Learning to represent continuous variables in heterogeneous neural networks

Ran Darshan*[1] and Alexander Rivkind*[2]

[1] *Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA*
[2] *Weizmann Institute of Science, Rehovot, Israel*
* *Both authors contributed equally*

Manifold attractors are a key framework for understanding how continuous variables, such as position or head direction, are encoded in the brain. In this framework, the variable is represented along a continuum of persistent neuronal states which forms a manifold attactor. Neural networks with symmetric synaptic connectivity that can implement manifold attractors have become the dominant model in this framework. In addition to a symmetric connectome, these networks imply homogeneity of individual-neuron tuning curves and symmetry of the representational space; these features are largely inconsistent with neurobiological data. Here, we developed a theory for computations based on manifold attractors in trained neural networks and show how these manifolds can cope with diverse neuronal responses, imperfections in the geometry of the manifold and a high level of synaptic heterogeneity. In such heterogeneous trained networks, a continuous representational space emerges from a small set of stimuli used for training. Furthermore, we find that the network response to external inputs depends on the geometry of the representation and on the level of synaptic heterogeneity in an analytically tractable and interpretable way. Finally, we show that a too complex geometry of the neuronal representation impairs the attractiveness of the manifold and may lead to its destabilization. Our framework reveals that continuous features can be represented in the recurrent dynamics of heterogeneous networks without assuming unrealistic symmetry. It suggests that the representational space of putative manifold attractors in the brain dictates the dynamics in their vicinity.

1 Introduction

Stimulus-specific neuronal activity is known to persist even in the absence of stimuli (see [Wang, 2001] for review). Such persistent states of neuronal circuits were hypothesized to be sustained by recurrent synaptic connections [Hebb, 1949, Durstewitz et al., 2000] and are refereed to as neural *attractors* [Hopfield, 1982, Amit, 1992]. In a wide variety of brain systems the variables that are represented by such

a persistent neuronal activity are continuously-valued [Funahashi et al., 1993, Romo et al., 1999]. In particular, neurons in the navigation system represent continua of animal's directional heading [Taube et al., 1990, Seelig and Jayaraman, 2015], speed [Kropff et al., 2015] and locations [O'Keefe and Dostrovsky, 1971, Hafting et al., 2005], while neuronal activity in prefrontal and posterior parietal neocortices correlates with stimulus orientation [Christophel et al., 2017] and its spatial location [Funahashi et al., 1993].

The common framework to study such continuous internal representations is the theory of computations by *manifold* attractor networks [Amari, 1977, Ben-Yishai et al., 1995, Seung, 1996, Burak and Fiete, 2009, Wimmer et al., 2014, Hansel and Mato, 2013, Chaudhuri et al., 2019, Gardner et al., 2021]. Here, the variable of interest is represented as a point in the space of neural activity, with the continuum of values forming a low dimensional manifold of attractor states in the high dimensional space of neural firing rates. Neural dynamics converge toward the manifold attractor, and are thus robust to perturbations that could kick the state away from it. On the other hand, due to the continuum of stable states, stability is *marginal* [Ben-Yishai et al., 1995] along the manifold - a perturbation in this direction does not face either converging nor repelling forces [Durstewitz et al., 2000].

Models of manifold attractors tend to extensively rely on symmetry assumptions [Amari, 1977, Brody et al., 2003, Machens and Brody, 2008, McNaughton et al., 2006, Burak and Fiete, 2009, Mastrogiuseppe and Ostojic, 2018, Beiran et al., 2020]. For example, in models of the head direction system [Zhang, 1996], or in representations in the primary visual [Ben-Yishai et al., 1995] and prefrontal cortices [Compte et al., 2000], the connectivity is constructed according to a rotation symmetry principle, in which recurrent interactions depends only on the distance between the preferred direction of the neurons. As a result, the connectivity profile of all neurons are the same up to a rotation of the angular feature (Fig.1Ai-ii, see also [Mastrogiuseppe and Ostojic, 2018] for an extension for these connectivity rules). Under some general conditions on the recurrent interactions, such as short-range excitation and long-range inhibition, a manifold of attractors appears (Fig.1Aiii). We call these classical models which are based on a symmetry assumption, symmetric-connectome attractor networks.

The symmetry assumption is highly unlikely in real biological systems, in which heterogeneity in synaptic connections are abundant [Braitenberg and Schüz, 2013]. Yet, in the aforementioned models any deviation from perfect symmetry results in shattering of the continuous attractor into a few isolated attractors and as a result to a fast deterioration of the computational capabilities of the network, such as the loss of persistent representation [Zhang, 1996, Tsodyks and Sejnowski, 1995, Renart et al., 2003, Itskov et al., 2011] or imperfect path integration [Burak and Fiete, 2009]. Furthermore, as a direct outcome of the symmetry in connectivity, the activity profile of neurons in symmetric-connectome models are identical. This is in sharp contrast to neurons in reality that can show many degrees of diversity [Barak et al., 2013, Finkelstein et al., 2015, Chaudhuri et al., 2019, Fisher et al., 2019].

Beyond the constraints on the synaptic connections and neuronal tuning profile, symmetric-connectome models imply, by their construction, a perfect *geometry* of the high dimensional neuronal

representation along the manifold. In particular, a one-dimensional manifold will have a perfect circular shape when projected to the leading principal components of neural activity (Fig.1Aiii) and average activity which is independent of the location. Evidence of such a perfection of geometry in the brain are lacking: while recent studies suggest that neuronal representations of continuous features might exhibit a topology of a ring [Chaudhuri et al., 2019, Rubin et al., 2019] or a torus [Gardner et al., 2021], in agreement with the manifold attractor hypothesis, there is no evidence for a perfect geometrical symmetry in these representations. Symmetric-connectome attractor models are thus inconsistent with synaptic heterogeneity and diverse tuning profiles of neurons and unable to support the imperfect geometries observed in neural manifolds.

To loosen the aforementioned idealistic assumptions, one may consider a manifold that emerges as a result of learning rather than via a pre-engineered connectivity. Indeed, trained on a wide variety of tasks, ranging from integration of evidence [Mante et al., 2013], to path integration [Sorscher et al., 2020, Cueva et al., 2019] and natural language processing [Maheswaranathan et al., 2019], recurrent neural networks (RNNs) exhibit manifold attractor dynamics. In these unconstrained setting, where symmetry in the connectivity is not imposed and when the training is done in the presence of synaptic heterogeneity, neither the connectivity nor the geometry of the representation is expected to exhibit a perfect symmetry. However, theoretical understanding of how manifold attractors emerge in these models and how the neuronal representation in the trained RNNs shapes their dynamical properties and computational power, remain elusive.

How can a continuum of persistent states co-exists with synaptic heterogeneity and diversity of neuronal representation observed in experiments? What is the effect of such diversity and of the manifold's geometry on the dynamics along the manifold and in its vicinity? We present a minimal and solvable model of a trained recurrent network that we analyzed analytically in the large network limit and in which we relax the symmetry assumption in both synaptic connectivity and manifold geometry. As in real biological systems, tuning curves and connectivity patterns in the model are heterogeneous. Furthermore, our framework encompasses biological manifolds with non-symmetric geometry, in sharp contrast to existing theories that can only deal with symmetric manifolds. Finally, we connect the static properties, such as synaptic heterogeneity and imperfect geometry of the internal representation, to the computational properties of the attractor network, such as its robustness to perturbation and its response to external stimulus. Our work thus shows that diversity in synaptic connectivity and imperfections in neural representation can coexist with the hypothesis of neural manifold attractors, obviating the need to rely on symmetry principles.

2 Results

We studied networks consisting of N firing rate units:

$$\tau \frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x} + (\boldsymbol{W} + g\boldsymbol{J})\phi(\boldsymbol{x}) + \epsilon \boldsymbol{u} \tag{1}$$

where x_i is the total input into the i-th neuron, and the recurrent connectivity is decomposed into a structured connectivity matrix, $\boldsymbol{W}$, and a random matrix, $\boldsymbol{J}$. The strength of the external input, $\boldsymbol{u}$, is controlled by a parameters ϵ, and the level of synaptic heterogeneity in the recurrent network is controlled by a parameter g. We used a sigmoidal transfer function, $\phi(x)$, but it is possible to generalize our theory to other transfer functions.

We will use the term symmetric-connectome network models to refer to the case where the structured recurrent component is constructed based on symmetry principles, instead of via training, and in the absence of synaptic heterogeneity (i.e. $g = 0$, Fig.1A). Indeed, many computational and theoretical studies use these type of network models to explain the emergence of continuous representations in visual cortex, continuous and persistent activity in prefrontal cortex and persistent activity and path integration in the navigation system. In these type of models, a continuous periodic feature, ψ, is encoded in the network and the recurrent connectivity is invariant to rotations (Fig.1Ai-ii). In the absence of external inputs, all states are equally stable and form a manifold of attractors, i.e., activity of the neurons lies on a one dimensional manifold (Fig.1Aiii). Each state is a packet, or a 'bump' of localized activity. With external cue in present, the symmetry is broken and an appropriate state is selected from the continuum of states on the manifold. Figure 1B depicts both update of the memorized feature when the stimulus is present and periods of persistent activity when it is absent.

In what follows we will show that while symmetric-connectome network models [Ben-Yishai et al., 1995, Burak and Fiete, 2009, Beiran et al., 2020, Mastrogiuseppe and Ostojic, 2018], such as the one depicted in Fig.1A, are a possible implementation for computations based on manifold attractors, they are by far not the only ones.

2.1 Trained manifold attractors

In contrast to symmetric-connectome models, we assume that the structured component is *trained* rather than set *a priori*. We proceed with writing the structured recurrent connectivity in the following form:

$$\boldsymbol{W} = \boldsymbol{W}_{fb}\boldsymbol{W}_{out}^T \tag{2}$$

which allows to interpret this recurrent component as a mapping from the neural activity into a two dimensional (2D) representation which we denote by $\boldsymbol{z}$, via $\boldsymbol{W}_{out}$:

$$\boldsymbol{z} = \boldsymbol{W}_{out}^T\phi(\boldsymbol{x}) \tag{3}$$

and $\boldsymbol{W}_{fb}$ projects it back to the neurons [Jaeger, 2001, Sussillo and Abbott, 2009]. Specifically, we train the structured connectivity such that the output, $\boldsymbol{z}$, lies on a desired, predefined, manifold. The continuous feature ψ is then read out from the network through the angle of the 2D vector $\boldsymbol{z}$. To clarify notations we distinguish the manifold of internal states $\boldsymbol{\phi} \in M_\phi$ (Fig1A,Ciii) which is embedded in the N-dimensional neuronal state, from the pre-defined trained manifold projected in 2D, $\boldsymbol{z} \in M_z$ (e.g. see Fig.1Aiii, Fig.3 and Methods).

