\mathbf{1}^T$ is a matrix filled with 1's, and $\boldsymbol{W}^{\cos}$ and $\boldsymbol{W}^{\sin}$ refer to cosine- and sine-shaped connectivity profiles (Fig. 2*G*). The network's circular symmetry makes the entries of these matrices depend only on the relative distance in preferred HD, and the entries are given by $W_{ij}^{\cos} = \frac{2}{N}\cos(\phi_i - \phi_j)$, and $W_{ij}^{\sin} = \frac{2}{N}\sin(\phi_i - \phi_j)$. The scaling factor $\frac{2}{N}$ was chosen to facilitate matching our analytical results from the continuum network to the network structure outlined here. We further considered a cosine-shaped external input of the form $I_t^{\text{ext}}(\phi_i) = I_t(dt)\cos(\Phi_t - \phi_i)$ that is peaked around an input location Φ_t. Here, $I_t(dt)$ denotes the input pattern in the infinitesimal time bin dt.

As described in *Results*, we assume the population activity $\boldsymbol{r}_t$ to encode the HD belief parameters μ_t and κ_t in the phase and amplitude of the activity's first Fourier component. As we show in *SI Appendix*, the described network dynamics thus lead to the following dynamics of the cosine-profile parameters μ_t and κ_t:

$$d\mu_t = w^{\text{asym}}(v_t)\, dt + \frac{I_t}{\kappa_t} \sin(\Phi_t - \mu_t), \quad [12]$$

$$d\kappa_t = \left(w^{\text{sym}} - \frac{1}{\tau}\right)\kappa_t\, dt - g(\boldsymbol{r}_t)\kappa_t\, dt + I_t \cos(\Phi_t - \mu_t). \quad [13]$$

To derive these dynamics, we make the following three assumptions. First, we assume the network to be rate based. Second, our analysis assumes a continuum of neurons, i.e., $N \to \infty$. For numerical simulations, and the network description below, we used a finite-sized network of size N that corresponds to a discretization of the continuous network. *SI Appendix*, Fig. S2 demonstrates only a very weak dependence of our results on the exact number of neurons in the network. Third, our analysis and simulations focused on the first Fourier mode of the bump profile and is thus independent of the exact shape of the profile (as long as Eq. **3** holds).

Network parameters for Bayesian inference. Having identified how the dynamics of μ_t and κ_t encoded by the network, Eqs. **12** and **13** depend on the network parameters, we now tuned these parameters to match these dynamics to those of the mean and certainty of the circKF, Eqs. **8** and **9**. Here, we first do so to achieve an exact match to the circKF, without the quadratic approximation. After that, we describe the quadratic approximation that is used in the main text and leads to the Bayesian ring attractor network. Specifically, an exact match to the circKF requires the following network parameters:

- Asymmetric recurrent connectivities are modulated by angular velocity observations, $w^{\text{asym}}(v_t) = \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t$, which shifts the activity profile without changing its amplitude (12, 13).
- HD observations z_t are represented as the peak position Φ_t of a cosine-shaped external input whose amplitude is modulated by the reliability of the observation, i.e., $I_t = \kappa_z\, dt$. The inputs might contain additional Fourier modes (e.g., a constant baseline), but those do not affect the dynamics in Eqs. **12** and **13**.
- The symmetric component of the recurrent excitatory input needs to exactly balance the internal activity decay, i.e., $w^{\text{sym}} - \frac{1}{\tau} = 0$.
- The decay nonlinearity is modulated by the reliability of the angular velocity observations and is given by $g(\boldsymbol{r}_t) = \frac{1}{2(\kappa_\phi + \kappa_v)} f\left(\kappa(\boldsymbol{r}_t)\right)$, where $f(\cdot)$ equals the nonlinearity that governs the certainty decay in the circKF, Eq. **10**. This can be achieved, for example, through interaction with an inhibitory neuron (or a pool of inhibitory neurons) with activation function $f(\cdot)$ that computes the activity bump's amplitude $\kappa(\boldsymbol{r}_t)$.

A network with these parameters is not an attractor network, as its activity decays to zero in the absence of external inputs.

To arrive at the Bayesian ring attractor, we approximate the decay nonlinearity by a quadratic approximation that takes the form $g(\boldsymbol{r}_t)\boldsymbol{r}_t \to w^{\text{quad}}\left(\frac{\pi}{N}\sum_{i=1}^N [r_t^{(i)}]_+\right) \cdot \boldsymbol{r}_t$, where $[\cdot]_+$ denotes the rectification nonlinearity. The resulting recurrent inhibition can be shown to be quadratic in the amplitude κ_t and has the further benefit of introducing an attractor state at a positive bump amplitude (below). In the large population limit, $N \to \infty$, this

leads to the amplitude dynamics (*SI Appendix* for derivation)

$$\mathrm{d}\kappa_t = \left(w^{\mathrm{sym}} - \frac{1}{\tau}\right)\kappa_t\,\mathrm{d}t - w^{\mathrm{quad}}\kappa_t^2\mathrm{d}t + I_t\cos(\Phi_t - \mu_t). \qquad [14]$$

The dynamics of the phase μ_t does not depend on the form of $g(\cdot)$ and thus remains to be given by Eq. **12**. If we set the network parameters to $w^{\mathrm{quad}} = \frac{1}{\kappa_\phi+\kappa_v}$ and $w^{\mathrm{sym}} - \frac{1}{\tau} = \frac{1}{\kappa_\phi+\kappa_v}$, while sensory input, i.e., angular velocity v_t and HD observations z_t, enters in the same way as before, the network implements the quadratic approximation to the circKF, Eqs. **8** and **11**.

