Bayesian inference of circular variables in ring attractor networks

Anna Kutschireiter[1]*, Melanie A Basnak[1], Jan Drugowitsch[1]

[1]Harvard Medical School, Department of Neurobiology, 200 Longwood Avenue, Boston, MA-02115, United States.

*Correspondence: anna_kutschireiter@hms.harvard.edu, jan_drugowitsch@hms.harvard.edu

Summary

Working memories are thought to be held in attractor networks in the brain. Because working memories are often based on uncertain information, memories should ideally come with a representation of this uncertainty for strategic use in behavior. However, the attractor states that hold these memories in attractor networks commonly do not represent such uncertainty. Focusing here on ring attractor networks for encoding head direction, we show that these networks in fact feature all the motifs required to represent uncertainty in head direction estimates. Specifically, they could do so by transiently modulating their overall activity by uncertainty, in line with a circular Kalman filter that performs near-optimal statistical circular estimation. More generally, we show that ring attractors can perform near-optimal Bayesian computation if they can flexibly deviate from their attractor states. Finally, we show that the basic network motifs sufficient for such statistical inference are already known to be present in the brain. Overall, our work demonstrates that ring attractors can in principle implement a dynamic Bayesian inference algorithm in a biologically plausible manner.

Keywords
Working memory; Bayesian inference; Ring attractor networks; Head direction neurons; Kalman filtering; Population coding; Drosophila; Central complex;

Introduction

Many brain functions - including motor control, classification, and pattern completion - have been attributed to attractor networks, and they have proven particularly useful in modeling working memory[1,2]. More specifically, these networks support neural population activity patterns that persist even in the absence of inputs, endowing them with the ability to retain past information across time[3]. A change in the memory's content then corresponds to a change in the network's population activity pattern. At these *attractor states*, the networks only store "point estimates" of these memories, without an associated sense of uncertainty. As this stands in conflict with the observation that memories include a sense of uncertainty (e.g., refs[4,5]), do we need to discard the idea of memories being stored in attractor networks? Our work shows that this does not need to be the case.

A ring attractor is a special case of an attractor whose set of stable activity profiles forms a ring in neural activity space and thus has the ability to represent circular variables. Head direction (HD) is a classic example of a circular variable that is encoded by a ring attractor network in the brain[6]. Many features of mammalian HD neurons are highly suggestive of ring attractors[7–11]. Moreover, recent work has revealed HD cells in the *Drosophila* brain, which not only function as a ring attractor, but also form a topographic map of HD[12–14]. Importantly, the brain often estimates HD under conditions of high uncertainty -- e.g., in unfamiliar environments, or in darkness. Ideally, these HD networks would respond differently to a new piece of information, depending on the current level of uncertainty in the HD estimate. Such an uncertainty-weighted response is a hallmark of Bayesian inference[15]. How exactly the ring attractor networks that track HD could implement Bayesian inference without an explicit notion of uncertainty, however, remains largely unknown.

To address potential neural mechanisms for doing so, we took a normative modeling approach, and established how ring attractor networks could maintain and update uncertainty along with the encoded estimate (see Fig. 1 for an overview of our approach). Specifically, we first asked how uncertain HD estimates ought to be updated from unreliable information, irrespective of how these estimates are encoded in the activity of neural populations. We then combined the resulting Bayesian ideal observer model with a neural representation of uncertain HD estimates to arrive at a neural network architecture that can well-approximate the required computations. Interestingly, this network has the general connectivity structure of a ring attractor network. However, its ability to perform near-Bayesian inference depends on its connectivity strengths. A tightly connected, “strict” attractor network performs worse than a weakly connected, “loose” attractor network. This is because strict attractor networks rapidly decay back to their attractor states, while loose attractor networks can persistently deviate from these states. As we show, these deviations are essential to perform the required Bayesian computations. Nonetheless, the networks do not need to be finely tuned to achieve close-to-optimal HD tracking performance. Indeed, a large range of loose networks can adequately combine uncertain HD estimates with unreliable sensory information. Lastly, we showed that model ring attractors can implement dynamic Bayesian inference even after we incorporate constraints from neural connectivity data. In summary, our work provides a principled theoretical foundation for how attractor networks can maintain a sense of uncertainty in their memories, even without an explicit notion of uncertainty. Although we focus on HD encoding as a concrete example, our results are potentially also relevant to other ring attractors in the brain (e.g., the grid cell representation of an animal’s path).

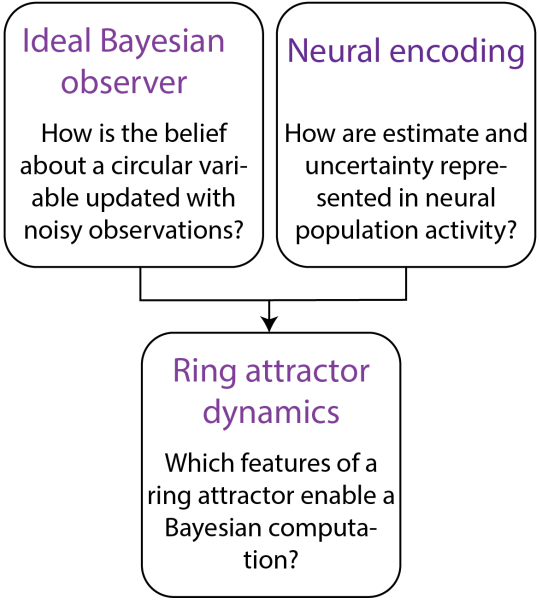
Ideal Bayesian observer
How is the belief about a circular variable updated with noisy observations?
Neural encoding
How are estimate and uncertainty represented in neural population activity?
Ring attractor dynamics
Which features of a ring attractor enable a Bayesian computation?

Figure 1. Our approach combines an ideal Bayesian observer model for circular variables with a 'bump' encoding to derive ring attractor dynamics that perform a Bayesian computation.

Results

Circular Kalman filtering: a Bayesian ideal observer model for tracking circular variables

We first focus on how uncertain memory ought to be updated from new sensory information irrespective of how this memory is encoded in the activity of a neural network. We do so with the example of HD tracking, by deriving the statistically best HD estimate from a continuous stream of unreliable absolute and relative HD information. This results in a generic algorithm - the circular Kalman filter - that tells us how an estimate of HD, or of any other dynamic circular variable (e.g., visual orientation, time of day, etc.), ought to be updated over time, and the role uncertainty plays in these updates. In the sections that follow we ask how this algorithm can be implemented by neural networks, and analyze the properties of these networks.

HD estimates are informed by two qualitatively different types of sensory inputs (Fig. 2a). Relative HD observations (or **angular velocity observations**), e.g., vestibular or proprioceptive signals, provide information about changes in HD. As they tend to be noisy, integrating them over time results in gradual error accumulation, and a HD estimate that increasingly deviates from the true HD. **Absolute head direction observations**, such as the position of a visual landmark, provide direct HD information that can be used to re-calibrate the HD estimate. Since these observations are also noisy, they should be combined with the internal HD estimate according to their respective reliabilities.

Here, we use dynamic Bayesian inference to properly handle the uncertainties arising from the aforementioned unreliable sensory inputs. We assume access to both angular velocity observations $v_t \in R$ and absolute HD observations $z_t \in [-\pi, \pi]$, which provide noisy information about true angular velocity $\dot{\phi}_t \in R$ and HD $\phi_t \in [-\pi, \pi]$, respectively. Specifically, angular velocity observations are corrupted by Gaussian noise that limits the precision of these observations (with precision κ_v, larger κ_v= more reliable), while absolute HD observations are corrupted by von Mises noise with precision κ_z, the Gaussian equivalent for circular variables. Dynamic Bayesian inference accounts for uncertainties arising from these noisy observations, by forming a **posterior belief** of HD $p(\phi_t|z_{0:t}, v_{0:t})$ that is continuously updated in light of new incoming sensory evidence. Importantly, this belief constitutes a whole probability distribution, rather than a single point estimate, which automatically includes a measure of uncertainty around the best HD estimate[15,16].

Estimating circular variables, such as HD, precludes the use of standard dynamic Bayesian inference schemes such as the Kalman filter[17,18] to update the posterior belief $p(\phi_t|z_{0:t}, v_{0:t})$ over time. Instead, statistical inference turns out to be analytically intractable[19] and needs to be approximated (see Methods). Here, we approximate this belief at each point in time by a von Mises distribution, $p(\phi_t|z_{0:t}, v_{0:t}) \approx VM(\mu_t, \kappa_t)$, which is fully characterized by its mean μ_t, which is the current best HD estimate, and its precision κ_t, which measures the estimate's certainty (Fig. 2b). As these two posterior, or belief, parameters fully specify the HD belief, updates of the belief in light of sensory evidence simplify to updating these two parameters. We derived the parameter update dynamics by a technique called projection filtering[20,21], resulting in

$$d\mu_t = v_t dt + \frac{\sqrt{2\kappa_z dt}}{\kappa_t} sin\,(z_t - \mu_t), \qquad (1)$$

$$d\kappa_t = -\frac{f(\kappa_t)}{2\kappa_v}\kappa_t dt + \sqrt{2\kappa_z dt}\, cos\,(z_t - \mu_t). \qquad (2)$$

Here, $f(\kappa_t)$ is a monotonically increasing nonlinear function that controls the speed of decay of one's certainty κ_t (see Eq. (10) in Methods). Equations (1) and (2) together define an algorithm that we call the **circular Kalman filter** (circKF)[21]. The circKF provides a general solution for estimating the evolution of a circular variable over time from noisy data.

To provide intuition for the filter's operation, let us represent the belief parameters in their polar coordinate form as a vector on the 2D plane (Fig. 2b inset). Then, the vector's direction determines the mean HD estimate μ_t, and its length the precision κ_t. Let us now consider how this vector is updated in light of angular velocity and absolute head direction observations.