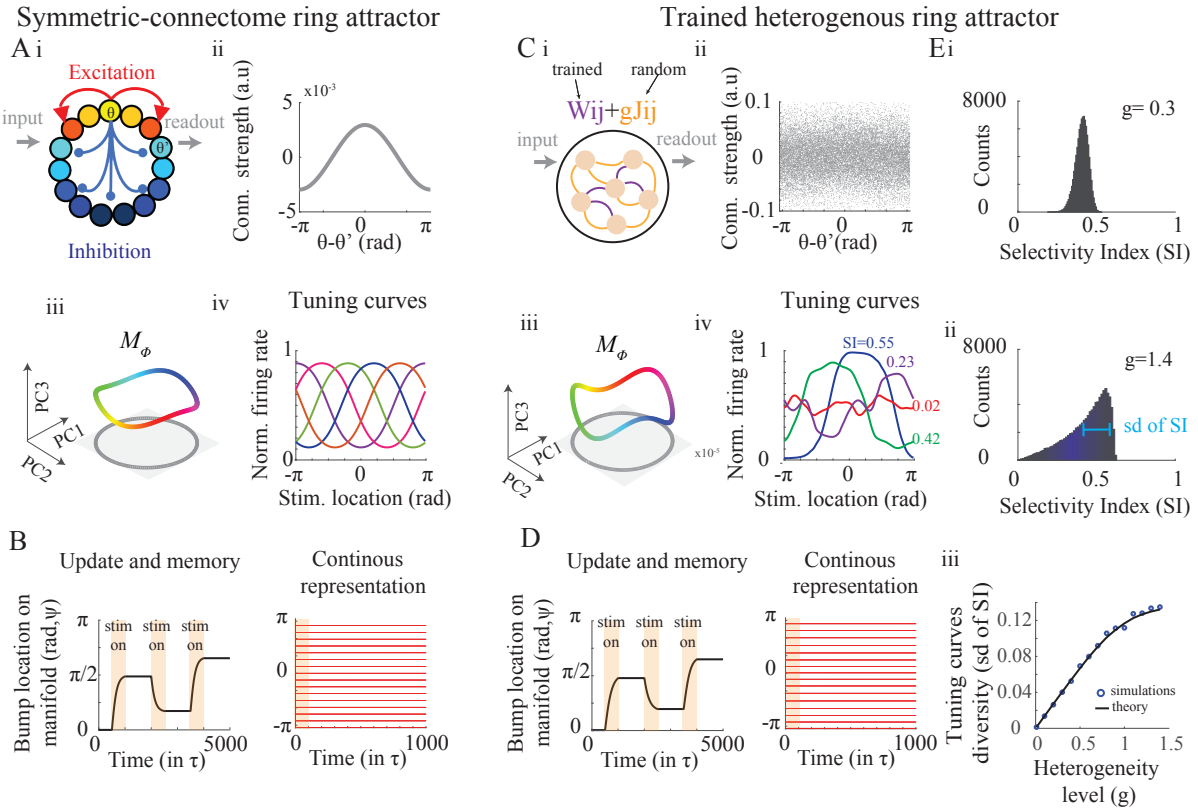

Figure 1: The symmetric-connectome ring attractor network and heterogeneous trained ring attracor network. A-B. Symmetric-connectome ring model. **Ai.** Cartoon of a ring network architecture. Neurons are aligned on a ring, in which the connectivity depends solely on the difference between the neuronal preferred directions, denoted by θ. **ii.** Rotation symmetry in the connectivity: Connectivity strengths coincide for all neurons in the network when plotted against the difference between the neuronal preferred directions. **iii.** The manifold attractor, denoted by M_ϕ, projected on the leading PCs of the neural representation (i.e. the tuning curves). Color indicates the decoded representation of the continuous feature (the angle of the decoder in Eq.(3)). Note that for a ring manifold the projection on the first two PCs exhibits a circular shape (gray). **iv.** Tuning curves of example neurons in the symmetric-connectome model are all identical, up to a symmetry for rotations. **B.** Left: Bump's location on the manifold (in radians) vs. time (in units of membrane time-constant). The internal representation of the feature (a 'bump' of activity) evolves when external input is applied (orange areas) and persists for a long time in the absence thereof (white areas). Right: Continuous internal representation of the feature. Red lines: position of the bump following initialization at that locations using the external input (in orange). The state persists in the absence of external inputs. **C-D.** Same as (A-B), but for a *trained* heterogeneous ring. Connectivity consists of random heterogeneous component, gJ_{ij} (orange connections) superimposed with the structured component, W_{ij} (purple); only the latter component is affected by training. Contrary to the case in panel **A**, connectivity is not fully determined by difference in preferred directions **(ii)** and tuning curves do exhibit diversity **(iv)**. **E.** Diversity in tuning curves increases with the level of synaptic heterogeneity. **Ei-ii.** Examples of distribution of selectivity index (SI), defined as the first Fourier component of the tuning curve (see Methods) for networks that were trained with different levels of synaptic heterogeneity. See also examples in Civ. **Eiii** Diversity in tuning curves (SD of SI, see Eii) increases with the heterogeneity level. Theory in Eq.(53)

Figure 1C shows an example of such a network that was trained in the presence of synaptic heterogeneity ($g > 0$) and without relying on symmetric-connectome. Specifically, we trained the network to produce a ring manifold, i.e. requiring a circle of radius A in the *decoder plane*: $\boldsymbol{z}(\psi) = A[\cos(\psi), \sin(\psi)]$. As a result of training, the activity in the network persists for a long time and the network memorizes the angular feature, until the representation gets updated through an external stimulus (Fig.1D). This is in contrast to models with pre-defined symmetric-connectome [Ben-Yishai et al., 1995, Mastrogiuseppe and Ostojic, 2018], where both memorization and update capabilities are impaired in presence of heterogeneous connectivity (Fig.S1).

The neuronal representation in the trained network is dramatically different from the symmetric-connectome counterpart. Due to the presence of synaptic heterogeneity, not all states are identical; both the population activity profile (the 'bump'), and the tuning curves of neurons in the model are highly heterogeneous. Figure 1Civ and Fig.S2 exemplifies such highly heterogeneous tuning curves. To quantify this, we solved the steady state of Eq.(1) and used it to derive the statistics of the tuning curves in the model (see Methods). The diversity in tuning curves increases monotonically with the level of synaptic heterogeneity in the network (Fig.1E). The analytical calculations are in good agreement with the simulations (compare solid line with circles).

Did the learning result in a structured component $\boldsymbol{W}$ that merely compensates for the heterogeneous component $g\boldsymbol{J}$, such that it restores the rotational symmetry of the synaptic connectivity? We find that this is not the case. The recurrent connectivity is not solely a function of the distance in preferred directions, as is the case for the symmetric-connectome models (compare Fig1Aii with Fig1Cii). This is true also when considering only the structured component of the recurrent interactions, which is not symmetric (Fig.S2). In fact, it is only after considerable averaging of synaptic inputs across neurons that we can observe that the connectivity profile depends on the distance in neuronal preferred directions (Fig.S2). Finally, symmetry is restored only in a specific case in which we train the network in the absence of any heterogeneity ($g = 0$).

Figure 1C shows the existence of a manifold attractor without relying on symmetry in the synaptic interactions or neuronal tuning curves. Yet, in this case due to the circular shape of the manifold in the 2D plane ($\boldsymbol{z}(\psi)$) the second order statistics of the neuronal representation is invariant for rotations and the average population activity is the same at each point on the manifold (black line in Fig.2A). Consequently, the principal components (PCs) of the neural representation are the spatial Fourier modes and the projection of the manifold on the leading PCs features a circular shape (note the gray circle in Fig.1Ciii). However, symmetry in the second order statistics is not necessary for the emergence of manifolds attractors. Indeed, figure 2 shows an example of a trained manifold which is not circular, the average population activity varies along the manifold (gray lines in Fig.2A), the distribution of preferred direction can be non-uniform (Fig.2F), the second order statistics of the representation show no rotation symmetry (Fig.2G) and the PCs are not pure Fourier modes (not shown). However, similarly to the ring manifold, these networks can memorize and update the representation of a continuous feature based on manifold attractor dynamics (Fig.3).

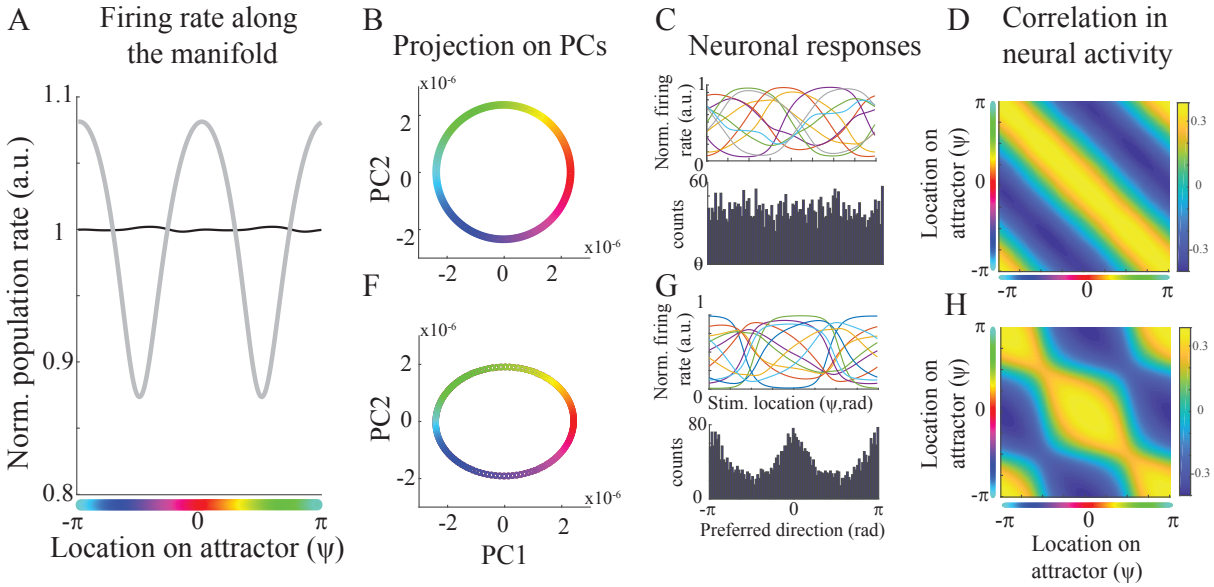

Figure 2: Neural representations of manifold attractors A. Normalized population activity along the manifold (Eq.(9)) for a ring manifold (black, B) and the ellipse manifold in E (gray). In a ring manifold the total population activity is constant at each location of the manifold (up to small fluctuations that arise in small networks). **B.** Projection of the neural representation on the top two leading principal components for a ring manifold (see parametarization of the trained manifold in Methods). The represented feature, ψ, is color coded. **C.** Top: Tuning curves of neurons in the trained network. Bottom: distribution of preferred directions in the network. **D.** Correlation across population activity at different locations on the manifold, $C(\psi,\psi') = \langle \phi(\boldsymbol{x}(\psi))\phi(\boldsymbol{x}(\psi'))\rangle$. For a perfect ring geometry the correlation function exhibits a rotation symmetry ($C(\psi,\psi') = C(\psi-\psi')$, i.e. matrix is circulant). **E-F**. Same as (B)-(D) but for an ellipse manifold in which there is no rotational symmetry in the representation