General ring-attractor networks with fixed point κ^* and decay speed β. In the absence of HD observations ($I_t = 0$), the amplitude dynamics in Eq. **14** has a stable fixed point at $\kappa^* = \frac{w^{\mathrm{sym}}-1/\tau}{w^{\mathrm{quad}}}$ and no preferred phase, making it a ring-attractor network. Linearizing the κ_t dynamics around this fixed point reveals that it is approached with decay speed $\beta = w^{\mathrm{sym}} - \frac{1}{\tau}$. Therefore, we can tune the parameters to achieve a particular fixed point κ^* and decay speed β by setting $w^{\mathrm{sym}} = \beta + 1/\tau$ and $w^{\mathrm{quad}} = \frac{\beta}{\kappa^*}$. A large value of β requires increasing both w^{sym} and w^{quad}, yields faster dynamics, and thus indicates more rigid attractor dynamics. In the limit of $\beta \to \infty$, the attractor becomes completely rigid in the sense that, upon any perturbation, it immediately moves back to its attractor state. In the main text, we assume conventional ring attractors to operate close to this rigid regime. For the Bayesian ring attractor, we find $\kappa^* = 1$ and $\beta = \frac{1}{\kappa_\phi+\kappa_v}$. Further, in our simulations in Fig. 3, we explored network dynamics with a range of κ^* and β values by adjusting network parameters accordingly.

Assessing the Impact of Neural Noise on Inference Accuracy. In *SI Appendix*, we show that neural noise results in an unbiased diffusion of μ_t and a diffusion and positive drift of κ_t. We assessed the impact of this noise on inference accuracy by simulating a network with $N = 64$ neurons and κ^* and β tuned to maximize noise-free inference accuracy ("Best network model" in Fig. 3*D*) and by adding Gaussian zero-mean white noise with variance $\sigma_{nn}^2\delta t$ in each time step δt to each neuron, for different levels of σ_{nn} (Fig. 3*F*, light blue lines). We computed the signal-to-noise ratio for each simulation as the average κ_t divided by the diffusion noise SD $\sigma_{nn}\sqrt{2/N}$ that additive neural noise causes in κ_t (*SI Appendix* for derivation). The impact of this noise can be reduced by retuning the network's connectivity strengths. We did so for each neural noise magnitude separately by a grid search over κ^* and β (*SI Appendix*, Fig. S4), similar to the previous section (Fig. 3*F*, purple lines).

Drosophila-like multipopulation network. We extended the single population network dynamics, Eq. **4**, to encompass five populations: a HD population, which we designed to track HD estimate and certainty with its bump parameter dynamics; two angular velocity populations (AV+ and AV-), which are tuned to HD and are differentially modulated by angular velocity input; an inhibitory population (INH); and a population that mediates external input (EXT), corresponding to HD observations. The network parameters were tuned such that the activity profile in the HD population tracks the dynamics of the circKF quadratic approximation, in the same way as for the single-population network, Eq. **4**. To limit the degrees of freedom, we further constrained the connectivity structure between HD and AV+/- and INH populations by the known connectome of the *Drosophila* HD system (hemibrain dataset in ref. 42) and tuned only across-population connectivity weights. For further details on the network dynamics and with- and across-population connectivity weights, please consult *SI Appendix*.

Simulation Details.

Numerical integration. Our simulations in Figs. 2–4 used artificial data that matched the assumptions underlying our models. In particular, the "true" HD ϕ_t followed a diffusion on the circle, Eq. **7**, and observations were drawn at each point in time from Eqs. **5** and **6**. To simulate trajectories and observations, we used the Euler–Maruyama scheme (71), which supports the numerical integration of stochastic differential equations. Specifically, for a chosen discretization time step Δt, this scheme is equivalent to drawing trajectories and observations from Eqs. **7**, **5**, and **6** directly while substituting $dt \to \Delta t$. The same time-discretization scheme was used to numerically integrate the SDEs of the circKF, Eqs. **8** and **9** and its quadratic approximation, Eq. **11**.

Performance measures. To measure performance, in Figs. 2*I*, 3 *B* and *D* and 4 *G* and *H*, we computed the circular average distance (72) of the estimate μ_T from the true HD ϕ_T at the end of a simulation of length $T = 20$ from $P = 5'000$ simulated trajectories by $m_1 = \frac{1}{P}\sum_{k=1}^{P}\exp\left(i\left(\mu_T^{(k)} - \phi_T^{(k)}\right)\right)$. The absolute value of the imaginary-valued circular average, $0 \le |m_1| \le 1$, is unitless and denotes an empirical accuracy (or "inference accuracy") and thus measures how well the estimate μ_T matches the true HD ϕ_T. Here, a value of 1 denotes an exact match. The reported inference accuracy is related to the circular variance via $\mathrm{Var}_{circ} = 1 - |m_1|$. In *SI Appendix*, Fig. S5, we provide histograms with samples $\mu_T - \phi_T$ with different numerical values of $|m_1|$ to provide some intuition for the spread of estimates for a given value of the performance measure.