Angular velocity observations. Without absolute head direction observations, i.e., $\kappa_z = 0$, the HD estimate μ_t is fully determined by integrating angular velocity observations v_t over time, i.e., angular path integration (Fig. 2c). In our vector representation, angular velocity information (the $v_t dt$ term in Eq. (1)) rotates the brain's HD belief by adding a tangential vector (Fig. 2c, green vector). The increasing error from such angular path integration comes with an associated drop in the belief's certainty κ_t ($f(\kappa_t)$-related term in Eq. (2)), which causes the belief vector to shrink (Fig. 2c, bottom). Interestingly, angular velocity observations always decrease certainty. The decrease might be more modest for more precise velocity observations (i.e., κ_v large), but nonetheless persists. Thus, if only angular velocity observations are present, the posterior certainty κ_t will inevitably decay towards zero (uniform posterior distribution, i.e., complete lack of knowledge), with a speed of decay that is determined by the angular velocity observation's "informativeness" κ_v.

Absolute head direction observations. Absolute HD information, like observing a visual landmark, directly informs about the current HD, and thus calibrates the internal HD estimate. To weigh the reliability of such information against the current HD estimate's certainty, its impact in Eqs. (1) and (2) is scaled by κ_z: if the cue's reliability κ_z is large, the observation z_t will substantially change the mean μ_t towards the direction of the cue. Conversely, if the current certainty κ_t is large compared to the cue's reliability, an absolute HD observation z_t will hardly update the existing estimate. In vector form, this weighting by reliability corresponds to adding an absolute HD information vector to the current belief vector (Fig. 2d/f; see Methods). The direction and length of this HD information vector are determined by the observation's position z_t and reliability κ_z, respectively (Fig. 2d, blue vector). Depending on how well the observation is aligned with the current belief (as measured by the cosine in Eq. (2)), the certainty κ_t can either increase or, in the case of a strongly conflicting stimulus, even *decrease* (Fig. 2e). This interesting result is a consequence of the circular nature of the inference task[22], and stands in contrast to the Kalman filter where absolute information *always* increases the estimate's certainty[23]. It is thus a key distinction between the Kalman filter and the circKF.

In a dynamic setting, both angular velocity and absolute HD observations are available as a continual stream. That is, at every point in time, the belief is updated according to Eqs. (1) and (2). In summary, angular velocity observations rotate the HD estimate and reduce certainty. Absolute HD observations, in contrast, update the HD estimate weighted by their reliability, and either increase certainty (if compatible with the current belief) or reduce certainty (if strongly conflicting with the current belief). These operations are continuously repeated to bring the current belief in line with the latest observations.

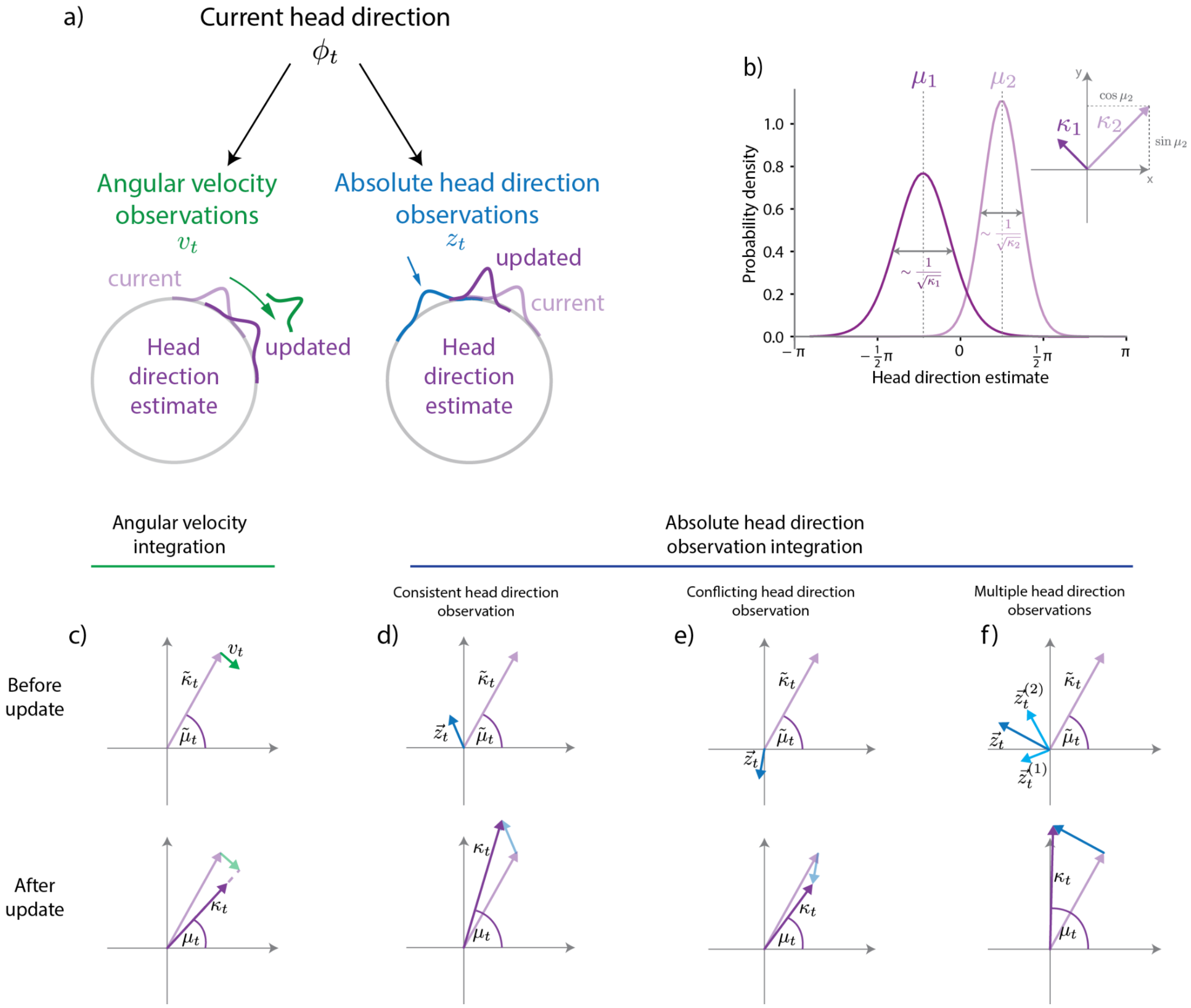

Figure 2. Tracking circular variables with the circular Kalman filter.

The circular Kalman filter performs dynamic Bayesian inference for circular variables. Its operation is illustrated here for tracking HD.

a) Two different types of observations inform the brain's estimated head direction ϕ_t: angular velocity observations v_t (green) provide noisy information about the true angular velocity $\dot{\phi}_t$, with precision κ_v (larger = more reliable), and absolute HD observations z_t (blue) provide noisy information about the true HD ϕ_t, with precision κ_z (larger = more reliable).

b) At every point in time, the belief $p(\phi_t|v_{0:t}, z_{0:t})$ about HD is approximated by the unimodal von Mises distribution, the Gaussian equivalent for circular variables. It is fully characterized by its mean parameter μ_t, which determines the position of the mode, and its precision parameter κ_t, which determines our belief's certainty. Interpreted as the polar coordinates in the 2D plane, these parameters provide a convenient vector representation of the belief (inset).

c) Angular velocity observations v_t rotate the current belief vector in the direction of the observations (angular path integration). Error accumulation from angular path integration comes with an associated drop in certainty and a corresponding drop in the vector's length (top vs. bottom).

d) Integrating absolute HD observations corresponds to adding the absolute HD observation vector (cyan) to the current belief vector (purple).

e) Absolute HD observations that are in conflict with the current belief (e.g., >120deg from the current estimate) result in a shortening of the belief vector (top vs. bottom) and an associated reduction of the belief's certainty.

f) Integration of multiple absolute HD cues, such as wind and vision, can be considered as a sum of multiple observation vectors.

Neural encoding of HD estimate and uncertainty

To link our ideal observer model to neural networks, we need to specify how the model's belief might be encoded by this activity pattern. In other words, we need to link our "algorithmic model" to a network model. Consider a ring attractor network where the peak of a localized increase in activity, or *bump*, encodes the estimate μ_t of the circular variable -- here, HD[7,24]. Here we assume that the bump's amplitude scales with the encoded certainty κ_t. This assumption is supported by some experimental evidence from the head direction system[10,25,26]. In any network where this assumption is correct, the activity of a neuron i with preferred head direction ϕ_i can be written as (Fig. 3a)

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

where μ_t and κ_t are the encoded belief's mean and certainty, and the "other components" might be a finite activity baseline or minor contributions of higher-order Fourier components to the activity. 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 modulated by certainty. This is satisfied, for example, by any unimodal bump profile (as the one in Fig. 2a) whose overall gain is governed by certainty. A particularly interesting case that matches Eq. (3) is a linear probabilistic population code[27,28] with von Mises-shaped tuning curves and independent Poisson neural noise (see SI text and Fig. S1).

This simple encoding scheme allows the network to encode both mean estimate μ and associated certainty κ, as required for implementing the Bayesian update rules (Eqs. (1) and (2)). Moreover, the natural parameters of the von Mises belief, $\theta_1 = \kappa_t \cos(\mu_t)$ and $\theta_2 = \kappa_t \sin(\mu_t)$, can be recovered by taking a weighted sum of the neural population's activity (Methods). This makes these parameters accessible to downstream neurons via simple (linear) neural operations.

Interestingly, θ_1 and θ_2 represent the von Mises distribution in terms of Cartesian vector coordinates in the 2D plane, whereas μ and κ are its polar coordinates (cf. Fig. 2b). Such a representation is related to the phasor representation of neural activity[29], which also translates bump position and amplitude to polar coordinates in the 2D plane (Fig. 3b). Since in our model the activity bump is scaled by certainty, the phasor representation of neural activity equals the vector representation of the von Mises distribution (Fig 3b,c).

Based on our ideal observer model (our "algorithmic model"), we know the vector operations required to implement the circular Kalman filter (Fig 2c-f). Thus, in what follows, we make use of this equality to implement the operations of the ideal observer model through neural dynamics. In other words, we show how the circKF algorithm could be implemented by a neural network.