To conclude, we obtain a family of manifold attractor networks that lack obvious symmetry by introducing an appropriate structured low-rank component of the form of Eq.(2) on top of synaptic heterogeneity. The hand-crafted symmetric connectome networks that have been theorized to underlie computations in the brain is only one specific choice of the structured connectivity $\boldsymbol{W}$, and without the heterogeneous part ($g = 0$), while other, learnable, solutions do not rely on any clear symmetry property.

2.2 Dynamics along the manifold

How does the absence of symmetry affect the properties of such learnable manifold atractors and what are their functional implications? Will the neuronal system be capable of representing a continuum of states, e.g. of head directions [Taube et al., 1990]? Will such a continuous manifold of persistent neural states be robust to perturbations that may push the neural activity out off the manifold and that are inevitable in biological systems? We find that both on and off manifold properties are affected when symmetry is lacking.

As a result of the difference in the timescale of the dynamics along the on- and off- manifold directions, they can be analyzed separately. We first focus on the on manifold direction. In case that the state is not perfectly persistent due to an external input ($\epsilon > 0$, Figs.S3F-H and Fig.3), an imperfect training (Fig.4), or an alteration of the connectivity (Fig.S1), the trajectories quickly converge to the manifold, and the dynamics along the manifold is governed by the following rule:

$$\frac{d\psi(t)}{dt} \approx \Delta(\psi)\tau_{eff}(\psi)^{-1} \tag{4}$$

which dissects the speed of motion along the manifold into two factors with simple interpretation: τ_{eff} represents the change in the time-constant that governs the dynamics along the manifold, while Δ reflects the inconsistency between the current neural state and a perfectly persistent one. This inconsistency, which we term the tangential error, is quantified via an auxiliary setting, which we refer to as a *recurrent autoencoder* (RAE). In such a setting, illustrated in Figure S3A-B, and explained in details in Methods (Section 5.3), we test if a point on the manifold or in its vicinity is a persistent state of the dynamics.

Steady state points of Eq.(1) obey $\Delta = 0$, and marginal stability along the manifold is determined by the condition $\frac{d\Delta}{d\psi} \equiv \Delta' = 0$ (Fig.S3A). Crucially, the evolution of ψ along the manifold is assumed to be dramatically slower than the convergence toward the manifold (see below for cases where this assumption does not hold).

We next proceed with applying the simple dynamical rule of Eq.(4) to analyze the network response to external stimuli and to study how a continuum of persistent neural states emerges in the trained networks.

2.2.1 Response to external input

Assuming that the learning is successful and a manifold attractor emerges in the recurrent network, the drift along the attractor is negligible (see Section 2.2.2). Upon introduction of external input the continuum of persistent states collapses to a single fixed point attractor (Fig.3,Fig.S3G-H) and the neural state, and hence the representation of the feature ψ, begins to evolve according to Eq.(4).

Figure 3A depicts an example in which after training an heterogeneous ring manifold we update the bump position by introducing a weak stimulus in a direction of $\psi_1 = \pi/2$. The update dynamics are in agreement with Eq.(4) (Fig.3B-compare colored points with the theoretical prediction in dashed lines). Here, the heterogeneity level g only affects dynamics via the effective time constant τ_{eff}, and not via the tangential error, which is given by $\Delta = -\epsilon A^{-1}\sin(\psi - \psi_1)$ (red curve in Fig.3B, see also Methods and [Ben-Yishai et al., 1995]). For a ring manifold we calculated the effective timescale and find that it is given by $\tau_{eff} \approx \tau(1-\beta(g))^{-1}$, with $\beta(g)$ accounting for the effects of synaptic heterogeneity. This factor is connected to the correlation among the individual neuronal gains:

$$\beta = g^2 \left\langle \sin^2(\psi)\phi'(x_i(\psi))\phi'(x_i(0)) \right\rangle_{i,\psi} \tag{5}$$

with $\phi'(x_i(\psi))$ being the gain of the neuron i at location ψ on the manifold and with $\langle . \rangle_{i\psi}$ denoting average over all locations and all neurons (Methods). In networks lacking heterogeneity, the factor β is zero and it is monotonically increasing with the heterogeneity level. We therefore find that the larger the level of synaptic heterogeneity in the trained network, the slower the response to the external input is (Fig.3C-D).

We next sought to investigate how the dynamics along the manifold is affected by its geometry. For an arbitrary manifold, in contrast to a ring manifold, the time required to update the internal representation from one location to another depends on the initial and final locations and not solely on the difference between them. This is depicted in the example of Figures 3E-G. This dependency on the specific location on the manifold can be decomposed to the dependency on the geometry of the manifold and the input, as captured by the tangential error, and on the effective timescale which varies along the manifold (Fig.3F). Figure 3H shows examples of the calculated timescales along various shape of manifolds (see derivation in Methods). Interestingly, we find that the effective timescale tend to be slower at locations along the manifold that are represented by a higher total firing rate (peaks at Fig.3H and see Fig.2A). This provides a quantitative support to the intuition that the more tuned the internal representation is (i.e. the larger the amplitude of the bump is), the harder it is to update its location and, therefore, the slower is its response to external input.

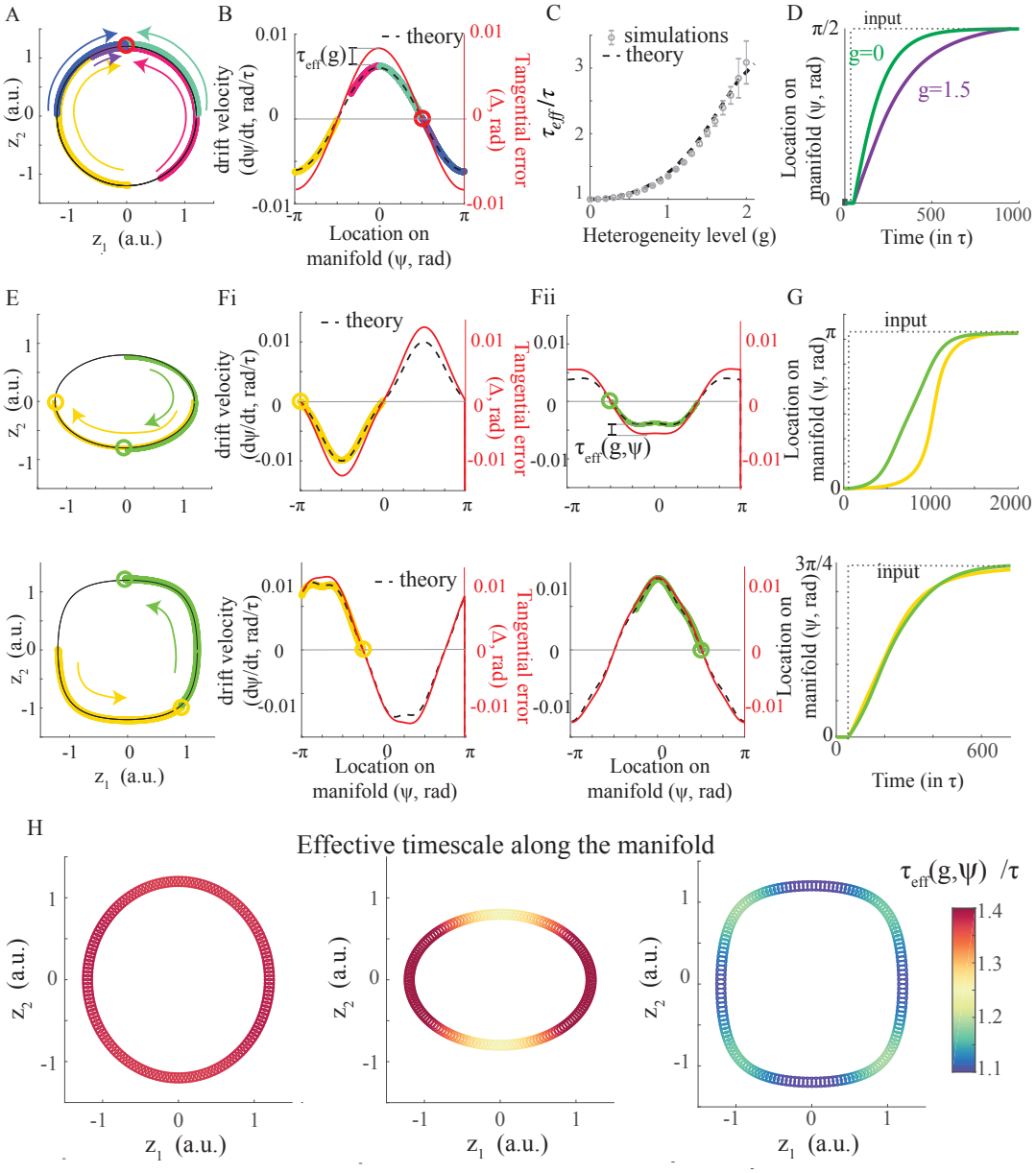

Figure 3: Translation on manifold attractors in response to external inputs A-D. Trained ring manifold. **A.** Projection of the neural activity on the decoder ($\boldsymbol{z}$ plane) for a trained ring manifold with external input at $\psi = \pi/2$ (red circle). All five trajectories (different colors), starting from random locations on the manifold, converge to $\psi = \pi/2$. Trajectories were slightly adjusted to not overlap for illustration purposes. **B.** Drift velocity of the trajectories in (A) (same color code) against location on the manifold. Red circle: angle of the external input. Red curve: Tangential error. Dashed line: Theory (Eqs.(4),(5)). The difference between the tangential error and the drift velocity is a simple scaling by τ_{eff}. **C.** The effective timescale increases with the level of heterogeneity in trained ring manifolds. Dashed line: theory from Eq.(5). Gray: S.E.M for 10 network realizations. **D.** Bump's location vs. time for networks that were trained with different levels of synaptic heterogeneity. The external input is presented at $t = 50\tau$ at $\psi = \pi/2$. The network's internal representation rotates towards $\pi/2$ with a velocity that depends on the level of synaptic heterogeneity. **E-G.** Top: Trained ellipse manifold from Fig.2E-G. Bottom: another example of a trained manifold (see Methods for parameters). **E.** Green trajectory: the bump's location drifts along the manifold towards the input at $\psi = -\pi/2$ (green circle). Yellow trajectory: drift towards the input which is now at $\psi = \pi$ (yellow circle). **F.** Same as (B) but for the two trajectories in (E). Note that both the predicted and the actual drift velocities are distorted with respect to tangential error. This is because the effective timescale now depends on the location along the manifold. **G.** Bump's location vs. time. Same trajectories as in (E). **H.** Effective timescale along the manifold for the three examples in (A-G). Note the increase in timescale along the corners, that correspond to high firing activity.