We estimated performance through such averages for a range of HD observation information rates in Figs. 2*I*, 3*B* and 4*G*. This information rate is a simulation time-step size-independent quantity, which measures the Fisher information that HD observations provide about true HD per unit time. For individual HD observations of duration dt, Eq. **6**, this Fisher information approaches $I_{z_t}(\phi_t) \to (\kappa_z\mathrm{d}t)^2/2$ with $\mathrm{d}t \to 0$ (31, Theorem 2). Per unit time, we observe $1/\mathrm{d}t$ independent observations, leading to a total Fisher information (or information rate) of $\gamma_z = \kappa_z^2 dt/2$. As in simulations, γ_z needs to remain constant with changing Δt to avoid increasing the amount of provided information, the HD observation reliability κ_z needs to change with the size of simulation time-step size Δt. To keep our plots independent of this time-step size, we thus plot performance as a function of the HD observation information rate rather than κ_z. For the inset of Fig. 3*B*, and for Figs. 3 *D* and *F*, we additionally performed a grid search over the fixed-point κ^* (Fig. 3 *B*, inset) or both the fixed-point κ^* and of the decay speed β (Figs. 3 *D* and *F*). For each setting of κ^* and β, we assessed the performance by computing an average over this performance for a range of HD observation information rates, weighted by how likely each observation reliability is assumed to be a priori. The latter was specified by a log-normal prior, $p(\gamma_z) = \mathrm{Lognormal}(\gamma_z|\mu_{\gamma_z}, \sigma_{\gamma_z}^2)$, favoring intermediate reliability levels. We chose $\mu_{\gamma_z} = 0.5$ and $\sigma_{\gamma_z}^2 = 1$ for the prior parameters, but our results did not strongly depend on this parameter choice. The performance loss shown in Fig. 3*D* also relied on such a weighted average across information rates γ_z for a particle filter benchmark (PF, SI for details). The loss itself was then defined as $1 - \frac{\text{Performance}}{\text{Performance PF}}$.

Update weights for HD observations. In Fig. 3*C*, we computed the weight with which a single HD observation with $|z_t - \mu_t| = 90°$ changes the HD estimate. We defined this weight as the change in HD estimate, normalized by the value of the maximum possible change, $w = \frac{\Delta\mu_t}{\pi} = \frac{1}{\pi}\tan^{-1}\frac{\kappa_z\,\mathrm{d}t}{\kappa_t}$. To make units intuitively comparable between the two axes, we chose to scale the y-axis in units of Fisher information of a single HD observation of duration $\Delta t = 10ms$, given by $I_{z_t}(\phi_t) = \gamma_z\,\Delta t$ where $\gamma_z = \kappa_z^2\,\Delta t/2$. Thus, the weight is plotted as a function of the Fisher information of a single HD observation (how reliable is the observation?) and the Fisher information of the current HD posterior belief (how certain is the current estimate?), which is given by $I_{\mu_t,\kappa_t}(\phi_t) = \kappa_t\frac{I_1(\kappa_t)}{I_0(\kappa_t)}$ (31).

Simulation parameters. In our network simulations, we set the leak time constant τ to an arbitrary, but nonzero, value. Effectively, this resulted in a cosine-shaped activity profile. Note that by setting higher-order recurrent connectivities accordingly, other profile shapes could be realized, without affecting the validity of our analysis from the neural activity vector $\boldsymbol{r}_t$, we retrieved the natural parameters $\boldsymbol{x}_t$ with a decoder matrix $A = (\cos(\boldsymbol{\phi}), \sin(\boldsymbol{\phi}))^T$, such that $\boldsymbol{x}_t = A\cdot\boldsymbol{r}_t$, and subsequently computed the position of the bump by $\phi_t = \arctan 2(x_2, x_1)$ and the encoded certainty (length of the population vector) by $\kappa_t = \sqrt{x_1^2 + x_2^2}$.

In all our simulations, times are measured in units of inverse diffusion time constant κ_ϕ, where we set $\kappa_\phi = 1s$ for convenience. We used the following simulation parameters. For Fig. 2*H*, we used $\kappa_v = 2$ and information

rate of HD observations of $\gamma_z = 10/s$ (equaling $\kappa_z \approx 45$; during "Visual cue" period) and $\kappa_z = 0$ (during "Darkness" period). For Figs. 2*I* and 3 *B* and *D* we used $\kappa_v = 1$, $T = 20$, and averaged results over $P = 5,000$ simulation runs. For Fig. 3*E*, we used $\kappa_v = 1$, information rate of $\gamma_z = 1/s$ (equaling $\kappa_z \approx 14$), $T = 10$. In the network simulations in Fig. 2 *H* and *I* and Fig. 3 *B* and *D*, we translated these parameters into network connectivity parameters according to our analytical results in *SI Appendix*, section 3B. Without loss of generality, we set all connectivity parameters that are not explicitly mentioned, to zero (including w^{const}). Please consult *SI Appendix* for details on the *Drosophila* network simulation parameters. We used $\Delta t = 0.01$ for all simulations.

Trajectory simulations and general analyses were performed on a MacBook Pro (Mid 2019) running 2.3 GHz 8-core Intel Core i9. Parameter scans were run on the Harvard Medical School O_2 HPC cluster. For all our simulations, we used Python 3.9.1 with NumPy 1.19.2.