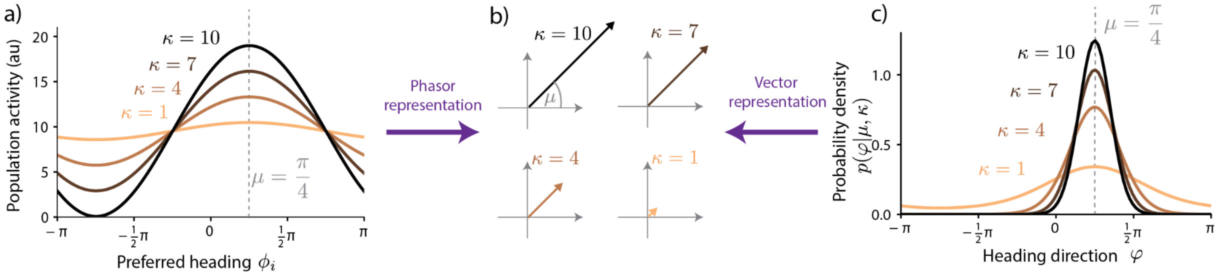

Figure 3. Encoding the HD belief in neural population activity.

a) Neural population activity profile (e.g., average firing rate) encoding the HD estimate $\mu = \pi/4$ with different values of certainty κ. Neurons are sorted by preferred head directions ϕ_i.

b) Vector representation of estimate $\mu = \pi/4$ for different values of certainty κ. This vector representation can be obtained by linearly decoding the population activity in a) ("phasor representation"). It also corresponds to the vector representation of the von Mises distribution in c), and thus connects neural activities with the probability distributions they encode.

c) Von Mises probability densities for different values of certainty κ and fixed HD estimate $\mu = \pi/4$. Note that, unlike the population activity in a), the density sharpens around the mean with increasing certainty.

Recurrent neural networks can track Bayesian HD estimates

Linking the belief parameters to neural population activity (Eq. (3)) reveals the population activity dynamics required to implement our ideal HD tracking model (Eqs. (1) & (2)). We now ask how these dynamics can be implemented by a recurrent neural network (RNN). We start with an idealized network with a single neural population, similar to many generic ring attractor networks (e.g., refs[3,7]). Later, we will build on this idealized network to construct a more distributed network that satisfies the known constraints of a biological ring attractor that encodes HD.

Simple and analytically accessible network dynamics that implement the circular Kalman filter (Eqs. (1) & (2)) are of the form

$$\dot{\mathbf{r}}_t = -\frac{1}{\tau}\mathbf{r}_t - g(\mathbf{r}_t)\mathbf{r}_t + W \cdot \mathbf{r}_t + \mathbf{I}_t^{\mathrm{ext}}, \qquad (4)$$

where r_t denotes a vector of neural activities, with neurons ordered by their preferred head directions $\varphi^{(i)}$, τ is the network time constant (leak), W is the recurrent connectivity matrix, and $I_t{}^{ext}$ is a vector of external inputs to the network. The synaptic inhibition nonlinearity $g(r_t)$ is closely related to the nonlinearity $f(\kappa_t)$ in Eq. (2): it is tuned such that its output increases with bump amplitude, and thus implements nonlinear global inhibition.

The network dynamics in Eq. (4) allow us to attribute specific wiring patterns (or motifs) to the effect they have on the population activity vector, mimicking the transformations required to implement computations in the circKF (Fig. 2c,d). In particular, probabilistic angular path integration is implemented by an interplay between recurrent connectivity (W), leak ($1/\tau$), and synaptic inhibition ($g(r_t)$). The matrix of recurrent connectivity W can be divided into symmetric (even) and asymmetric (odd) components (Fig. 4a). The even component holds the bump of activity at its current location in the absence of any other input. Meanwhile, the odd component can push the bump of activity around the ring -- e.g., in response to an angular velocity observation (Fig. 4b). Leak and global inhibition together cause the amplitude of the bump to decay over time (Fig. 4c), corresponding to the progressive decay in certainty in the absence of new HD information. Absolute HD observations enter the network via the external input vector I_t^{ext}, in form of a cosine-shaped bump with amplitude modulated by perceptual reliability κ_z (Fig. 4d). This input activity effectively implements the vector addition required for proper absolute HD observation integration. Then, the external information's weight is determined by the ratio between input amplitude and bump amplitude, in line with the weighting between cue reliability and own certainty required by the circKF. Bump position and amplitude dynamics derived from a network with these basic motifs well-approximate the parameter dynamics of the circKF (Eqs. (1) & (2); Fig. 4e; see Eqs. (13) & (14) in Methods for bump parameter dynamics).

In the limit of infinitely many neurons, the network can even be tuned to implement the circKF exactly. Importantly, a network that exactly implements the circular Kalman filter is *not* a ring attractor network: in the absence of external absolute HD input, $I_t^{ext} = 0$, the activity bump decays towards its single attracting state with zero amplitude (Fig. 4e). In contrast, ring attractor networks commonly settle on a constant activity bump with non-zero amplitude (the "attractor state") if input is removed. In our "exact inference" network, activity decay is by design, and reflects the continuously decreasing certainty arising from pure angular path integration in the absence of absolute HD observations.

Activity dynamics change qualitatively if we choose $g(r_t)$ such that the second term in Eq. (4) becomes a quadratic function of the bump amplitude (quadratic inhibition). This change in $g(\cdot)$ introduces ring attractor states with non-zero network activity, and has the additional advantage of making the network dynamics analytically accessible (see Methods). As a result, we can tune the network parameters such that this network implements **a quadratic approximation** to the circKF. This approximation becomes precise in the limit of large posterior certainties κ_t. In other words, the bump's amplitude dynamics will correctly reflect the posterior's certainty for large bump amplitudes, but will deviate from it in the small-certainty/small-amplitude limit.

Such a Bayesian ring attractor has two operating regimes: a regime close to the attractor state with constant, low bump amplitude, encoding approximately constant certainty, and a dynamic, high amplitude regime away from the attractor state, where the network correctly implements dynamic Bayesian inference. Numerical simulations confirm the existence of these two regimes: the network tracks the HD estimate and its associated certainty just like the circular Kalman filter in the dynamic regime, but features a slightly lower HD tracking precision, and overestimates its confidence, close to the attractor state (Fig. 4f, orange vs. blue; SI Fig. S2 shows that performance is largely independent of ring attractor population size). We will analyze these two regimes further in the next section.

In summary, the following three network motifs support the implementation of Bayesian inference in ring attractor networks (Fig. 4b-d): (i) asymmetric (odd) recurrent connectivity with strength modulated by angular velocity observations v_t, (ii) global inhibition that is approximately quadratic in bump amplitude, and (iii) a cosine-shaped external input at the position of the absolute HD observation, whose strength is modulated by the reliability κ_z of this observation. Motifs (i) and (ii) implement probabilistic angular path integration, whereas motif (iii) updates the network's current HD estimate in light of uncertain absolute HD observations. Interestingly, these motifs are common in many generic ring attractor networks, and have been discussed in terms of their function individually (see e.g. refs[7,30]). Here, we show that, together, they can implement approximate dynamic Bayesian inference for circular variables - inference that becomes more precise in the limit of large amplitudes, away from the attractor state.

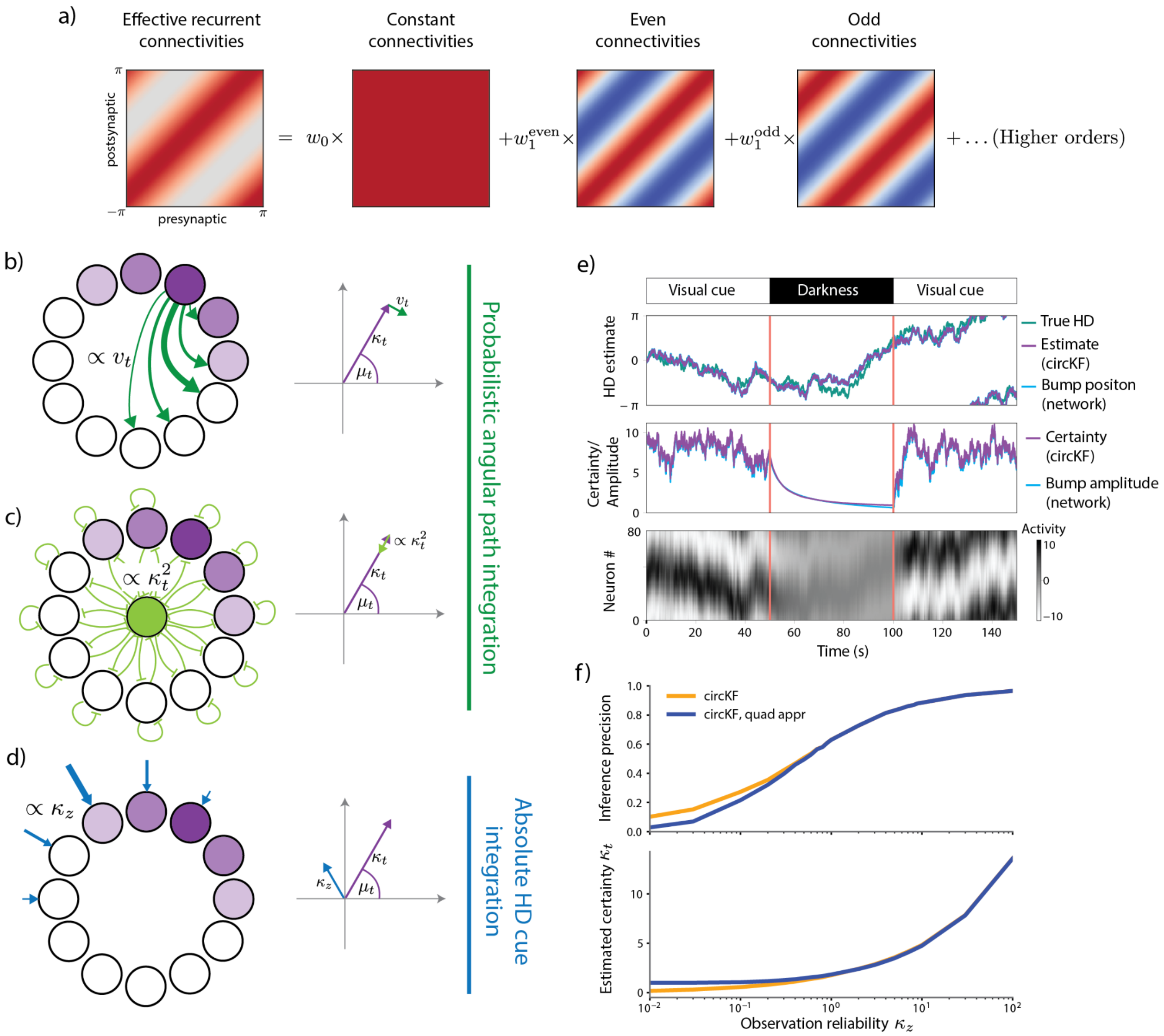

Figure 4. A recurrent neural network implementation of the circular Kalman filter.