2.2.2 Build-up of a continuum of persistent neural states

We next ask how a continuous internal representation emerges from sampling a discrete points on the manifold. In our model we train the network by sampling M points of the manifold $\psi_1 ... \psi_M$, which enforces these points to become fixed points of the neural dynamics (Fig.4A, Methods). However, the neural states at unforeseen values of ψ are not persistent and tend to converge to one of the learned fixed points (Fig.4B-C). While up until here we assumed that the number of sampled points is large ($1 \ll M \ll N$), in this section we consider finite number of sampled points. Specifically, we are interested to assess how fast the attractiveness of individual points diminishes with more samples added (Fig.4B-C), prompting emergence of a continuous manifold, and how the heterogeneity level modulates this effect (Fig.4B-C). According to Eq.(4), this is equivalent to analyze how quickly the slope of the tangential error at the sampled points approaches zero. We find that interpolation towards a continuum of the unforeseen values of the feature ψ happens very quickly with the number of samples.

We quantify this by analyzing the trained ring manifold (Fig.1B), which is especially amenable for a full analytical treatment. Here, it can be shown analytically that the rate in which the on-manifold dynamics approaches marginal stability, and hence a continuous representation of the feature, depends on the decay rate of the principal components of the neuronal representation. Specifically, linearizing Eq.(4) around the sampled points yields that the eigenvalue of the linearized dynamics is:

$$\Lambda_\psi \approx \Delta' \tau_{eff}^{-1} \approx -\frac{(M-1)C_{M-1}}{C_1}\tau_{eff}^{-1} \tag{6}$$

where C_k denotes the k'th score of the neural activity correlation matrix, or, equivalently, variance

explained (VE) by the k'th PCs of the neuronal representation (see Eq.(76) in Methods for the exact equation, including small M). We next calculated the decay in VE with the PC number. Figure 4D shows that the VE decays fast with successive PCs, even for high synaptic heterogeneity. Therefore, as the factor $\frac{C_{M-1}}{C_1}$ decays fast with M, exponentially in case of smooth correlation functions (e.g. [Katznelson, 2004]), the derivative of the tangential error gets smaller, and the dynamics along the manifold becomes marginally stable, with no attractive or repelling forces in present ($\Lambda_\psi \approx 0$, Fig.4B-C,E-F). The analytical derivation is in good agreement with the simulations (compare solid lines with circles in Fig 4E-F).

Finally, another indication for convergence to a continuous attractor is a small drift velocity between the sampled points (see Eq.(4)), This is verified numerically in Fig.S5A, showing that the drift velocity is small, and also decays exponentially with the number of points.

While it is not straightforward to generalize Eq.(6) to a manifold with arbitrary geometry, like for example those presented in Fig.3, we find that also in such cases the continuity along the manifold is obtained very fast, exponential with the number of sampled points (Fig.S5B-C). We thus conclude that the rate of approaching continuity of feature representation in the trained networks is extremely fast, even in the presence of heterogeneity in synaptic connectivity and asymmetries in the neural representation.

2.3 Convergence toward the manifold attractor

Neural dynamics must resist perturbations or stimuli that aim to push the neuronal activity away from the manifold attractor. How does heterogeneous synaptic interactions affect the dynamics in the N-1 dimensions that are orthogonal to the 1D manifold direction and in which stimulus or perturbations should be suppressed? Here, we distinguish between the neuronal activity that affects the bump's dynamics, i.e. its location and amplitude, and the vast majority of activity which is orthogonal to decoder plane and hence do not affect the bump's dynamics (Fig5A).

In the case of a heterogeneous ring manifold we find that synaptic heterogeneity stabilizes perturbations in the bump's amplitude. This is exemplified in Figure 5D-E. Here, bump's amplitude is perturbed by a noisy external input. In good agreement with theory, both static perturbations and fluctuations are damped more when levels of synaptic heterogeneity are increased. This effect is determined by the second largest eigenvalue of the linearized dynamics of Eq.(1) (Fig.5C), corresponding to the amplitude direction for which simulations and theory are compared in Figure 5E (green line, see Methods). Qualitatively, this additional stabilization could be attributed to an increase of neuronal activity, causing more neurons to saturate and become less prone to perturbations. Conversely, the internal dynamics of the network, which are not observable via bump's amplitude or angle, become less stable when synaptic heterogeneity grows and eventually become chaotic (purple line in Fig.5E and [Sompolinsky et al., 1988]).

We continue with exploring how the shape of the manifold affects the stability of the bump's magnitude. We demonstrate that the convergent property towards the manifold might be compromised and even ruined (Fig.6A) by geometry changes, and that these dynamical phenomena can be accurately

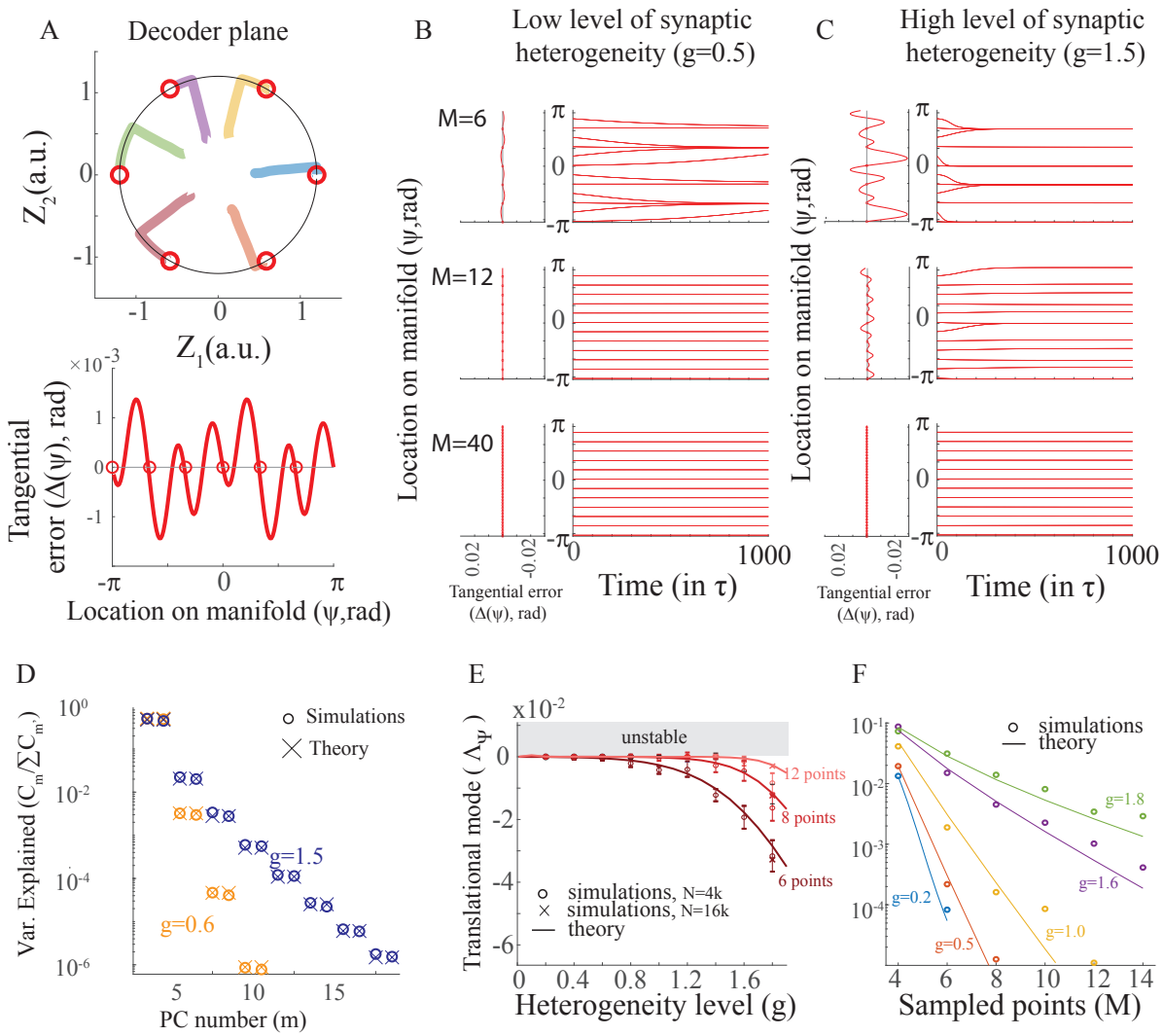

Figure 4: Continuity of manifold quickly emerges in trained heterogeneous ring network. A. Top: Decoder view of a trained ring manifold when training M=6 equally distributed points on the manifold. Circles: trained points. Colored lines: trajectories starting from different initial conditions quickly converge towards the manifold and then drift slowly to one of the trained points. Bottom: The tangential error of the RAE for the trained network. **B.** Left: The tangential error against the the bump location. Right: Bump location when starting from different points on the manifold as function of time. Level of synaptic heterogeneity in the network is g=0.5. Top: training the network by sampling M=6 points of the manifold. Middle: same but with M=12. Bottom M=40. **C.** Same as (B), but with g=1.5. **D.** Variance explained (VE) against the principal components number for low (orange) and high (blue) levels of synaptic heterogeneity. Circles: VE as calculated from a simulated network. Crosses: Theoretical prediction (Eq.(45)). **E.** Stability along the manifold direction against the heterogeneity level for M=6, 8 and 12 sample points. Theory: Eq.(6). Note the scale of the y-axis. **F.** Same as (E) but against the number of sampled points. Note the exponential decay to marginal stability (zero eigenvalue), and hence continuity.

inferred from the static neural representation along the manifold through a dramatic dimensionality reduction (see Section 5.8).