Data, Materials, and Software Availability. Computer simulations and data analysis were performed with custom Python code, which has been deposited in Zenodo, DOI: 10.5281/zenodo.7615975.

ACKNOWLEDGMENTS. We would like to thank Habiba Noamany for assisting us in navigating the neuPrint database and for informed comments on the manuscript. We would further like to thank Johannes Bill and Albert Chen for discussions and feedback on the manuscript, Philipp Reinhard for going on a typo hunt in *SI Appendix*, and the entire Drugowitsch lab for valuable and insightful discussions. The work was funded by the NIH (R34NS123819; J.D. & R.I.W), the James S. McDonnell Foundation (Scholar Award #220020462; J.D.), the Swiss NSF (grant numbers P2ZHP2 184213 and P400PB 199242; A.K.), and a Grant in the Basic and Social Sciences by the Harvard Medical School Dean's Initiative award program (J.D. & R.I.W.).

1. X. J. Wang, Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–463 (2001).
2. A. Compte, Computational and in vitro studies of persistent activity: Edging towards cellular and synaptic mechanisms of working memory. *Neuroscience* **139**, 135–151 (2006).
3. D. Hansel, H. Sompolinsky, "Modeling feature selectivity in local cortical circuits" in *Methods in Neuronal Modeling: From Ions to Networks, Computational Neuroscience Series*, C. Koch, I. Segev, Eds. (1998), p. 69.
4. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci U.S.A.* **79**, 2554–2558 (1982).
5. R. L. Rademaker, C. H. Tredway, F. Tong, Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *J. Vision* **12**, 21 (2012).
6. H. H. Li, T. C. Sprague, A. H. Yoo, W. J. Ma, C. E. Curtis, Joint representation of working memory and uncertainty in human cortex. *Neuron* **109**, 3699–3712.e6 (2021).
7. P. Bays, S. Schneegans, W. J. Ma, T. Brady, Representation and computation in working memory. PsyArXiv (2022).
8. M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
9. A. T. Piet, A. E. Hady, C. D. Brody, A. El Hady, C. D. Brody, Rats adopt the optimal timescale for evidence integration in a dynamic environment. *Nat. Commun.* **9**, 1–12 (2018).
10. C. R. Fetsch, A. H. Turner, G. C. DeAngelis, D. E. Angelaki, Dynamic reweighting of visual and vestibular cues during self-motion perception. *J. Neurosci.* **29**, 15601–15612 (2009).
11. J. J. Knierim, K. Zhang, Attractor dynamics of spatially correlated neural activity in the limbic system. *Ann. Rev. Neurosci.* **35**, 267–285 (2012).
12. K. Zhang, Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *J. Neurosci.* **16**, 2112–2126 (1996).
13. 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*, G. Tesauro, D. Touretzky, T. Leen, Eds. (MIT Press, 1994), vol. 7.
14. A. D. Redish, A. N. Elga, D. S. Touretzky, A coupled attractor model of the rodent head direction system. *Network: Comput. Neural Syst.* **7**, 671–685 (1996), 10.1088/0954-898X/7/4/004.
15. A. Peyrache, M. M. Lacroix, P. C. Petersen, G. Buzsáki, Internally organized mechanisms of the head direction sense. *Nat. Neurosci.* **18**, 569–575 (2015).
16. J. D. Seelig, V. Jayaraman, Neural dynamics for landmark orientation and angular path integration. *Nature* **521**, 186–191 (2015).
17. T. D. Kim, M. Kabir, J. I. Gold, Coupled decision processes update and maintain saccadic priors in a dynamic environment. *J. Neurosci.* **37**, 3632–3645 (2017).
18. D. Turner-Evans *et al.*, Angular velocity integration in a fly heading circuit. *eLife* **6**, e23496 (2017).
19. Z. Ajabi, A. T. Keinath, X. X. Wei, M. P. Brandon, Population dynamics of the thalamic head direction system during drift and reorientation (bioRxiv, Tech. rep. 2021).
20. Si. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* **27**, 77–87 (1977).
21. B. Ermentrout, Neural networks as spatio-temporal pattern-forming systems. *Rep. Progr. Phys.* **61**, 353–430 (1998).
22. S. Heinze, A. Narendra, A. Cheung, Principles of insect path integration. *Curr. Biol.?: CB* **28**, R1043–R1058 (2018).
23. X. Xie, R. H. Hahnloser, H. S. Seung, Double-ring network model of the head-direction system. *Phys. Rev. E - Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **66**, 9–9 (2002).
24. J. Green *et al.*, A neural circuit architecture for angular integration in Drosophila. *Nature* **546**, 101–106 (2017).
25. D. C. Knill, A. Pouget, The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
26. G. P. Dehaene, R. Coen-Cagli, A. Pouget, Investigating the representation of uncertainty in neuronal circuits. *PLOS Comput. Biol.* **17**, e1008138 (2021).
27. R. E. Kalman, A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.* **82**, 35–45 (1960).
28. R. E. Kalman, R. S. Bucy, New results in linear filtering and prediction theory. *J. Basic Eng.* **83**, 95–108 (1961).
29. G. Kurz, F. Pfaff, U. D. Hanebeck, "Kullback-Leibler Divergence and moment matching for hyperspherical probability distributions" in *2016 19th International Conference on Information Fusion (FUSION). No. July* (2016), pp. 2087–2094.
30. D. Brigo, B. Hanzon, F. Le Gland, Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli* **5**, 495–534 (1999).