a) Rotation-symmetric recurrent connectivities (here: neurons are sorted according to their preferred HD) can be decomposed into constant, cosine-shaped (even), sine-shaped (odd) and higher-order frequency components (basis function). Red and blue denote excitatory and inhibitory components, respectively.

b) Network motifs sufficient to implement the circular Kalman filter (b-d). Rotations of the HD estimate are mediated by sine-shaped (or odd) recurrent connectivities, whose strength is modulated by angular velocity observations.

c) Decay in amplitude arises from leak and global inhibition.

d) A cosine-shaped input to the network provides external absolute HD cue input. The strengths of this input is modulated by observation reliability κ_z.

e) The dynamics of the network implement the dynamics of the ideal observer's belief, as shown in a simulation of a network with 80 neurons. Here, we assume that vision provides the network with absolute HD information. When a 'visual cue' was present, both absolute

HD observations and angular velocity observations were available. During ‘darkness’, only angular velocity observations were available.

f) The network implementation with quadratic leak approximation (circKF, quadratic approx) tracks the HD estimate with the same precision (top; higher = lower circular distance to true HD) as the circular Kalman filter (circKF, Eqs. (1) and (2)) if absolute HD observations are reliable (large κ_z), but with slightly lower precision once they become less reliable (small κ_z). This drop co-occurs with an overestimate in the estimate’s confidence κ_t (bottom). Plots are averages over 5’000 simulations (see Methods for simulation details).

Ring attractors approximate Bayesian inference for HD tracking through amplitude dynamics

Our Bayesian ring attractor network qualitatively differs in two ways from classical ring attractor networks for working memory[1,7,31]. First, classical networks are not explicitly designed to represent uncertainty, and therefore assign no interpretation to their bump’s amplitude. Second, ring attractors are usually designed to operate close to their attractor states, where the bump amplitude tends to vary little. We now ask how important it is for network activity - including bump amplitude - to deviate from these attractor states to implement Bayesian inference.

Our RNN with quadratic inhibition is a ring attractor network that exhibits attracting states with a finite amplitude. In networks of such structure, bump position changes upon absolute HD inputs in proportion to the ratio between input amplitude (≙ reliability of absolute HD cue) and bump amplitude (≙ own certainty) (see Eq. (13) in Methods). Jointly tuning the network parameters, in particular recurrent weights, network time constant, and inhibitory weights, allows us to change the bump amplitude dynamics (Eq. (15) in Methods) to explore different regimes of network operation. Specifically, we tuned the network parameters to modulate two amplitude characteristics. The first is the attractive amplitude fixed point $\kappa *$ of the population activity bump (specified by the parameter $\kappa *$, which defines both the bump amplitude and the uncertainty it encodes). The second is the effective decay speed β, which describes how fast the amplitude approaches this fixed point (larger β implies faster dynamics, Fig. 5a).

In the limit of fast dynamics, $\beta \to \infty$, the network becomes a “strict attractor” with a bump amplitude that never moves away from its fixed point. As the bump amplitude encodes the HD estimate’s uncertainty, such a strict attractor never updates its uncertainty, and consequently lacks proper reliability weighting of absolute HD observations. In general, we expect that such a network is not able to estimate HD as accurately as one that correctly implements Bayesian inference, as it does not properly account for the observation’s reliability. Strict attractors with a numerically optimized fixed point amplitude $\kappa *$ (Fig. 5b inset) can still perform HD estimation reasonably well (Fig. 5b, green curve), but perform systematically worse than the tuned network (Fig. 5b, circKF with quadratic approximation, blue curve) or the circKF (Fig. 5b, orange curve). Here, we measure performance by how much the estimate (bump position) deviates on average from the ground-truth HD (circular average distance, see Methods). Adjusting the fixed point amplitude for each level of information reliability *individually* further boosted the network’s estimation performance, and effectively re-established proper Bayesian reliability weighting of incoming absolute HD observations, similar to the circKF (Fig. 5c). Even then, strict attractors lack temporal adjustment of their internal certainty estimates, as required by Bayes-optimal evidence integration.

We next asked whether we could increase performance by loosening the attractor. Such a relaxation would permit the network to operate farther away from its attracting state, to which it will decay in the absence of absolute HD observations (the dynamics that implement the quadratic approximation of the circKF are a special case of this network, see Fig. 5a). Indeed, the slower dynamics (slower decay speed β) of such a "loose attractor" boosted overall performance (Fig. 5d). In fact, with network parameters tuned numerically to maximize performance, HD estimation performance becomes practically indistinguishable from that of the ideal Bayesian observer (light blue dot in Fig. 5d, light blue line in Fig. 5e). In this regime, the HD estimate and bump amplitude dynamics become almost identical to the dynamics of certainty representations in the circKF. On the other hand, more rigid networks with faster decay speed β (such as the strict attractor as an extreme case) clearly deviate from the circKF, despite optimized fixed-point values (Fig. 5e). Interestingly, the optimal network parameters do not necessarily coincide with the quadratic approximation of the circKF, which we found by analytically matching the certainty dynamics in the large-certainty limit rather than by numerical optimization. In fact, a wide range of network parameters lead to a relatively small performance loss (<10%, Fig. 5d). Therefore, accurate parameter tuning might be unnecessary, as long as the network dynamics remain sufficiently slow.

Overall, this demonstrates that proper HD estimation relies on weighting absolute HD observations both globally (Figs. 5 b,c), i.e. according to the average level of reliability, and dynamically (Figs. 5d,e), according to the dynamics of one's own certainty. Nonetheless, reasonable performance can be achieved over a wide range of network parameters. This may indicate a "built-in" implicit reliability weighting in attractor networks through their amplitude dynamics. As we just demonstrated, this requires sufficiently slow attractor dynamics around the fixed point and the possibility for deviations from the attractor state. This may explain why ring attractor networks perform evidence integration reasonably well in practice, even though they are unlikely to be precisely tuned to the task.

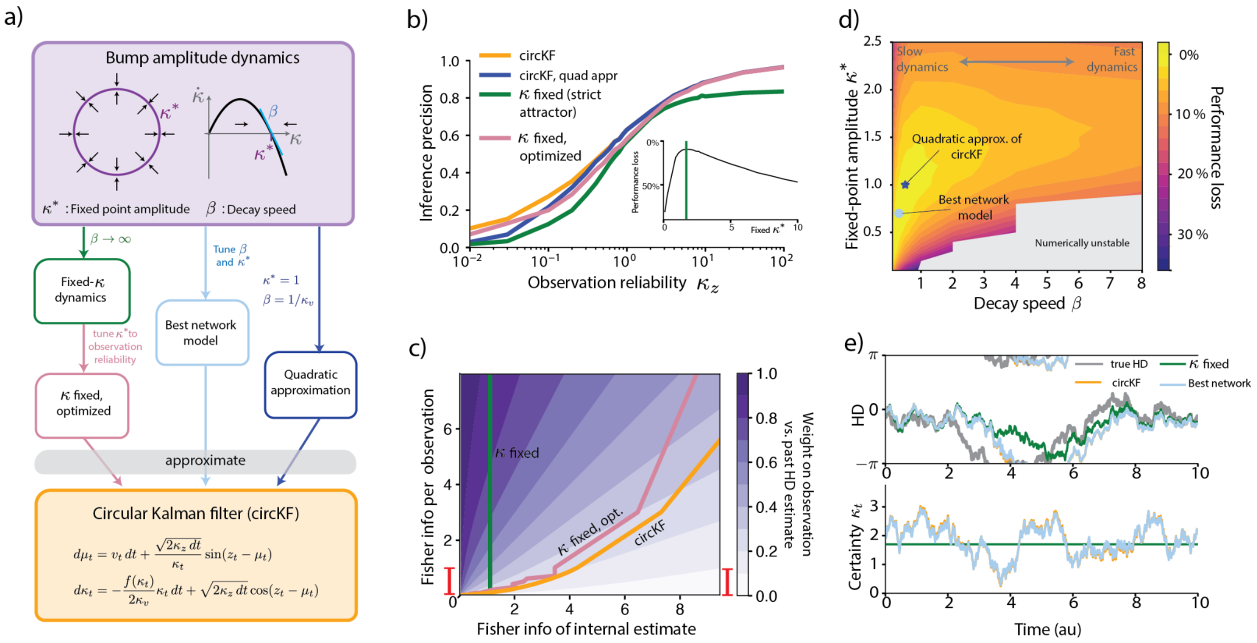

Figure 5. Attractor models with slow dynamics approximate Bayesian inference

- **a)** A linear RNN with quadratic inhibition can operate in different regimes. Its bump amplitude dynamics can be characterized by fixed point amplitude $\kappa *$ and decay speed β. Note that the bump position dynamics is described by the same equation across all compared regimes (Eq. (13) in Methods). However, the position dynamics depend on bump amplitude, whose dynamics differ across regimes (Eq. (15) in Methods). This causes HD tracking behavior to differ across network regimes.
- **b)** HD estimation performance as measured by inference precision (as defined by $1 - circVar$, see Methods). Here, the blue curve shows performance of the analytically tuned ring attractor network, implementing the quadratic approximation to the circKF (yellow). For the strict attractor (green curve), we chose $\kappa *$ to numerically maximize performance averaged across all levels of observation reliability, weighted by a prior $p(\kappa_z)$ on this reliability (see Methods). For the optimized, but still strict, network (pink curve), we found the performance-maximizing $\kappa *$ separately for each level of observation reliability.
- **c)** The weight with which a single observation contributes to the HD estimate varies with informativeness of both the absolute HD observations and the current HD estimate. We here illustrate this for an absolute HD observation that is orthogonal to the current HD estimate, resulting in the largest possible estimate change ($|z_t - \mu_t| = 90$deg in Eq. (1)). The weight itself quantifies how much the observation impacts the HD estimate as a function of how informative this observation is (vertical axis, measured by Fisher information of a 10ms observation) and our certainty in the HD estimate (horizontal axis, also measured by Fisher information) before this observation. A weight of one implies that the observation replaces the previous HD estimate, whereas a weight of zero implies that the observation does not impact this estimate. The close-to-optimal update weight of the circKF (yellow) forms a nonlinear curve through this parameter space. Fisher information per observation is directly related to the observation reliability κ_z, and the vertical red bar shows the equivalent range of observation reliabilities, $\kappa_z \in [10^{-2}, 10^2]$, shown in panel b.