Figure 6 depicts two families of manifold geometry, parametrized by a continuous parameter h so that $h = 0$ corresponds to a perfect ring and positive or negative h implies gradual deformation. When considering manifolds with involved geometries, instabilities in the amplitude direction may appear (Figure 6A). Without equivalence between the points in a general manifold, local stability may vary along the attractor. However, we focus the analysis on points with reflectional symmetry (black circles in Fig.6B,C), as we find that instability tends to show up at these points (Fig.S6, Methods). Indeed, figure 6C shows that the amplitude direction destabilizes for various ranges of manifold deformations. This effect can be non-monotonic in both the complexity of the manifold's shape (in h, Fig.6C) and in the level of synaptic heterogeneity in the network (not shown). In what follows, we will provide evidence for the relation between loss of stability and the geometric complexity.

Crucially, in manifolds in which the amplitude direction approaches instability, such as the manifold depicted in the top panel of Fig.6C, the separation of timescales, which Eq.(4) is build upon, no longer holds. Contrarily to the attractor networks in Fig.3, here a weaker attraction, or equivalently higher susceptibility to input in the amplitude direction, enables input driven trajectories that do not follow the shape of the unperturbed manifold. Hence, upon introduction of input, the representation might jump toward the new state rather than rotate smoothly along the manifold (Fig.S4).

To conclude, the convergent dynamics towards the manifold depends both on the heterogeneity level and the geometry of the manifold. For some cases of complex geometry this convergence is either partially compromised or completely ruined, resulting in jumps instead of smooth transitions along the manifold, or even in its complete destabilization. Further insights to this phenomenon are related to the effective dimensionality of the neural representation and of the resulting dynamics and are described in the following section.

2.4 The manifold geometry controls the complexity of the dynamics

We conclude the paper by analysing how the complexity of the dynamics in the vicinity of the manifold depends on the manifold geometry. We quantify this by asking how many PCs of the neural representation are needed to capture the first few leading modes of local dynamics around the manifold.

While in symmetric-connectome manifold attrators the leading modes of the local dynamics are determined by the rank of the structured matrix, two dimensional in the case of no heterogeneity in our model (Methods, [Ben-Yishai et al., 1995, Hansel and Sompolinsky, 1998]), the situation in networks in which manifold attractors emerge in the presence of heterogeneity is more delicate. Namely, it can formally be only reduced to dimension $2M$ (number of sampled points times the rank of the structured interaction matrix. see Methods and [Rivkind and Barak, 2017]). At first sight, this implies that increasing M results in a more complex dynamics around the manifold, involving more modes of the dynamics that participate

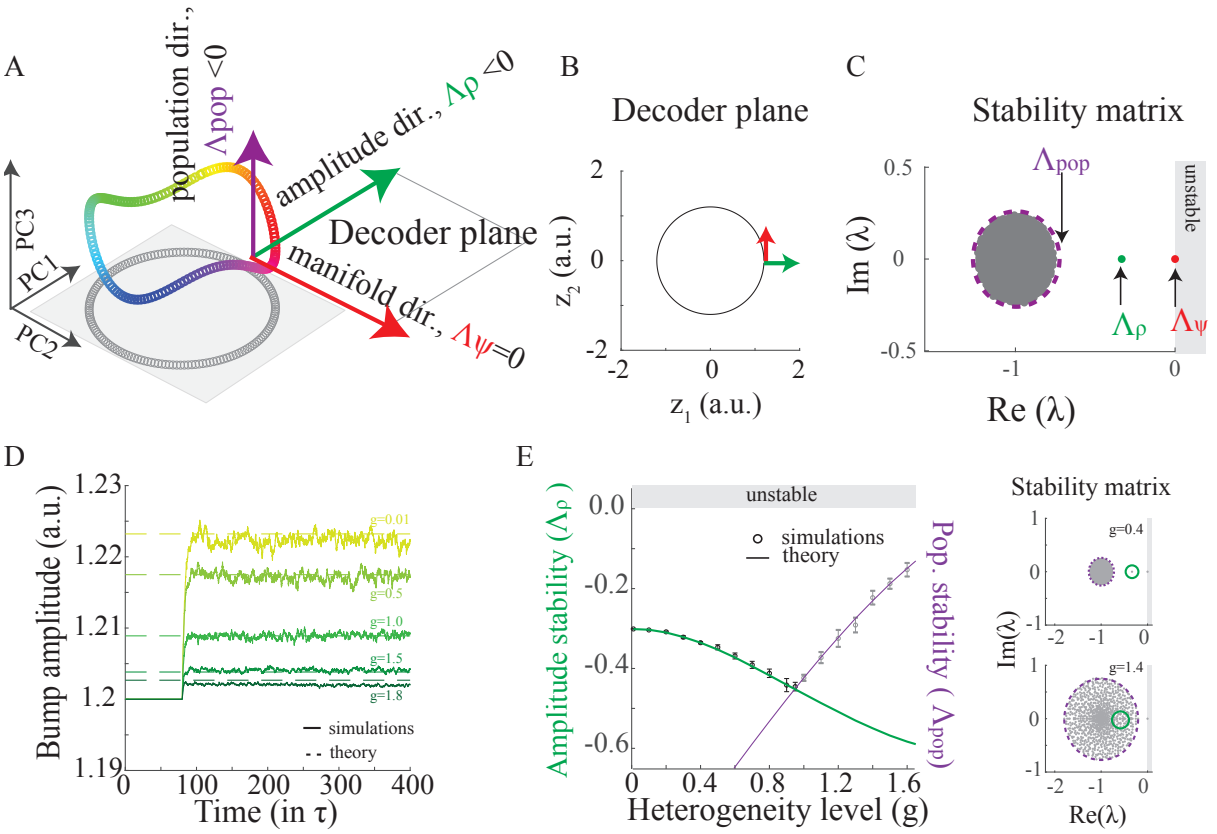

Figure 5: The effect of synaptic heterogeneity on the convergence toward the manifold. A.-C. Illustration of the directions used to determine the stability of the manifold and the expected behavior for each dynamical direction. **A.** Linearizing the dynamics around the manifold results in modes that affect the stability of the manifold (red) and amplitude (green) directions, as well as other modes of the neural dynamics (purple) that are invisible to the decoder plane and, hence, do not affect the bump's dynamics. For a ring manifold the decoder plane is align to the first 2 PCs. **B.** Projection of the manifold on the decoder plane and illustration of the amplitude and manifold directions for which we do the linearization. We expect a zero maximal eigenvalue in the manifold direction ($\Lambda_\psi \approx 0$ for marginal stability, Eq.(6)), and negative maximal eigenvalues in the amplitude (Λ_ρ) and the other population directions (Λ_{pop}). **C.** Spectrum of the linearized dynamics of Eq.(1) (imaginary and real parts of the eigenvalues, λ's, of the stability matrix in Eq.(13)) around the trained points. We denote the maximal eigenvalue at each direction by Λ. Instability of the dynamics happens when one or more eigenvalues cross zero (gray area). **D.-E.** Synaptic heterogeneity stabilizes the amplitude direction. **D.** Bump amplitude against time following a perturbation at time 80τ for different levels of synaptic heterogeneity. The larger g is, the more stable is the amplitude and fluctuations are damped. **E.** Right: Eigenvalues of the network stability matrix in the vicinity of the attractor. Left: Second largest eigenvalue of the stability matrix (obtained from simulations) against the heterogeneity level. Mean+-sd of simulations of 10 network realizations. Green: predicted (largest) maximal eigenvalue in the amplitude direction (Λ_ρ, see Methods 5.7.3). Purple: predicted maximal eigenvalue of the population direction; see Eq.(14) for Λ_{pop}. For a ring manifold the projection on the first two leading PCs coincide with the decoder plane