31. A. Kutschireiter, L. Rast, J. Drugowitsch, Projection filtering with observed state increments with applications in continuous-time circular filtering. *IEEE Trans. Signal Proc.* **70**, 686–700 (2022).
32. R. F. Murray, Y. Morgenstern, Cue combination on the circle and the sphere. *J. Vision* **10**, 15–15 (2010).
33. R. Wilson, L. Finkel, A neural implementation of the Kalman filter. *Adv. Neural Inf. Proc. Syst.* **22**, 9 (2009).
34. R. Ben-Yishai, R. L. Bar-Or, H. Sompolinsky, Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3844–3848 (1995).
35. A. Johnson, K. Seeland, A. D. Redish, Reconstruction of the postsubiculum head direction signal from neural ensembles. *Hippocampus* **15**, 86–96 (2005).
36. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–8 (2006).
37. J. M. Beck, P. E. Latham, A. Pouget, Marginalization in neural circuits with divisive normalization. *J. Neurosci.* **31**, 15310–15319 (2011).
38. C. Lyu, L. F. Abbott, G. Maimon, Building an allocentric travelling direction signal via vector computation. *Nature* **601**, 92–97 (2022).
39. A. Compte, Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
40. A. A. Faisal, L. P. J. Selen, D. M. Wolpert, Noise in the nervous system. *Nat. Rev. Neurosci.* **9**, 292–303 (2008).
41. Y. Burak, I. R. Fiete, Fundamental limits on persistent activity in networks of noisy neurons. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17645–17650 (2012).
42. B. K. 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).
43. L. K. Scheffer *et al.*, A connectome and analysis of the adult Drosophila central brain. *eLife* **9**, e57443 (2020).
44. D. B. Turner-Evans *et al.*, The neuroanatomical ultrastructure and function of a biological ring attractor. *Neuron* **108**, 145–163.e10 (2020).
45. T. Merkle, R. Wehner, Desert ants use foraging distance to adapt the nest search to the uncertainty of the path integrator. *Behav. Ecol.* **21**, 349–355 (2010).
46. T. Merkle, Uncertainty about nest position influences systematic search strategies in desert ants. *J. Exp. Biol.* **209**, 3545–3549 (2006).
47. M. G. Campbell, A. Attinger, S. A. Ocko, S. Ganguli, L. M. Giocomo, Distance-tuned neurons drive specialized path integration calculations in medial entorhinal cortex. *Cell Rep.* **36**, 109669 (2021).
48. K. Cheng, S. J. Shettleworth, J. Huttenlocher, J. J. Rieser, Bayesian integration of spatial information. *Psychol. Bull.* **133**, 625–637 (2007).
49. R. Knight *et al.*, Weighted cue integration in the rodent head direction system. *Philosop. Trans. R. Soc. B: Biol. Sci.* **369**, 20120512 (2014).
50. X. Sun, M. Mangan, S. Yue, An analysis of a ring attractor model for cue integration. *Lecture Notes Comput. Sci. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNAI* (2018), vol. 10928, pp. 459–470.
51. X. Sun, S. Yue, M. Mangan, A decentralised neural model explaining optimal integration of navigational strategies in insects. *eLife* **9**, e54026 (2020).
52. J. M. Esnaola-Acebes, A. Roxin, K. Wimmer, Flexible integration of continuous sensory evidence in perceptual estimation tasks. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2214441119 (2022).
53. W. J. Ma, M. Husain, P. M. Bays, Changing concepts of working memory. *Nat. Neurosci.* **17**, 347–356 (2014).
54. R. Moreno-Bote *et al.*, Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417 (2014).
55. S. Carroll, K. Josić, Z. P. Kilpatrick, Encoding certainty in bump attractors. *J. Comput. Neurosci.* **37**, 29–48 (2014).
56. R. G. Robertson, E. T. Rolls, P. Georges-François, S Panzeri, Head direction cells in the primate pre-subiculum. *Hippocampus* **9**, 206–219 (1999).
57. O. Baumann, J. B. Mattingley, Medial parietal cortex encodes perceived heading direction in humans. *J. Neurosci.* **30**, 12897–12901 (2010).
58. A. Finkelstein *et al.*, Three-dimensional head-direction coding in the bat brain. *Nature* **517**, 159–164 (2015).
59. M. B. Zugaro, E. Tabuchi, C. Fouquier, A. Berthoz, S. I. Wiener, Active locomotion increases peak firing rates of anterodorsal thalamic head direction cells. *J. Neurophysiol.* **86**, 692–702 (2001).
60. M. E. Shinder, J. S. Taube, Self-motion improves head direction cell tuning. *J. Neurophysiol.* **111**, 2479–2492 (2014).
61. J. J. Knierim, H. S. Kudrimoti, B. L. McNaughton, Interactions between idiothetic cues and external landmarks in the control of place cells and head direction cells. *J. Neurophysiol.* **80**, 425–446 (1998).
62. Y. E. Fisher, J. Lu, I. D'Alessandro, R. I. Wilson, Sensorimotor experience remaps visual input to a heading-direction network. *Nature* **576**, 121–125 (2019).
63. A. T. Keinath, R. A. Epstein, V. Balasubramanian, Environmental deformations dynamically shift the grid cell spatial metric. *eLife* **7**, e38169 (2018).
64. M. Milford, G. Wyeth, D. Prasser, "RatSLAM: A hippocampal model for simultaneous localization and mapping" in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA 2004 (IEEE, New Orleans, LA, USA, 2004), vol. 1, pp. 403–408.*