Update weights for the tuned network (circKF with quadratic approximation) are not shown as they would be visually indistinguishable from that of the circKF, and only deviate from it for very uninformative observations (see SI Fig. S3).

d) Overall inference performance loss (compared to a particle filter; performance measured by avg. inference precision, as in **b**, 0%: same average inference precision as particle filter, 100%: average. inference precision = 0), averaged across all levels of observation reliability (weighted by prior $p(\kappa_z)$, see Methods) as a function of the bump amplitude parameters $\kappa *$ and β. For too small fixed point amplitudes and too fast dynamics, numerical simulations become unstable (grey area).

e) Simulated example trajectories of HD estimate/bump positions of HD estimate/bump positions (top) and certainties/bump amplitudes (bottom). To avoid cluttering, we are not showing the quadratic approximation of the circKF (visually indistinguishable from circKF and best network).

A biological ring attractor can implement the circular Kalman filter

Having established the network motifs sufficient for implementing dynamic Bayesian inference in ring attractor networks, and the network parameter regimes that lead to good HD tracking performance, we finally asked if biological networks are in principle able to implement such inference. A biological implementation is plausible because the critical motifs of our model networks are actually common in many generic ring attractor networks. The most well-studied biological ring attractor network is the HD system of the fruit fly *Drosophila*[13]. Here we show how the motifs of this network -- and, by extension, any biological ring attractor network -- could potentially implement dynamic Bayesian inference.

The ring attractor in the *Drosophila* HD system is composed of three core cell types, called EPG, PEN1 and *Δ*7 neurons[32–34], cf. Fig. 6a-c. Head direction is represented as a bump of neural activity in the EPG population[12]. 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[14,35]. This motif effectively establishes the velocity-modulated odd recurrent connectivity required to initiate turns in ring attractor networks (Fig. 6d). Moreover, EPG neurons are recurrently connected with inhibitory *Δ*7 neurons, which establishes broad inhibition (Fig. 6e). Finally, EPG neurons receive inhibitory inputs from so-called ER neurons, which send absolute HD information to EPG neurons[36–38] (Fig. 6f). In summary, the fly's HD system is equipped with the basic motifs to implement a Bayesian ring attractor.

To demonstrate that these motifs can in principle implement a Bayesian ring attractor, we analytically tuned the relative connection strength between (rather than within) the populations of our idealized network in Fig. 6c such that the dynamics of the bump parameters in the HD population implement the quadratic approximation to the circKF (see Eqs. (16)-(20) in Methods for network dynamics, SI for derivation). To achieve this, we additionally assumed that the broad inhibition implemented by the inhibitory population (Fig. 6c,e) was achieved by a subtractive signal that resulted from a multiplicative interaction between activities of INH and HD neurons. This multiplicative interaction achieves the quadratic certainty decay required to approximate the circKF. We found that this network achieves a HD tracking performance indistinguishable to that of our idealized Bayesian ring attractor network (SI Figure S4). 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.

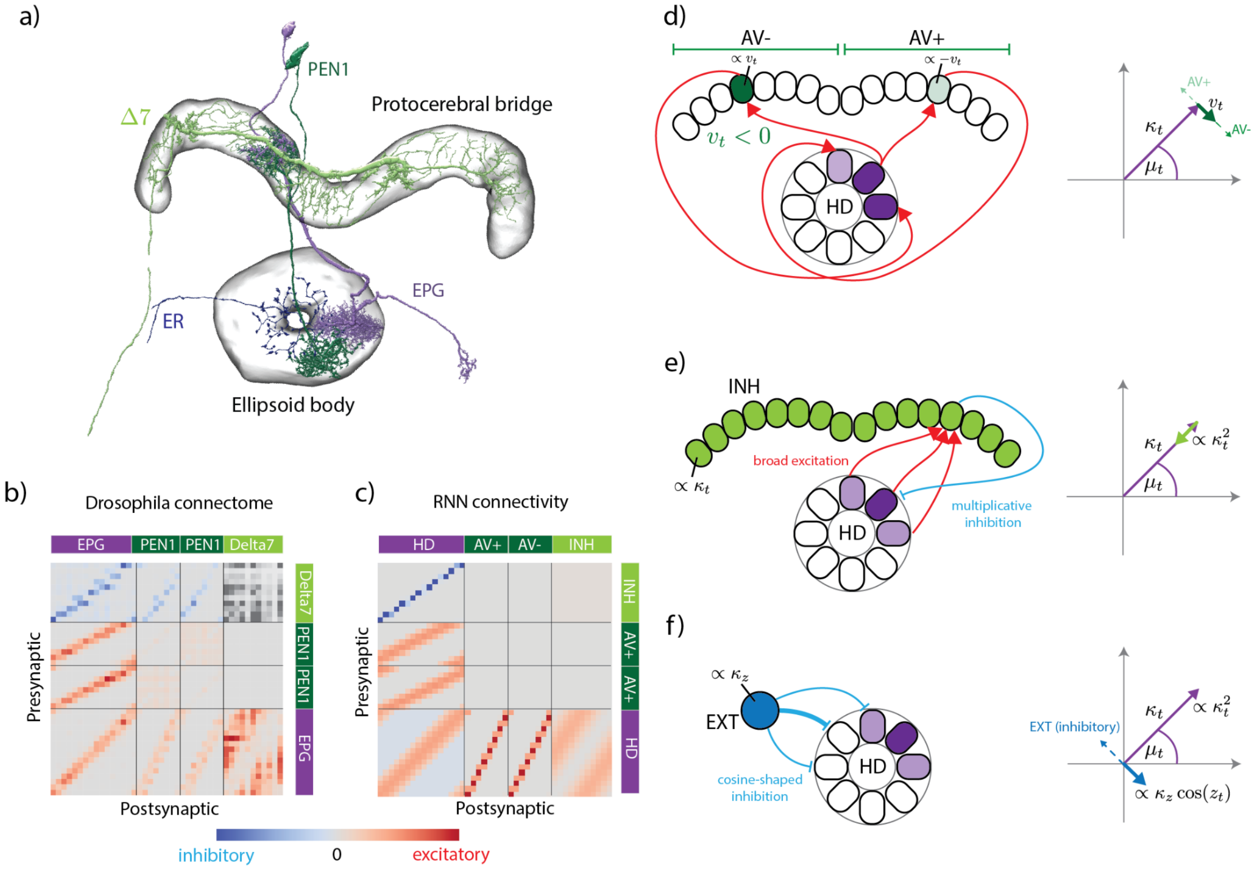

Figure 6. A Drosophila-like network implementing the circular Kalman filter.

- **a)** Cell types in the *Drosophila* brain that could contribute to implementing the circular Kalman filter.
- **b)** Connectivity between EPG, *Δ*7 and PEN1 neurons, as recovered from the hemibrain:v1.2.1 database[33]. ER neurons were omitted because they only form the inputs to the recurrently connected ring attractor. Here, neurons were grouped according to anatomical region as a proxy for preferred HD, and we used the total number of synaptic connections between two neurons to indicate connection strength. *Δ*7 to *Δ*7 connectivities are omitted, as the polarity of these connections (inhibitory or excitatory) remains unclear.
- **c)** The RNN connectivity profile that implements an approximation of Bayesian inference algorithm 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) in the two hemispheres as well as a shifted feedback connectivity from AV to HD populations effectively implement

the odd (or shifted) connectivity needed to turn the bump position (here: clockwise shift for anti-clockwise 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, implement 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 preferred direction *opposite* the location of the absolute HD observation, effectively implementing a vector sum of belief with absolute HD input.

Discussion

We have shown that ring attractor networks - prominent models for working memory of circular variables - can encode and compute with a sense of uncertainty, even when their attractor states are unable to do so. They can achieve this by operating in a dynamic regime away from these attractor states. In this regime, their bump amplitude can vary and thus can encode uncertainty. Such deviations from the attractor state are only possible in loose attractors with sufficiently weak connectivity strengths. Stronger connectivity leads to strict attractors that operate closer to their attractor states and feature worse performance. For a canonical working memory of a circular variable - our sense of head direction - we have shown that network motifs common to ring attractor networks are sufficient to implement the basic computations for dynamic Bayesian inference: (i) angular velocity-modulated odd recurrent connectivities implement incremental changes to the HD estimate, (ii) global inhibition implements the required decay in certainty over time, and (iii) reliability-modulated external input implements reliability-weighted absolute HD integration. We expect these findings to translate to working memories of other circular variables, like those that follow circadian rhythms, or encode memory about visual orientations[5,39]. We further found that close-to-optimal estimation does not require exact tuning of the ring attractor network's connectivities, as long as the networks feature the aforementioned motifs and are flexible enough to deviate from their attractor states. Lastly, we demonstrated that a network with realistic biological constraints still supports the implementation of such a Bayesian ring attractor. Our findings thus suggest that ring attractor models can implement Bayesian computations for working memory.