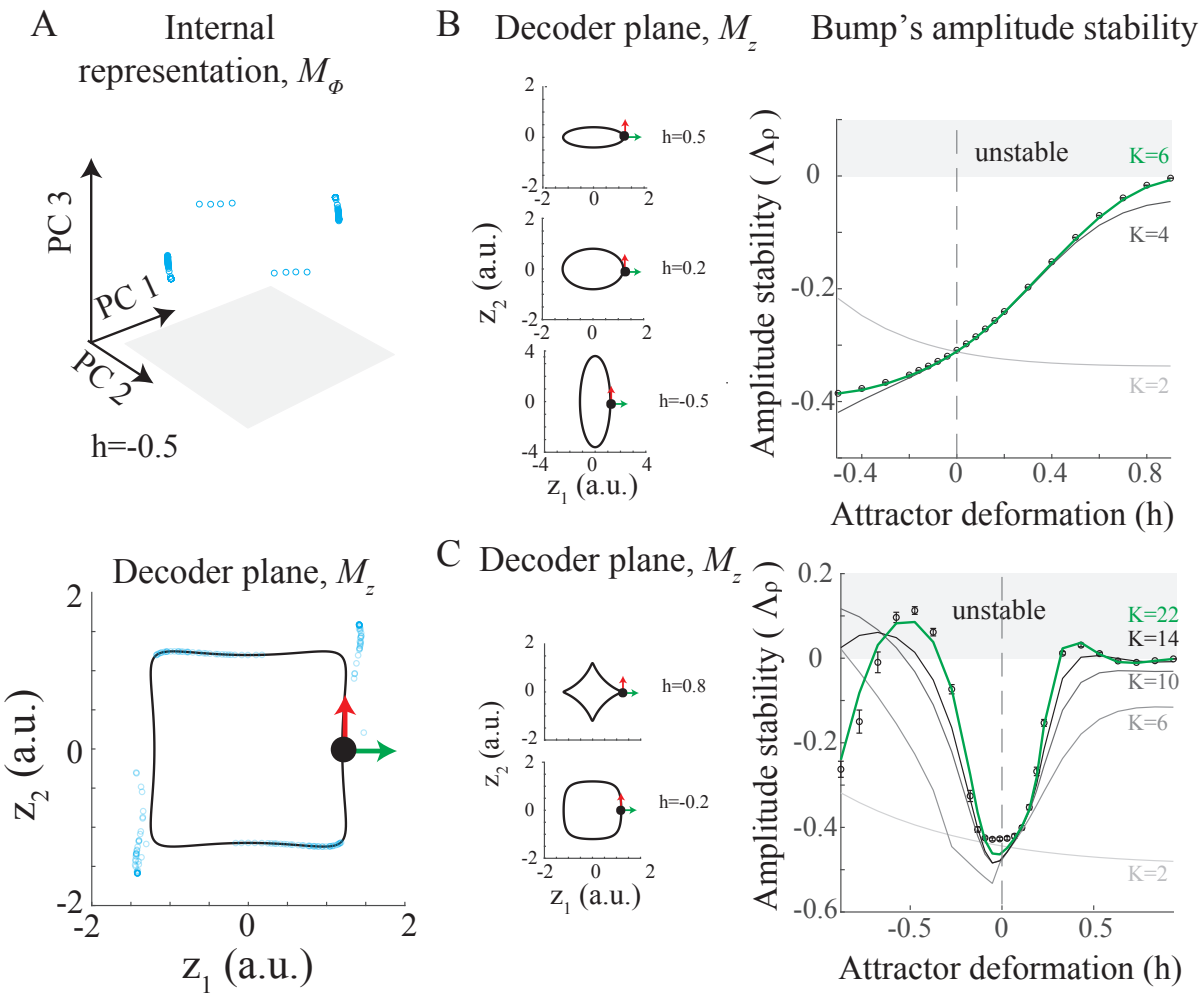

Figure 6: Geometry of manifold attractors controls the convergence in the amplitude direction. **A.** Destruction of a manifold attractor due to instability of the amplitude direction. Top: A snapshot of the neural activity, where all points initialized on the trained manifold, projected on the leading PCs. In steady state all points converge to one of four isolated attractors. Bottom: same as top, but projected on the decoder plane, $\boldsymbol{z}$. Black curve: predefined manifold (see parameters in Methods). **B.** Ellipse manifolds, parametrized as in Eq.(7), with $h = 0$ corresponds to a ring manifold. Left: Predefined trained manifolds. Stability is analyzed at 0 radians (black point). Amplitude direction in green and manifold direction in red. Right: Maximal eigenvalue of the stability matrix along the amplitude direction against the deformation of the manifold, h. Circles: maximal eigenvalue calculated by the spectrum of the full stability matrix, Eq.(13) (see Methods). Solid lines: theoretical prediction through the low-dimensional mean field stability matrix (Eq.(93)) with a cutoff at K=2,4 and 6 PCs. **C**. Same as (B), but for a different manifold (see Methods for parameters). Note that some manifold geometries are unstable due to instabilities in the off-manifold direction, as shown in (A) for h=-0.5.

in the dynamics of the bump. However, further simplification is possible and it can be shown that the leading modes of the local dynamics can be captured by only $K \ll M \ll N$ leading PCs (Methods), allowing for a dramatic dimensionality reduction of the dynamics from the full N dimensional system to a low, K-dimensional, local dynamics.

Without a clear symmetry in the second order statistics, as is the case for the trained heterogeneous ring manifold (Fig.2), carrying out a complete mean-field analysis for an arbitrary manifold, while technically doable, turns out to be cumbersome and not necessarily instructive. Instead, we predict the leading modes of the dynamics by applying the self-consistency mean-field equations using the empirically obtained K leading PCs of the static neural representation (Methods).

We find that for the special case of a ring geometry, taking $K = 2$ PCs is enough to predict the leading modes of the dynamics (Figs.6B,C, Fig.7A). It is only when considering more involved geometry of the manifold that the number of required PCs is larger than 2. Indeed, in Fig.6 we show that when $h \neq 0$, more than 2 PCs of the manifold are needed to predict the leading dynamical mode of the bump's amplitude. However, as long as the deformation of the manifold is small, the dynamics alters continuously and moderately with h. In particular, in the cases presented in Fig.6C and Fig.7B-C only 4 to 6 PCs suffice to explain the local dynamic in the vicinity of the manifold, and it is only for more involved geometries that we find that more PCs are required to predict the modes of the dynamics (Fig.6C and Fig.7D).

Interestingly, destabilization was typically associated with large K needed to predict instability (Fig.6C). This offers a clue that destabilization mechanisms and complexity of geometry, as indicated by large K, are interrelated.

To conclude, we find that it is possible to learn manifold attractors with neither symmetry in synaptic connectivity, nor symmetry in the shape of the manifold. The dimensionality of the dynamics in the vicinity of the attractor is determined by the complexity of the attractor geometry, with the special case of ring geometry being especially amenable to analytical formulation and analysis, but otherwise not exceptional.

3 Discussion

Hypothesis of computations by manifold attractors relays on idealized symmetry assumption in synaptic connectivity, which constitute the theoretical backbone for the emergence of such attractors in models and specifically in non-linear recurrent networks. However, this assumption is inconsistent with heterogeneous synaptic interactions and single neuron properties, as well as with diverse neural responses that are widely observed in various brain areas. Furthermore, as a result of the symmetry in the connectivity, such symmetric-connectome networks can only support limited repertoire of internal representations, as it requires symmetric geometry thereof. Thus, the validity of the symmetry assumption in real biological

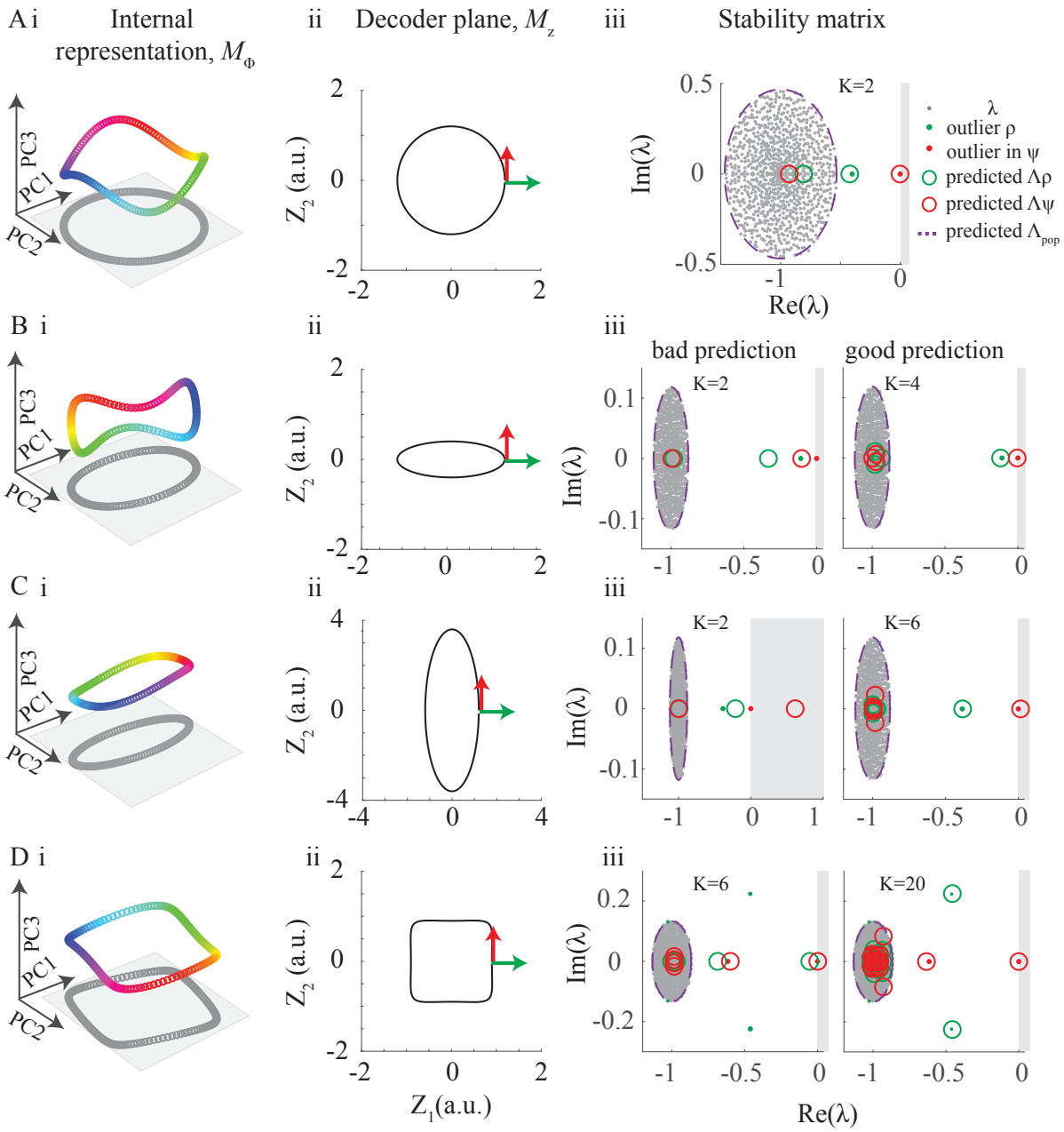

Figure 7: Geometry of manifold attractors controls the complexity of the dynamics in its vicinity. **Ai.** Projection of a trained ring manifold onto the leading PCs with color coded for the continuous feature. Gray: projection onto the 2 leading PCs. **Aii.** Decoder view of a predefined trained ring manifold. **Aiii.** Spectrum of linearized dynamics of the manifold in (ii). Purple: predicted maximal eigenvalue of the population direction; see Eq.(14) for Λ_{pop}. Red and green circles are the predicted eigenvalues according to the low-dimensional mean field stability matrix (Eq.(93)). Here $K = 2$ principal components suffice to get good mean field prediction for the outliers. Grey region corresponds to instability ($Re\lambda > 0$) **B.** Same as (A), but for an ellipse manifold. Here, prediction of the outliers is improved by increasing the number of principal components used for estimating the low-dimensional mean field stability matrix **C-D.** Same as (B) but for different manifolds

systems is highly speculative, raising the question of the applicability of this hypothesis to real biological systems.

We argue that symmetry is not required to have an approximately continuous manifold in recurrent networks. Instead, training networks results in manifolds of states that can be considered persistent for any practical purpose, without symmetry in the recurrent interactions or in the neuronal representation. Specifically, exponential decay of attracting forces along the manifold is predicted analytically for tractable case and verified numerically to hold for more involved settings. Relatively small number, order of ten, of learning samples, nulls any motion on the manifold.