65. M. Mulas, N. Waniek, J. Conradt, Hebbian plasticity realigns grid cell activity with external sensory cues in continuous attractor models. *Front. Comput. Neurosci.* **10** (2016).
66. S. A. Ocko, K. Hardcastle, L. M. Giocomo, S. Ganguli, Emergent elasticity in the neural code for space. *Proc. Natl. Acad. Sci. U.S.A.* **115** (2018).
67. H. J. I. Page *et al*., A theoretical account of cue averaging in the rodent head direction system. *Philosop. Trans. R. Soc. B: Biol. Sci.* **369**, 20130283 (2014).
68. H. J. I. Page, K. J. Jeffery, Landmark-based updating of the head direction system by retrosplenial cortex: A computational model. *Front. Cell. Neurosci.* **12**, 191 (2018).
69. A. J. Cope, C. Sabo, E. Vasilaki, A. B. Barron, J. A. R. Marshall, A computational model of the integration of landmarks and motion in the insect central complex. *PLoS One* **12**, e0172325 (2017).
70. R. J. van Beers, A. C. Sittig, J. J. D. vd. Gon, Integration of proprioceptive and visual position-information: An experimentally supported model. *J. Neurophysiol.* **81**, 1355–1364 (1999).
71. P. E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations, Applications of Mathematics* (Springer, Berlin, ed. 3, 2010), No. 23.
72. K. V. Mardia, P. E. Jupp, *Directional Statistics* (John Wiley & Sons, 2000), p. 3.

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:

Contents

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

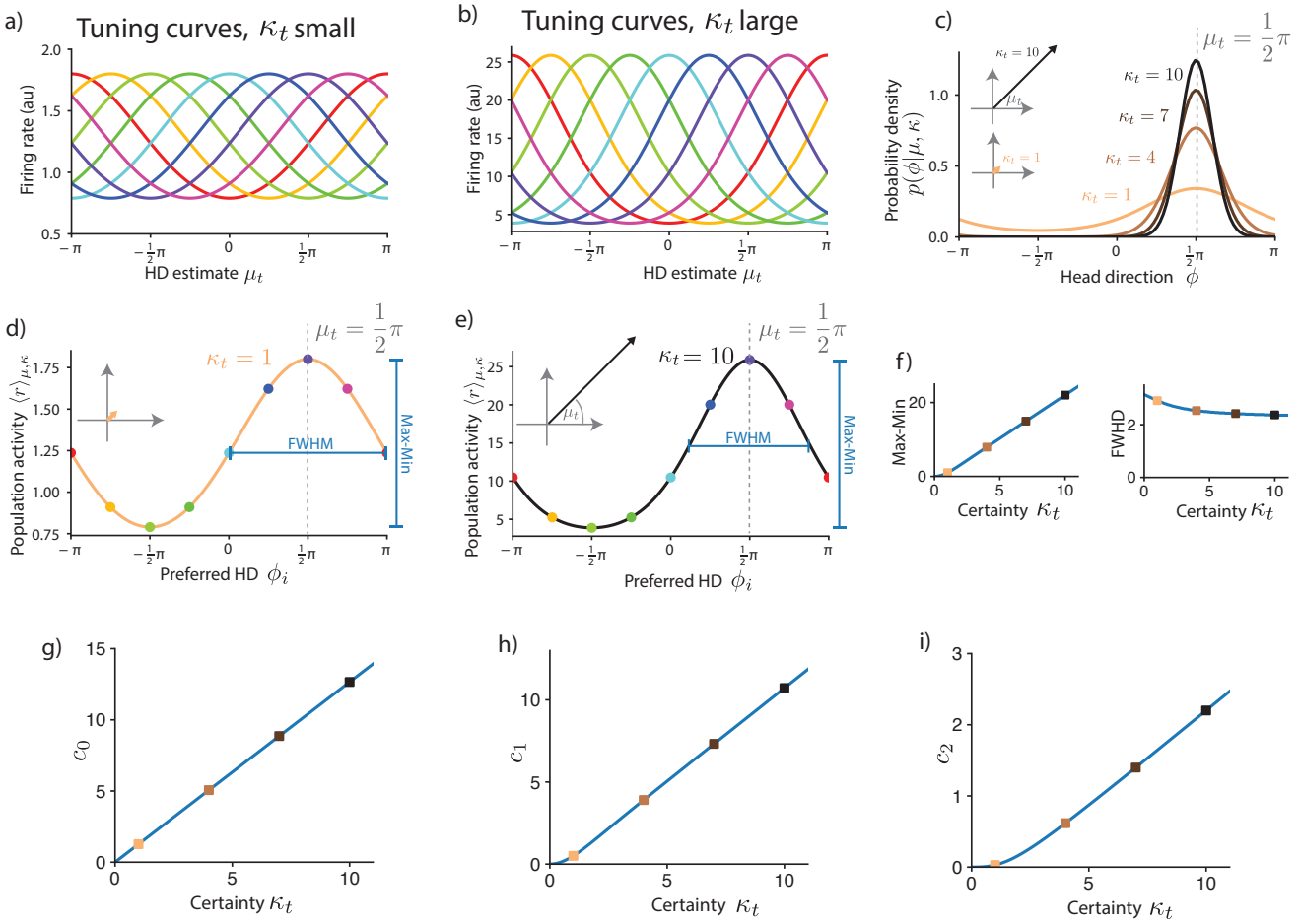

Fig. S1. Encoding the HD with linear probabilistic population codes. a) Tuning curves with respect to encoded HD estimate for small values of encoded certainty κ_t are cosine-shaped. Here, we show tuning curves of 8 example neurons with $\kappa_t = 1$ (colors indicate preferred HD ϕ_i). **b)** Tuning 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.

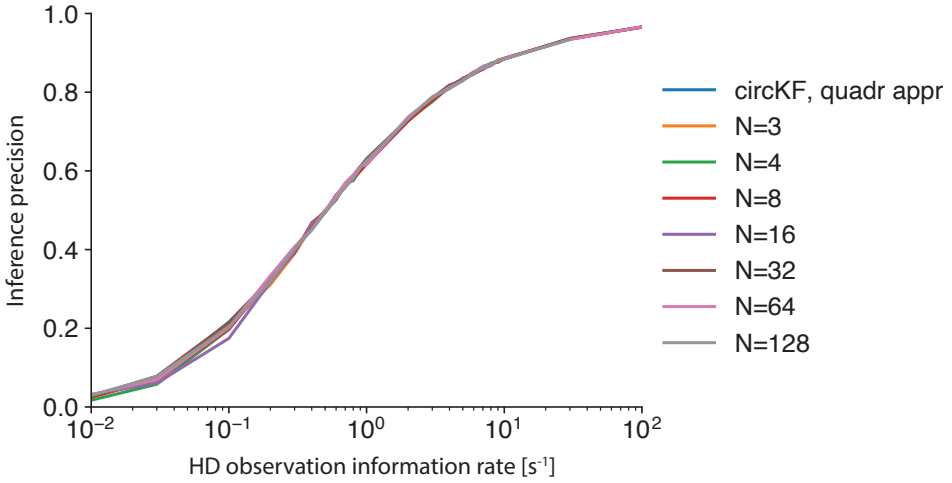

Fig. S2. Network inference performance is mostly independent of the number of neurons N in the Bayesian ring attractor network. Here, for each value of the observation reliability κ_z and number of neurons in the network N we compute the circular average distance of the network's HD estimate μ_T from 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 $\Delta t = 0.01$.

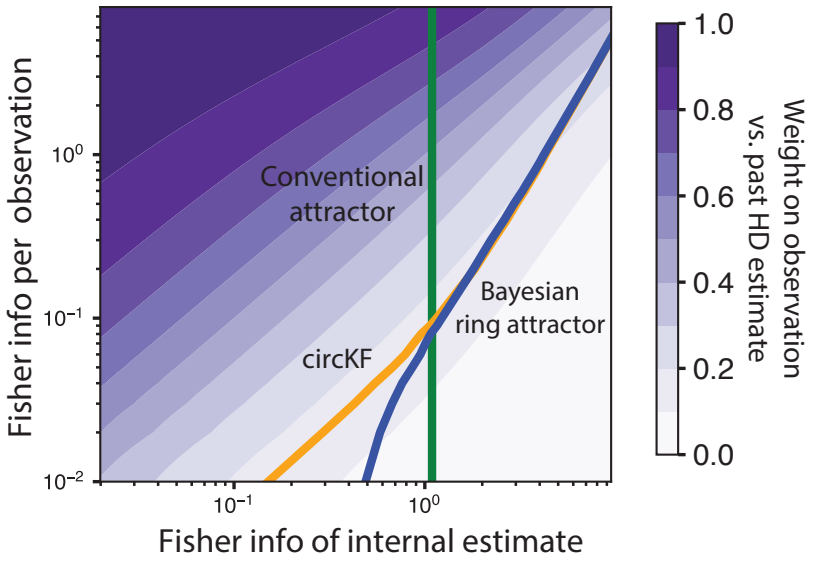

Fig. S3. The weight with which a single observation contributes to the HD estimate varies with informativeness of both the HD observations and the current HD estimate. Same plot as main text Fig. 4, only on a log-log scale. Here, we additionally plot the resulting updates for the circKF, to demonstrate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