A key element of our approach is the representation of uncertainty as the amplitude of a neural activity bump. This differentiates our work from recent network models that only performed reliability-weighted cue integration at the level of the inputs[40,41], without considering the resulting certainty of the HD estimate. In our framework, this certainty determines the weight with which new external evidence enters the estimate through the bump amplitude. As such, it plays a central computational role for updating the estimate, rather than being a passive measure of precision[25,26]. It predicts that the speed with which the activity bump reacts to changing absolute HD observations should depend on the HD estimation's certainty, and thus bump amplitude: low bump amplitudes (low certainty) should lead to rapid bump shifts, whereas high bump amplitudes (high certainty) show lead to slower ones. Recent experimental evidence[10] suggests that bump amplitude varies in navigating rodents, and this amplitude modulates the speed with which their HD system reacts to changing absolute HD observations - in line with our predictions.

By restricting ourselves to an analytically tractable ring attractor network, we were able to almost exactly map the certainty dynamics of the ideal-observer circKF to the bump amplitude dynamics. Having the network implement the circKF rather than a standard Kalman filter fully accounts for the circular symmetry of HD estimation. Thus, unlike previous work[23], our network does not suffer from imprecise inference once absolute HD observations strongly deviate from the current HD estimate. As a result, it yields fundamentally different predictions for strongly conflicting absolute HD direction cues (Fig. 1e). Specifically, since in the circular Kalman filter a conflicting absolute observation (>90 deg from the current estimate) could yield a reduction in certainty, our network dynamics would predict a transient decrease in bump amplitude following a conflicting observation. Further, our network automatically adjusts its cue integration weights (Fig. 5c) to perform close-to-optimal Bayesian inference for absolute HD observations of varying reliability - from highly reliable to very unreliable or even completely absent observations. This stands in contrast to previous approaches[42], that required hand-tuned weights to show that continuous ring attractors can track orientation and compute the running circular average of an absolute HD stimulus. Lastly, our network is to our knowledge the first to fully account for the effect of *probabilistic* angular path integration in a principled way: unlike, e.g., the disc attractor in ref[43], the bump amplitude decay in our network matches the quadratic certainty decay of the ideal Bayesian observer in absence of absolute HD observations. We would expect to observe such a decay in biological ring attractors implementing Bayesian inference once absolute HD observations are removed.

Even though our Bayesian HD tracking algorithm requires keeping track of the HD estimate's uncertainty, we have shown that imperfectly tuned ring attractor networks can track head direction reasonably well. In fact, even strict attractor networks with a fixed amplitude, and fixed associated uncertainty, can perform close-to-Bayesian cue integration (Figs. 5b; cf. also ref[44]). This result raises the question of why neurons should encode uncertainty in the first place. First and foremost, for some animals, uncertainty influences their behavior directly to improve their performance (e.g., refs[45–47]). As a prime example, the homing behavior of the desert ant[48] suggests that the performance gained from tracking one's uncertainty justifies the added complexity for doing so. Further, uncertainty appears to impact the neural encoding of other navigation-related variables. For example, when absolute visual cues are in conflict with path integration cues, grid cells in mouse medial entorhinal cortex are more likely to remap when the visual cues are more reliable[49]. Identifying how uncertainty ought to be reflected in their neural activity, as we do here, is required for a comprehensive understanding of the role of uncertainty in the brain's computations.

In summary, our work shows how ring attractors could implement dynamic Bayesian inference, even in networks that obey some biological constraints, such as the *Drosophila*'s HD system. We expect similar network motifs to be present in the HD systems of other animals, such as that of mice[9,10], monkeys[50], humans[51], or even in systems that yield three-dimensional HD cells, as those of bats[52]. More generally, we demonstrated how classic network motifs, like those common in ring attractor networks, can perform close-to-optimal Bayesian inference when considered in combination, and expect our results to generalize to other circular variables that are represented in ring attractor networks.

Acknowledgements

We thank Rachel Wilson for fruitful discussions and input throughout the whole research phase, and for her comments on the manuscript. We thank Habiba Noamany (Wilson lab, HMS) for assisting us navigating the neuprint database, and for informed comments on the manuscript. We would further like to thank Johannes Bill & Albert Chen (Drugowitsch lab, HMS) for discussions and feedback on the manuscript, Philipp Reinhard for going on a typo hunt in the SI, and the entire Drugowitsch lab for valuable and insightful discussions.

The work was funded by the NIH (R34NS123819; J.D.), the James S. McDonnell Foundation (Scholar Award #220020462; J.D.), the Swiss National Science Foundation (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.).

Author contributions

Conceptualization, A.K., M.A.B., J.D.; Methodology, A.K., J.D.; Software, A.K.; Formal analysis, A.K., J.D.; Investigation, A.K, M.A.B., J.D.; Resources, J.D; Writing - Original Draft: A.K., J.D.; Writing - Review & Editing: A.K., M.A.B., J.D.; Visualization, A.K.; Supervision, J.D.; Funding Acquisition, A.K., J.D.

Declaration of interests

The authors declare no competing interests.

Methods

Ideal observer model: the circular Kalman filter

Our ideal observer model - the circular Kalman filter (circKF)[21] - performs dynamic Bayesian inference for circular variables. It computes the posterior probability 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 $v_{0:t} = \{v_0, v_{dt}, \ldots v_t\}$ with $v_\tau \in \mathbb{R}$, and absolute HD observations $z_{0:t} = \{z_0, z_{dt}, \ldots z_t\}$ with $z_\tau \in [-\pi, \pi]$. Specifically, we assume that these observations are generated from some true angular velocity $\dot{\phi}_t$ and HD ϕ_t, whose observations are corrupted by zero-mean noise at each point in time, via

$$v_t|\dot{\phi}_t \sim \mathcal{N}\left(\dot{\phi}_t, \frac{1}{\kappa_v\, dt}\right), \tag{5}$$

$$z_t|\phi_t \sim \mathcal{VM}\left(\phi_t, \sqrt{2\kappa_z\, dt}\right). \tag{6}$$

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 denote the precision of the angular velocity and absolute HD observations, respectively. Note that as $dt \to 0$, the precision of both angular velocity and absolute 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". More formally, the square-root scaling of the absolute HD observation precision with $\sqrt{2\kappa_z\, dt}$ ensures that the Fisher information of the observations about the true HD scales linearly in time and κ_z in the continuum limit $dt \to 0$ (ref[21], Theorem 2). The same applies to the dt^{-1} scaling of the Gaussian variance of the angular velocity observations, again achieving a Fisher information that scales linearly in time.

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

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

with a diffusion coefficient that decreases with κ_ϕ. In Results, we assume $\kappa_\phi \to 0$, implying that HD can change arbitrarily across consecutive time steps, which was sufficient to convey intuition into the algorithm's workings. However, when simulating stochastic HD trajectories, we assume they evolve according to Eq. (7) with $\kappa_\phi > 0$, which needs to be accounted for when performing inference. Thus, we here assume a non-zero κ_ϕ for completeness and reproducibility.

The circKF in Eqs. (1) and (2) assumes that the posterior distribution over HD 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 non-circular variables.

An alternative parametrization of the von Mises distribution to its mean μ_t and precision κ_t, are its natural parameters, $\theta_t = (\kappa_t \cos \mu_t, \kappa_t \sin \mu_t)^T$. Geometrically, the natural parameters can be interpreted as the Cartesian coordinates of a "probability vector", and (μ_t, κ_t) as its polar coordinates (Fig. 2b). As we show in the SI, the natural parameter parametrization makes including absolute HD observations (Eq. (6)) in the circKF straightforward. In fact, it becomes a vector addition. In contrast, including angular velocity observations (Eq. (5)) is mathematically intractable, such that the circKF relies on an approximation method called projection filtering[20] to find closed-form dynamic expressions for posterior mean and certainty (see ref[21] for technical details, and the SI for a more accessible description of the circKF).

Taken together, the circKF for the model specified by Eqs. (5)-(7) reads:

$$d\mu_t = \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t\, dt + \frac{\sqrt{2\kappa_z\, dt}}{\kappa_t} \sin(z_t - \mu_t), \tag{8}$$

$$d\kappa_t = -\frac{f(\kappa_t)}{2(\kappa_\phi + \kappa_v)} \kappa_t dt + \sqrt{2\kappa_z\, dt}\, \sin(z_t - \mu_t), \tag{9}$$

where $f(\kappa_t)$ is a monotonically increasing nonlinear function,

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

and $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of the first kind of order 0 and 1, respectively. Setting $\kappa_\phi \to 0$ yields Eqs. (1) and (2). Importantly, setting $\kappa_\phi \to 0$ does not conceptually change the general vector operations we present in Fig. 2.

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 non-linearity by a quadratic decay:

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

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 $\mathbf{r}_t \in \mathbb{R}^N$ of a neural population with N neurons. Thereby, we focused on the simplest kind of network model that supports such an approximation, which is of the form:

$$d\mathbf{r}_t = -\frac{1}{\tau}\mathbf{r}_t\, dt - g(\mathbf{r}_t)\mathbf{r}_t\, dt + W \cdot \mathbf{r}_t\, dt + \mathbf{I}_t^{\text{ext}}, \tag{12}$$

where τ is the network time constant, $g : \mathbb{R}^N \to \mathbb{R}_+$ is a scalar nonlinearity, and the elements of $\mathbf{r}_t$ are assumed to be ordered by the respective neuron's preferred HD, $\phi_1, \ldots, \phi_N$ (see Eq. (3)). We decomposed the recurrent connectivity matrix into $W = \frac{w_0}{2} W^{\text{const}} + w_1^{\text{even}} W^{\cos} + w_1^{\text{odd}} W^{\sin}$, where W^{const} denotes a matrix with constant entries, and $W^{\cos}$ and $W^{\sin}$ refer to cosine- and sine-shaped connectivity profiles (Fig. 4a). Specifically, due to the network's circular symmetry, the entries of these matrices only depend on the relative distance in preferred HD, and are given by $W_{ij}^{\text{const}} = \frac{2}{N}$,

$W_{ij}^{\text{cos}} = \frac{2}{N}\cos(\phi_i - \phi_j)$, and $W_{ij}^{\text{sin}} = \frac{2}{N}\sin(\phi_i - \phi_j)$. 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 maximum input in the infinitesimal time bin dt.