Beside ability to maintain persistent state, functional manifold attractor must be responsive to external input in a tractable way. We found that the response to external input is determined primarily by attractors geometry, and more specifically by its projection into decoder plane. Namely, to make a coarse prediction of how the memorized feature will evolve upon introduction of input, one does not need to know about the internal connectivity of the neural network. For more accurate estimation of the input driven dynamics another factor is needed, which is an effective timescale that modulates the motion along the manifold and is expected to affect the ability of the network to perform angular integration (see below). This factor does depend on internal connectivity and can be devised self-consistently in our model, leading to a fine and accurate prediction of dynamics (see Fig. 3).

Finally, in contrast to the bump's location, which must be responsive to stimulus, its amplitude is expected to remain approximately indifferent to the external input and perturbations [Durstewitz et al., 2000]. To that extent, we found that amplitude stability is preserved even when symmetry of neuronal representation is compromised (Fig. 6). However, we do find that destabilization of the bump's amplitude is a limiting factor in the emergence of manifold attractors. When geometry of neuronal representation becomes "too complex" the bump's amplitude may become unstable. Such a complexity is associated with increasing number of dimensions of the neural representations that is required to predict dynamics in its vicinity (Fig. 6). This result suggests that the geometry of putative manifolds in the brain can be more involved than a ring, but not too involved for maintaining a stable representation.

3.1 Heterogeneity in networks that memorize and path integrate

Stored spatial information and path integration have been hypothesised to rely on the concept of computations by manifold attractors. Neurons in brain areas that support these computations are known to be highly heterogeneous [Funahashi et al., 1993, Romo et al., 1999, Finkelstein et al., 2015, Fisher et al., 2019, Chaudhuri et al., 2019]. Neuronal activity in symmetric-connectome networks, however, are inconsistent with these studies as neurons in these models do not exhibit any heterogeneity. In fact, adding heterogeneity to these networks results in a systematic drift of the memory and to a failure to integrate and accurately respond to external cues (Fig.S1).

A few studies explored the ability of manifold attractors to cope with diversity in neuronal tuning

and heterogeneity in synaptic connectivity. Two of these studies explored a way to reduce the drift of the internal representation by adding a slow component to the dynamics, such as short term facilitation [Itskov et al., 2011, Hansel and Mato, 2013]. In trained networks, such dynamical mechanisms may account for rapid alterations to connectivity for which training proves too slow. This possibility is left for future work. In Renart et al.[Renart et al., 2003] the authors showed that homeostatic mechanisms can homogenize the network despite considerable heterogeneity in single cell and synaptic properties. Our results indicates that even with heterogeneity overcome, its dynamical consequences persist: Converging forces towards the manifold in the amplitude direction increases with the level of heterogeneity (e.g. Fig.5) and the input response exhibits slowdown (Fig.3). Furthermore, the notion of homogenization becomes irrelevant once the requirement for symmetric geometry is relaxed, and neural states become manifestly diverse (Fig.2,7).

Instead of relying on rotational symmetry to construct a continuous attractor, a different approach was taken recently by [Mastrogiuseppe and Ostojic, 2018]. The recurrent connectivity in this model is based on a Hopfield network [Hopfield, 1982] and symmetry in strength of two attractors translates into continuity. However, this approach leads to a limited repertoire of tuning curves, differing only by their amplitude, and does not cope with asymmetry in synaptic connections. Furthermore, adding randomness to the recurrent connectivity in this model results in shattering of the continuous attractor [Mastrogiuseppe and Ostojic, 2018]. This shattering can be mitigated by training, and is tractable by our analysis (see Methods).

In application to systems that track external inputs and integrate angular velocity, such as the head direction system [Hulse and Jayaraman, 2020], our analysis can be used to obtain the maximal tracking velocity. As the response to external input slows down in the presence of synaptic heterogeneity, we expect the maximal tracking velocity to decrease with the level of synaptic heterogeneity in the network. As heterogeneity is correlated to diversity in tuning curves in our model, our work suggests that integration is impaired in networks where neuronal responses are too diverse.

In the fly's head direction system the network supporting the integration of idiothetic and allothhetic signals is assumed to be on the order of 10-50 neurons [Hulse et al., 2020]. In this and other [Simony et al., 2008] small networks almost any deviation from a symmetry assumption, such as heterogeneity in tuning curves and in the connectome, is catastrophic for the computational capabilities of the network. While here we applied a recurrent autoencoder paradigm (Fig. S3 and Methods) to analyze manifold attractors in large networks, in which mean-field estimates are obtainable for Eq.(4), the paradigm itself is valid for networks as small as a few neurons.

Finally, noisy dynamics such as those expected in spiking networks or in chaotic rate networks, will result in a diffusion of the internal representation along the manifold. Such noise accumulation is known to be a limiting factor in working memory and in integration systems [Burak and Fiete, 2012], and is attributed to the marginal stability of the manifold direction. Our analysis implies that not all the marginally stable manifolds were born equal: we found that both synaptic connectivity and the geometry

of neuronal representaton affect the drift along the manifold. It will thus be interesting to explore how these effects generalize to diffusive dynamics.

3.2 Beyond symmetrical geometry in manifold attractors

Symmetric-connectome models can support only the representation of manifolds with symmetry in their geometry. This is exemplified in Fig.1A, where a symmetric-connectome ring attractor model is depicted. The tuning curves are identical and the projections on the two leading principal components (PCs) is circular. However, it is unclear that this is the case for putative manifolds in the brain. For example, in the head direction system [Rubin et al., 2019, Chaudhuri et al., 2019] leading PCs do not feature such a perfect circular shape. Same holds for the recently discovered manifolds of torus topology in the enthorinhal cortex [Gardner et al., 2021]. Notably, the geometry of the manifold can be deformed due to external inputs. Indeed, recent studies suggest that the manifold's geometry depends on external signals like the richness of the environments or locations of learned rewards [Boccara et al., 2019, Low et al., 2020].

Our work provides a link between the geometry of the manifold and its dynamical properties. While estimating the geometry of the manifold can be challenging, for example due to small number of recorded neurons or sampling biases [Rubin et al., 2019, Chaudhuri et al., 2019], our work suggests that estimating the manifold's geometry, and not only its topology, is essential for predicting the computational properties of the network. In particular, our work suggests that the dynamics in the vicinity of locations which are represented by a higher total firing rates (Fig.2) are less responsive for updates (Fig.3) and more prune to perturbations (Fig.S 6), with potential implications for computations such as tracking and integration.

Finally, recent computational works showed that a manifold attractor emerges in networks with heterogeneous connectivity when training a recurrent network to integrate angular velocity [Cueva et al., 2019, Sorscher et al., 2020]. Interestingly, the authors in [Sorscher et al., 2020] found a distorted 2D manifold structure when they trained networks to path integrate. Our work provides a theoretical understanding for the connection between the neuronal representation, such as diversity in neuronal responses and the geometry of the manifold, and the emergence of a marginal direction and the dynamics in the vicinity of the attractor in such trained networks.

3.3 Analytical theory of trained neural networks

While continuous attactors were observed numerically [Seung, 1998, Seung et al., 2000, Mante et al., 2013, Sorscher et al., 2020, Cueva et al., 2019, Maheswaranathan et al., 2019], it is not clear how they emerge from finite number of training examples. Here we established a link between interpolation capabilities and the spectrum of neuronal activity. Specifically, Eq.(6) connects the decay rate of the PCs of

neuronal tuning curves to the rate of approaching a marginal stability. From a signal processing perspective (e.g. [Shannon, 1949]), this result can be interpreted as quantifying frequency *aliasing*. The ideal decoder samples the leading Fourier mode (i.e $\sin\psi$, $\cos\psi$). If a finite number of samples, M, used for learning, then the $M-1$-th spectral component is not orthogonal to the leading mode and it folds on the desired decoder (Eq. (6)). Interestingly, extrapolation, which is known to be remarkably harder than interpolation, can be also analysed from this viewpoint: when some restricted interval $[\psi_1, \psi_2]$ is used for learning instead of the entire domain $[0, 2\pi)$, the desired spatial Fourier mode loses its orthogonality to virtually all other modes and not just to the $M-1$-st one, resulting, in poor extrapolation.

Training large neural networks to have a handful of discrete fixed point attractors is known to results in low dimensional dynamics [Rivkind and Barak, 2017]. On the other hand, it was not clear how this result generalizes to continuous attractors, and specifically whether the dimension of dynamics becomes infinitely large at the limit of a large number of training points. Here, we found that the dynamics in the vicinity of continuous neural manifolds is approximable by a small number of dynamical modes, much smaller than the number of training points and is related to the leading principal components of static neural representation along the manifold. Interestingly, we observed numerically that destabilisation of manifold attractor is attributed to growing number of PCs needed to explain the dynamics. Future work may focus on solidifying this relation analytically.

Examining the trained recurrent manifold attractor network from an autoencoder perspective gives an interesting observation on the connection between manifold attractor networks and kernel approach in machine learning. The autoencoder in our model can be seen as mapping the low dimensional manifold, $\boldsymbol{z} \in \mathcal{M}_z$, to high dimensional feature space $\boldsymbol{\Phi}$ and then mapping it back to $\hat{\boldsymbol{z}}$ using linear regression. As such, the training of a network with an intermediate layer of infinite size is reminiscent of the kernel-based linear regression method. The kernel, however, in the trained manifold attractor case is not one of the classical kernels used in machine learning, like the Gaussian or exponential kernels. It is given by the correlation (kernel) matrix, and thus determined by the mapping to the feature space, as well as by the randomness in the recurrent connections. Interestingly, a recent study made a connection between how fast different modes are learned and the complexity of the neural code [Bordelon et al., 2020]. Similar methods might apply to our framework and, to this extent, generalizing our theoretical result on the fast convergence rate toward marginal stability along the manifold to more general topology could prove especially insightful.

Following these lines, it is also appealing to attribute a regularizing role for synaptic heterogeneity in the learning process. As the number of neurons are larger than the sample points, $M \ll N$, many solutions are possible and we choose the least square solution. Increasing the heterogeneity level thus stabilizes the solution, in a way that is reminiscent of the ridge parameter in ridge regression. Similarly, another intriguing comparison with the machine learning literature is with denoising and contractive autoencoders, methods that are used to capture local manifold structure of data [Alain and Bengio, 2014]. While synaptic heterogeneity in our setting is a recurrent and correlative noise, it has similar contractive flavor as those in the denoising and contractive autoencoders, with a higher level of synaptic heterogene-

ity resulting in an increase in contraction towards the sampled points on the manifold (Eq.(6)), and of contraction in the direction orthogonal to the manifold (Fig.(5)).