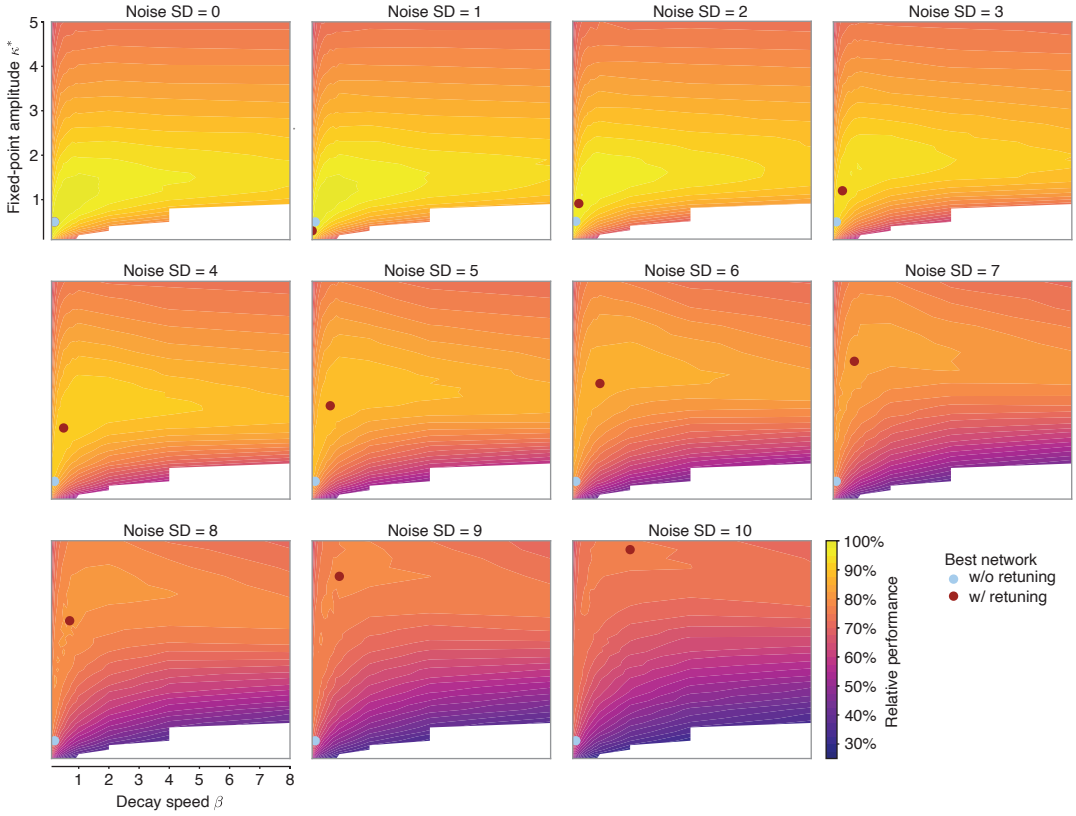

Fig. S4. Neural noise changes the optimal fixed point amplitude and decay speed. We simulated a network of $N = 64$ neurons with different levels of additive Gaussian noise with variance $\sigma_{nn}^2 \delta t$ to each neuron within each time step δt, for different fixed point amplitudes κ^* and decay 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.

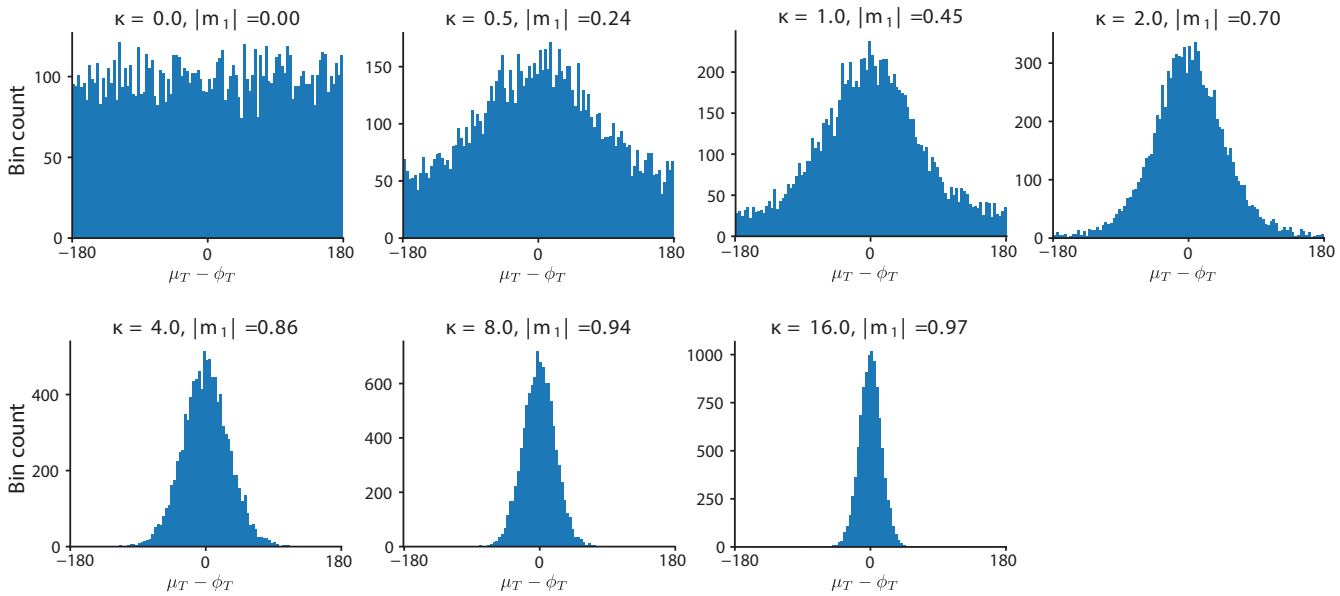

Fig. S5. Visualizing the HD tracking perfomance measure. To provide a better intuition for the used HD tracking performance measure we here show how a specific distribution of HD tracking errors (horizontal axis, in degrees) relates to this performance measure. In particular, we drew 10000 samples from 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.