As described in Results, we assume the population activity $\mathbf{r}_t$ 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 the SI, the described network dynamics thus lead to the following dynamics of the cosine-profile parameters μ_t and κ_t:

$$d\mu_t = w_1^{\text{odd}}\,dt + \frac{I_t}{\kappa_t}\sin(\Phi_t - \mu_t), \tag{13}$$

$$d\kappa_t = \left(w_1^{\text{even}} - \frac{1}{\tau}\right)\kappa_t\,dt - g(\mathbf{r}_t)\kappa_t\,dt + I_t\cos(\Phi_t - \mu_t). \tag{14}$$

To derive these dynamics, we assumed the following:

1. The network is *rate-based*.

2. 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 Fig. S2 demonstrates only a very weak dependence of our results on the exact number of neurons in the network.

3. 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 the μ_t and κ_t encoded by the network (Eqs. (13) & (14)) 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) & (9)). Specifically, we find for the network parameters:

- Odd recurrent connectivities are modulated by angular velocity observations, $w_1^{\text{odd}} = \frac{\kappa_v}{\kappa_\phi + \kappa_v} v_t$, which shifts the activity profile without changing its amplitude[7,8].

- Absolute 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., $I_t = \sqrt{2\kappa_z\,dt}$. The inputs might contain additional Fourier modes (e.g., a constant baseline), but those do not affect the dynamics in Eqs. (13) and (14).

- The even component of the recurrent excitatory input needs to exactly balance the internal activity decay, i.e., $w_1^{\text{even}} - \frac{1}{\tau} = 0$.

- The decay nonlinearity is modulated by the reliability of the angular velocity observations, and is given by $g(\mathbf{r}_t) = \frac{1}{2(\kappa_\phi + \kappa_v)} f\big(\kappa(\mathbf{r}_t)\big)$, where $f(\cdot)$ equals the nonlinearity that governs the certainty decay in the circKF (Eq. (10)). This can be achieved, e.g., through interaction with an inhibitory neuron (or a pool of inhibitory neurons) with activation function $f(\cdot)$ that computes the activity bump's amplitude $f\big(\kappa(\mathbf{r}_t)\big)$.

Quadratic approximation

To gain additional analytical tractability, we further approximated the recurrent inhibition by one that takes the form $g(\mathbf{r}_t)\mathbf{r}_t \to w^{\text{quad}}\left(\pi\sum_{i=1}^{N}[r_t^{(i)}]_+\right)\cdot\mathbf{r}_t$, 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 aplitude (see below). In the large population limit, $N\to\infty$, this leads to the amplitude dynamics (see SI for derivation)

$$d\kappa_t = \left(w_1^{\text{even}} - \frac{1}{\tau}\right)\kappa_t\,dt - w^{\text{quad}}\kappa_t^2 dt + I_t\cos(\Phi_t - \mu_t). \tag{15}$$

The dynamics of the phase μ_t does not depend on the form of $g(\cdot)$ and thus remains to be given by Eq. (13). If we set the network parameters to $w^{\text{quad}} = \frac{1}{\kappa_\phi+\kappa_v}$ and $w^{\text{even}} - \frac{1}{\tau} = \frac{1}{\kappa_\phi+\kappa_v}$, while sensory input, i.e. angular velocity v_t and absolute HD observations z_t, enter in the same way as before, the network implements the quadratic approximation to the circKF (Eqs. (8) & (11)).

Ring-attractor networks

In absence of absolute HD observations ($I_t = 0$), the amplitude dynamics in Eq. (15) has a stable **fixed point** at $\kappa^* = \frac{w^{\text{even}}-1/\tau}{w^{\text{quad}}}$ and no preferred phase, making it a ring-attractor network. Linearizing the κ_t dynamics around this fixed points reveals that it is approached with **decay speed** $\beta = w^{\text{even}} - \frac{1}{\tau}$. A large value of β denotes faster dynamics and thus indicates more rigid attractor dynamics. In the limit of $\beta\to\infty$ we consider the attractor to be a "strict" attractor that, upon any perturbation, immediately moves back to its attractor state. For the quadratic approximation network, we find $\kappa^* = 1$ and $\beta = \frac{1}{\kappa_\phi+\kappa_v}$. Further, in our simulations in Fig. 5, we explored network dynamics with a range of κ^* and β values by adjusting network parameters accordingly.

Multiple population network

We extended the single population network dynamics, Eq. (12), 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 absolute HD observations. The resulting network dynamics become (see SI for details):

$$\begin{aligned}
\dot{\mathbf{r}}_t^{HD} &= -\frac{1}{\tau_{HD}}\mathbf{r}_t^{HD} + W_{HD\leftarrow HD}\cdot\mathbf{r}_t^{HD} + W_{HD\leftarrow AV+}\cdot\mathbf{r}_t^{AV+} + W_{HD\leftarrow AV-}\cdot\mathbf{r}_t^{AV-} && (16)\\
&\quad + (W_{HD\leftarrow INH}\mathbf{r}_t^{INH}) \circ \mathbf{r}_t^{HD} + \mathbf{I}_t^{ext}, && (17)\\
\tau_{AV^+}\dot{\mathbf{r}}_t^{AV^+} &= -\mathbf{r}_t^{AV^+} + (o^{AV} + v_t)W_{AV^+\leftarrow HD}\cdot\mathbf{r}_t^{HD}, && (18)\\
\tau_{AV^-}\dot{\mathbf{r}}_t^{AV^-} &= -\mathbf{r}_t^{AV^-} + (o^{AV} - v_t)W_{AV^-\leftarrow HD}\cdot\mathbf{r}_t^{HD}, && (19)\\
\tau_{INH}\dot{\mathbf{r}}_t^{INH} &= -\mathbf{r}_t^{INH} + W_{INH\leftarrow HD}\cdot[\mathbf{r}_t^{HD}]_+ + W_{INH\leftarrow INH}\cdot\mathbf{r}_t^{INH}. && (20)
\end{aligned}$$

Here, the $W_{to\leftarrow from}$ denote connectivities within and between populations, and o^{AV} is a constant activity baseline in the AV populations.

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. (12). 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[33]). Specifically, we focused on the connectivities between EPG, PEN1 and $\Delta 7$ neurons (which in our model corresponds to HD, AV+/- and INH neurons), sorted according to anatomical regions within the ellipsoid body and the protocerebral bridge (Fig. 6b). Thereby, we used total number of synaptic connections between two regions as a proxy for connection strength. We further assumed that interactions within AV+/- populations and between AV+/- and INH populations were negligible. The resulting connectivity profile in Fig. 6c was determined by matching the *Drosophila* connectome as closely as possible, while allowing for modulation of the across-population connection strengths c_0^{HD},c_1^{HD}, $c^{AV^\pm\leftarrow HD}$, $c^{HD\leftarrow AV^\pm}$., $c_0^{INH\leftarrow HD}$, $c_1^{INH\leftarrow HD}$, c_0^{INH}, c_1^{INH}, and $c^{HD\leftarrow INH}$. We specify the specific analytic functions we used to create the connectivity matrix in Fig. 6c in the SI, where we also compute the connection strengths analytically.

Simulation details

Numerical integration

Our simulations in Figs. 4 and 5 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[54], 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), its quadratic approximation, Eq. (11), and the network dynamics, Eqs. (12) and (16)-(20).

Performance measures

To measure performance, in Figs. 4f, 5b and 5d we computed the circular average distance[53] 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, $0 \le |m_1| \le 1$ denotes an empirical precision (or 'inference precision'), 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 inference precision is related to the circular variance via Var$_{circ} = 1 - |m_1|$. In SI 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 all absolute HD observation reliabilities κ_z in Figs. 4f and 5b. For the inset of Fig. 5b, and for Fig. 5d, we additionally performed a grid search over the fixed-point amplitude κ^* (inset of Fig. 5b), or both the fixed-point amplitude κ^* and of the inverse time constant β (Fig. 5d). For each setting of κ^* and β we assessed the performance by computing an average over this performance for a range of observation reliability κ_z, weighted by how likely each observation reliability is a-priori assumed to be. The latter was specified by a log-normal prior, $p(\kappa_z) =$ Lognormal$(\mu_{\kappa_z}, \sigma^2_{\kappa_z})$, favouring intermediate reliabilitiy levels. We chose $\mu_{\kappa_z} = 0.5$ and $\sigma^2_{\kappa_z} = 1$ for the prior parameters, but our results did not strongly depend on this parameter choice. The performance loss shown in Fig. 5d also relied on such a weighted average across κ_z's for a particle filter benchmark (PF, see SI for details). The loss itself was then defined as $1 - \frac{\text{Performance}}{\text{Performance PF}}$.

Update weights

In Fig. 5c, we computed the weight with which a single 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, $w = \frac{\Delta\mu_t}{\pi} = \frac{1}{\pi}\tan^{-1}\frac{\alpha(\kappa_z\, dt)}{\kappa_t}$. Here, $\alpha(\kappa_z\, dt)$ denotes a function that ensures a linear scaling of the Fisher information with sampling time step (see ref[21], Theorem 2, for details about this function). Thus, by design of the observation model, the Fisher information of a single observation with reliability κ_z during a time interval Δt is given by $I_{z_t}(\phi_t) = \kappa_z\,\Delta t$. We plot the weight as a function of the Fisher information of a single update (how reliable is the observation?) and the Fisher information of the current HD estimate (how certain is the current estimate?), which is given by

$$I_{\mu_t,\kappa_t}(\phi_t) = \mathbb{E}\left[\left(\frac{\partial}{\partial\phi}\log\mathcal{VM}(\phi,\mu_t,\kappa_t)\right)^2\right] = \kappa_t\frac{I_1(\kappa_t)}{I_0(\kappa_t)}. \tag{21}$$

Details on numerical simulations

In our network simulations, we set the network decay constant τ to an arbitrary, but non-zero, 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 above. From the neural activity vector $\mathbf{r}_t$, we retrieved the natural parameters $\boldsymbol{\theta}_t$ with a decoder matrix $A = (\cos(\boldsymbol{\phi}^{(i)}), \sin(\boldsymbol{\phi}^{(i)}))^T$, such that $\boldsymbol{\theta}_t = A\cdot\mathbf{r}_t$, and subsequently computed the position of the bump by $\phi_t = \arctan 2(\theta_2, \theta_1)$, and the encoded certainty (length of the population vector) by $\kappa_t = \sqrt{\theta_1^2 + \theta_2^2}$.

In all our simulations, times are measured in units of inverse diffusion time constant κ_ϕ, where we set $\kappa_\phi = 1s$ for convenience. Figures were generated based on simulations with the following parameters:

- Figure 4e: $\kappa_v = 2$, $\kappa_z = 10$ (during 'Visual cue' bout), $\kappa_z = 0$ (during 'Darkness' bout), $\Delta t = 0.01$.
- Figure 4f, 5b, 5d: $\kappa_v = 1$, $T = 20$, $\Delta t = 0.01$. Results are averages over $P = 5000$ simulation runs.
- Figure 5e: $\kappa_v = 1$, $\kappa_z = 1$, $T = 10$, $\Delta t = 0.01$.

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. Jupyter notebooks, Python scripts, and data to reproduce the figures will be made available upon acceptance of the manuscript.

References

1. Wang, X.-J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–463 (2001).
2. Compte, A. Computational and in vitro studies of persistent activity: Edging towards cellular and synaptic mechanisms of working memory. *Neuroscience* **139**, 135–151 (2006).
3. Hansel, D. & Sompolinsky, H. Modeling Feature Selectivity in Local Cortical Circuits. *Methods Neuronal Model. Ions Netw.* 69 (1998).
4. Rademaker, R. L., Tredway, C. H. & Tong, F. Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *J. Vis.* **12**, 21 (2012).
5. Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J. & Curtis, C. E. Joint representation of working memory and uncertainty in human cortex. *Neuron* **109**, 3699-3712.e6 (2021).
6. Knierim, J. J. & Zhang, K. Attractor Dynamics of Spatially Correlated Neural Activity in the Limbic System. *Annu. Rev. Neurosci.* **35**, 267–285 (2012).
7. Zhang, K. Representation of Spatial Orientation by the Intrinsic Dynamics of the Head-Direction Cell Ensemble: A Theory. *J. Neurosci.* **16**, 2112–2126 (1996).
8. Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. A Model of the Neural Basis of the Rat's Sense of Direction. *Adv. Neural Inf. Process. Syst.* 10 (1995).
9. Peyrache, A., Lacroix, M. M., Petersen, P. C. & Buzsáki, G. Internally organized mechanisms of the head direction sense. *Nat. Neurosci.* **18**, 569–575 (2015).
10. Ajabi, Z., Keinath, A. T., Wei, X.-X. & Brandon, M. P. Population dynamics of the thalamic head direction system during drift and reorientation. 2021.08.30.458266 (2021) doi:10.1101/2021.08.30.458266.
11. Redish, A. D., Elga, A. N. & Touretzky, D. S. A coupled attractor model of the rodent head direction system. *Netw. Comput. Neural Syst.* **7**, 671–685 (1996).
12. Seelig, J. D. & Jayaraman, V. Neural dynamics for landmark orientation and angular path

integration. *Nature* **521**, 186–191 (2015).

13. Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the Drosophila central brain. *Science* **356**, 849–853 (2017).
14. Turner-Evans, D. *et al.* Angular velocity integration in a fly heading circuit. *eLife* **6**, e23496 (2017).
15. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
16. Dehaene, G. P., Coen-Cagli, R. & Pouget, A. Investigating the representation of uncertainty in neuronal circuits. *PLOS Comput. Biol.* **17**, e1008138 (2021).
17. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME J. Basic Eng.* **82**, 35–45 (1960).
18. Kalman, R. E. & Bucy, R. S. New Results in Linear Filtering and Prediction Theory. *J. Basic Eng.* **83**, 95–108 (1961).
19. Kurz, G., Gilitschenski, I. & Hanebeck, U. D. Recursive Bayesian filtering in circular state spaces. *IEEE Aerosp. Electron. Syst. Mag.* **31**, 70–87 (2016).
20. Brigo, D., Hanzon, B. & Le Gland, F. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli* **5**, 495–534 (1999).
21. Kutschireiter, A., Rast, L. & Drugowitsch, J. Projection Filtering with Observed State Increments with Applications in Continuous-Time Circular Filtering. *IEEE Trans. Signal Process.* 1–1 (2022) doi:10.1109/TSP.2022.3143471.
22. Murray, R. F. & Morgenstern, Y. Cue combination on the circle and the sphere. *J. Vis.* **10**, 15–15 (2010).
23. Wilson, R. & Finkel, L. A Neural Implementation of the Kalman Filter. *Adv. Neural Inf. Process. Syst.* **22**, 9 (2009).
24. Ben-Yishai, R., Bar-Or, R. L. & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci.* **92**, 3844–3848 (1995).

25. Georgopoulos, A., Kettner, R. & Schwartz, A. Primate motor cortex and free arm movements to visual targets in three- dimensional space. II. Coding of the direction of movement by a neuronal population. *J. Neurosci.* **8**, 2928–2937 (1988).
26. Johnson, A., Seeland, K. & Redish, A. D. Reconstruction of the postsubiculum head direction signal from neural ensembles. *Hippocampus* **15**, 86–96 (2005).
27. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–8 (2006).
28. Beck, J. M., Latham, P. E. & Pouget, A. Marginalization in Neural Circuits with Divisive Normalization. *J. Neurosci.* **31**, 15310–15319 (2011).
29. Lyu, C., Abbott, L. F. & Maimon, G. *A neuronal circuit for vector computation builds an allocentric traveling-direction signal in the* Drosophila *fan-shaped body*. http://biorxiv.org/lookup/doi/10.1101/2020.12.22.423967 (2020) doi:10.1101/2020.12.22.423967.
30. Xie, X., Hahnloser, R. H. R. & Seung, H. S. Double-ring network model of the head-direction system. *Phys. Rev. E - Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **66**, 9–9 (2002).
31. Compte, A. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cereb. Cortex* **10**, 910–923 (2000).
32. Turner-Evans, D. B. *et al.* The Neuroanatomical Ultrastructure and Function of a Biological Ring Attractor. *Neuron* S0896627320306139 (2020) doi:10.1016/j.neuron.2020.08.006.
33. Scheffer, L. K. *et al.* A connectome and analysis of the adult Drosophila central brain. *eLife* **9**, e57443 (2020).
34. Hulse, B. K. *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).
35. Green, J. *et al.* A neural circuit architecture for angular integration in Drosophila. *Nature* **546**, 101–106 (2017).

36. Omoto, J. J. *et al.* Visual Input to the Drosophila Central Complex by Developmentally and Functionally Distinct Neuronal Populations. *Curr. Biol.* **27**, 1098–1110 (2017).
37. Fisher, Y. E., Lu, J., D'Alessandro, I. & Wilson, R. I. Sensorimotor experience remaps visual input to a heading-direction network. *Nature* (2019) doi:10.1038/s41586-019-1772-4.
38. Kim, S. S., Hermundstad, A. M., Romani, S., Abbott, L. F. & Jayaraman, V. Generation of stable heading representations in diverse visual scenes. *Nature* 1–6 (2019) doi:10.1038/s41586-019-1767-1.
39. van Bergen, R. S., Ji Ma, W., Pratte, M. S. & Jehee, J. F. M. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18**, 1728–1730 (2015).
40. Sun, X., Mangan, M. & Yue, S. An analysis of a ring attractor model for cue integration. *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.* **10928 LNAI**, 459–470 (2018).
41. Sun, X., Yue, S. & Mangan, M. A decentralised neural model explaining optimal integration of navigational strategies in insects. *eLife* **9**, e54026 (2020).
42. Esnaola-Acebes, J. M., Roxin, A. & Wimmer, K. *Bump attractor dynamics underlying stimulus integration in perceptual estimation tasks*. http://biorxiv.org/lookup/doi/10.1101/2021.03.15.434192 (2021) doi:10.1101/2021.03.15.434192.
43. Carroll, S., Josić, K. & Kilpatrick, Z. P. Encoding certainty in bump attractors. *J. Comput. Neurosci.* **37**, 29–48 (2014).
44. Takiyama, K. Bayesian estimation inherent in a Mexican-hat-type neural network. *Phys. Rev. E* **93**, 052303 (2016).
45. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
46. Piet, A. T., Hady, A. E., Brody, C. D., El Hady, A. & Brody, C. D. Rats adopt the optimal timescale for evidence integration in a dynamic environment. *Nat. Commun.* **9**, 1–12 (2018).

47. Fetsch, C. R., Turner, A. H., DeAngelis, G. C. & Angelaki, D. E. Dynamic reweighting of visual and vestibular cues during self-motion perception. *J. Neurosci.* **29**, 15601–15612 (2009).
48. Merkle, T. & Wehner, R. Desert ants use foraging distance to adapt the nest search to the uncertainty of the path integrator. *Behav. Ecol.* **21**, 349–355 (2010).
49. Campbell, M. G., Attinger, A., Ocko, S. A., Ganguli, S. & Giocomo, L. M. Distance-tuned neurons drive specialized path integration calculations in medial entorhinal cortex. *Cell Rep.* **36**, 109669 (2021).
50. Robertson, R. G., Rolls, E. T., Georges-François, P. & Panzeri, S. Head direction cells in the primate pre-subiculum. *Hippocampus* **9**, 206–219 (1999).
51. Baumann, O. & Mattingley, J. B. Medial Parietal Cortex Encodes Perceived Heading Direction in Humans. *J. Neurosci.* **30**, 12897–12901 (2010).
52. Finkelstein, A. *et al.* Three-dimensional head-direction coding in the bat brain. *Nature* **517**, 159–164 (2015).
53. Mardia, K. V. & Jupp, P. E. *Directional Statistics*. 3 (John Wiley & Sons, 2000). doi:10.1002/9780470316979.
54. Kloeden, P. E. & Platen, E. *Numerical solution of stochastic differential equations*. (Springer, 2010).