In summary, our work shows that continuous attractors can cope with a large level of synaptic heterogeneity and asymmetries in the geometry of the attractors, allowing to construct mechanistic models of manifold attractors in the brain and predict the dynamics from their internal representation.

4 Acknowledgements

We would like to thank Larry Abbott, Ehud Ahissar, Arseny Finkelstein, James Fitzgerald, David Hansel, Ann Hermundstad, Sandro Romani and Inbar Saraf-Sinik for their valuable feedback. AR is hosted by Ehud Ahissar for post doctoral training in Weizmann Institute of Science.

References

[Ahmadian et al., 2015] Ahmadian, Y., Fumarola, F., and Miller, K. D. (2015). Properties of networks with partially structured and partially random connectivity. *Physical Review E*, 91(1):012820.

[Alain and Bengio, 2014] Alain, G. and Bengio, Y. (2014). What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593.

[Amari, 1977] Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, 27(2):77–87.

[Amit, 1992] Amit, D. J. (1992). *Modeling brain function: The world of attractor neural networks*. Cambridge university press.

[Barak et al., 2013] Barak, O., Sussillo, D., Romo, R., Tsodyks, M., and Abbott, L. (2013). From fixed points to chaos: three models of delayed discrimination. *Progress in neurobiology*, 103:214–222.

[Beiran et al., 2020] Beiran, M., Dubreuil, A., Valente, A., Mastrogiuseppe, F., and Ostojic, S. (2020). Shaping dynamics with multiple populations in low-rank recurrent networks. *arXiv preprint arXiv:2007.02062*.

[Ben-Yishai et al., 1995] Ben-Yishai, R., Bar-Or, R. L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848.

[Boccara et al., 2019] Boccara, C. N., Nardin, M., Stella, F., O'Neill, J., and Csicsvari, J. (2019). The entorhinal cognitive map is attracted to goals. *Science*, 363(6434):1443–1447.

[Bordelon et al., 2020] Bordelon, B., Canatar, A., and Pehlevan, C. (2020). Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR.

[Braitenberg and Schüz, 2013] Braitenberg, V. and Schüz, A. (2013). *Anatomy of the cortex: statistics and geometry*, volume 18. Springer Science & Business Media.

[Brody et al., 2003] Brody, C. D., Romo, R., and Kepecs, A. (2003). Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Current opinion in neurobiology*, 13(2):204–211.

[Burak and Fiete, 2009] Burak, Y. and Fiete, I. R. (2009). Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput Biol*, 5(2):e1000291.

[Burak and Fiete, 2012] Burak, Y. and Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*, 109(43):17645–17650.

[Chaudhuri et al., 2019] Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., and Fiete, I. (2019). The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature neuroscience*, 22(9):1512–1520.

[Christophel et al., 2017] Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., and Haynes, J.-D. (2017). The distributed nature of working memory. *Trends in cognitive sciences*, 21(2):111–124.

[Compte et al., 2000] Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex*, 10(9):910–923.

[Cueva et al., 2019] Cueva, C. J., Wang, P. Y., Chin, M., and Wei, X.-X. (2019). Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks. *arXiv preprint arXiv:1912.10189*.

[Durstewitz et al., 2000] Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature neuroscience*, 3(11):1184–1191.

[Finkelstein et al., 2015] Finkelstein, A., Derdikman, D., Rubin, A., Foerster, J. N., Las, L., and Ulanovsky, N. (2015). Three-dimensional head-direction coding in the bat brain. *Nature*, 517(7533):159–164.

[Fisher et al., 2019] Fisher, Y. E., Lu, J., D'Alessandro, I., and Wilson, R. I. (2019). Sensorimotor experience remaps visual input to a heading-direction network. *Nature*, 576(7785):121–125.

[Funahashi et al., 1993] Funahashi, S., Chafee, M. V., and Goldman-Rakic, P. S. (1993). Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature*, 365(6448):753–756.

[Gardner et al., 2021] Gardner, R. J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N. A., Dunn, B. J., Moser, M.-B., and Moser, E. I. (2021). Toroidal topology of population activity in grid cells. *bioRxiv*.

[Hafting et al., 2005] Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806.

[Hansel and Mato, 2013] Hansel, D. and Mato, G. (2013). Short-term plasticity explains irregular persistent activity in working memory tasks. *Journal of Neuroscience*, 33(1):133–149.

[Hansel and Sompolinsky, 1998] Hansel, D. and Sompolinsky, H. (1998). 13 modeling feature selectivity in local cortical circuits.

[Hebb, 1949] Hebb, D. (1949). The organization of behavior; a neuropsychological theory.

[Hopfield, 1982] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

[Hulse et al., 2020] Hulse, B. K., Haberkern, H., Franconville, R., Turner-Evans, D. B., Takemura, S., Wolff, T., Noorman, M., Dreher, M., Dan, C., Parekh, R., et al. (2020). A connectome of the drosophila central complex reveals network motifs suitable for flexible navigation and context-dependent action selection. *bioRxiv*.

[Hulse and Jayaraman, 2020] Hulse, B. K. and Jayaraman, V. (2020). Mechanisms underlying the neural computation of head direction. *Annual review of neuroscience*, 43:31–54.

[Itskov et al., 2011] Itskov, V., Hansel, D., and Tsodyks, M. (2011). Short-term facilitation may stabilize parametric working memory trace. *Frontiers in computational neuroscience*, 5:40.

[Jaeger, 2001] Jaeger, H. (2001). The echo state approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148:34.

[Katznelson, 2004] Katznelson, Y. (2004). *An introduction to harmonic analysis*. Cambridge University Press.

[Kropff et al., 2015] Kropff, E., Carmichael, J. E., Moser, M.-B., and Moser, E. I. (2015). Speed cells in the medial entorhinal cortex. *Nature*, 523(7561):419–424.

[Low et al., 2020] Low, I. I., Williams, A. H., Campbell, M. G., Linderman, S. W., and Giocomo, L. M. (2020). Dynamic and reversible remapping of network representations in an unchanging environment. *bioRxiv*.

[Machens and Brody, 2008] Machens, C. K. and Brody, C. D. (2008). Design of continuous attractor networks with monotonic tuning using a symmetry principle. *Neural computation*, 20(2):452–485.

[Maheswaranathan et al., 2019] Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S., and Sussillo, D. (2019). Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Advances in neural information processing systems*, 32:15696.

[Mante et al., 2013] Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84.

[Mastrogiuseppe and Ostojic, 2018] Mastrogiuseppe, F. and Ostojic, S. (2018). Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623.

[McNaughton et al., 2006] McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., and Moser, M.-B. (2006). Path integration and the neural basis of the'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678.

[Nyquist, 1932] Nyquist, H. (1932). Regeneration theory. *Bell System Technical Journal*, 11:126–147.

[O'Keefe and Dostrovsky, 1971] O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*.

[Rajan et al., 2010] Rajan, K., Abbott, L., and Sompolinsky, H. (2010). Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, 82(1):011903.

[Renart et al., 2003] Renart, A., Song, P., and Wang, X.-J. (2003). Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, 38(3):473–485.

[Rivkind and Barak, 2017] Rivkind, A. and Barak, O. (2017). Local dynamics in trained recurrent neural networks. *Physical review letters*, 118(25):258101.

[Romo et al., 1999] Romo, R., Brody, C. D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470–473.

[Rubin et al., 2019] Rubin, A., Sheintuch, L., Brande-Eilat, N., Pinchasof, O., Rechavi, Y., Geva, N., and Ziv, Y. (2019). Revealing neural correlates of behavior without behavioral measurements. *Nature communications*, 10(1):1–14.

[Seelig and Jayaraman, 2015] Seelig, J. D. and Jayaraman, V. (2015). Neural dynamics for landmark orientation and angular path integration. *Nature*, 521(7551):186–191.

[Seung, 1996] Seung, H. S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344.

[Seung, 1998] Seung, H. S. (1998). Learning continuous attractors in recurrent networks. In *Advances in neural information processing systems*, pages 654–660. Citeseer.

[Seung et al., 2000] Seung, H. S., Lee, D. D., Reis, B. Y., and Tank, D. W. (2000). Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron*, 26(1):259–271.

[Shannon, 1949] Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

[Simony et al., 2008] Simony, E., Saraf-Sinik, I., Golomb, D., and Ahissar, E. (2008). Sensation-targeted motor control: every spike counts? focus on:"whisker movements evoked by stimulation of single motor neurons in the facial nucleus of the rat". *Journal of neurophysiology*, 99(6):2757–2759.

[Sompolinsky et al., 1988] Sompolinsky, H., Crisanti, A., and Sommers, H.-J. (1988). Chaos in random neural networks. *Physical review letters*, 61(3):259.

[Sorscher et al., 2020] Sorscher, B., Mel, G. C., Ocko, S. A., Giocomo, L., and Ganguli, S. (2020). A unified theory for the computational and mechanistic origins of grid cells. *bioRxiv*.

[Sussillo and Abbott, 2009] Sussillo, D. and Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557.

[Taube et al., 1990] Taube, J. S., Muller, R. U., and Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435.

[Tsodyks and Sejnowski, 1995] Tsodyks, M. V. and Sejnowski, T. (1995). Rapid state switching in balanced cortical network models. *Network: Computation in Neural Systems*, 6(2):111–124.

[Wang, 2001] Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463.

[Wimmer et al., 2014] Wimmer, K., Nykamp, D. Q., Constantinidis, C., and Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, 17(3):431–439.

[Zhang, 1996] Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126.

5 Methods

Simulations All simulations were done using an Euler method with $dt = 0.1\tau$ and $\tau = 1$. In some cases we took $dt = 1\tau$ and checked that similar results hold for smaller dt. To train the network we simulated Eq.(1) for 1000τ and derived the least square solution. To simplify the analytical calculations we chose the following sigmoidal transfer function $\phi(x) = erf(x)$. Due to the symmetry of the transfer function we then simulated only $M/2$ points of the manifold: ψ_m with $m = 1/M...\pi/M$.

Manifold parameterization For the predefined manifolds in the decoder plane we chose the following parametrized closed curve:

$$\boldsymbol{f}(\gamma) = \frac{A}{1 + h/(a-1)}[\cos(\gamma + \frac{h}{a-1}\cos((a-1)\gamma), \sin(\gamma) - \frac{h}{a-1}\sin((a-1)\gamma)] \tag{7}$$

where the normalization $1/a + h$ guarantees that the amplitude is A at 0 radians.

External input In case of non-zero external input, it is given according to Section 5.1, with $\tilde{\boldsymbol{u}}(\gamma) = f(\gamma)/\mathrm{A}$.

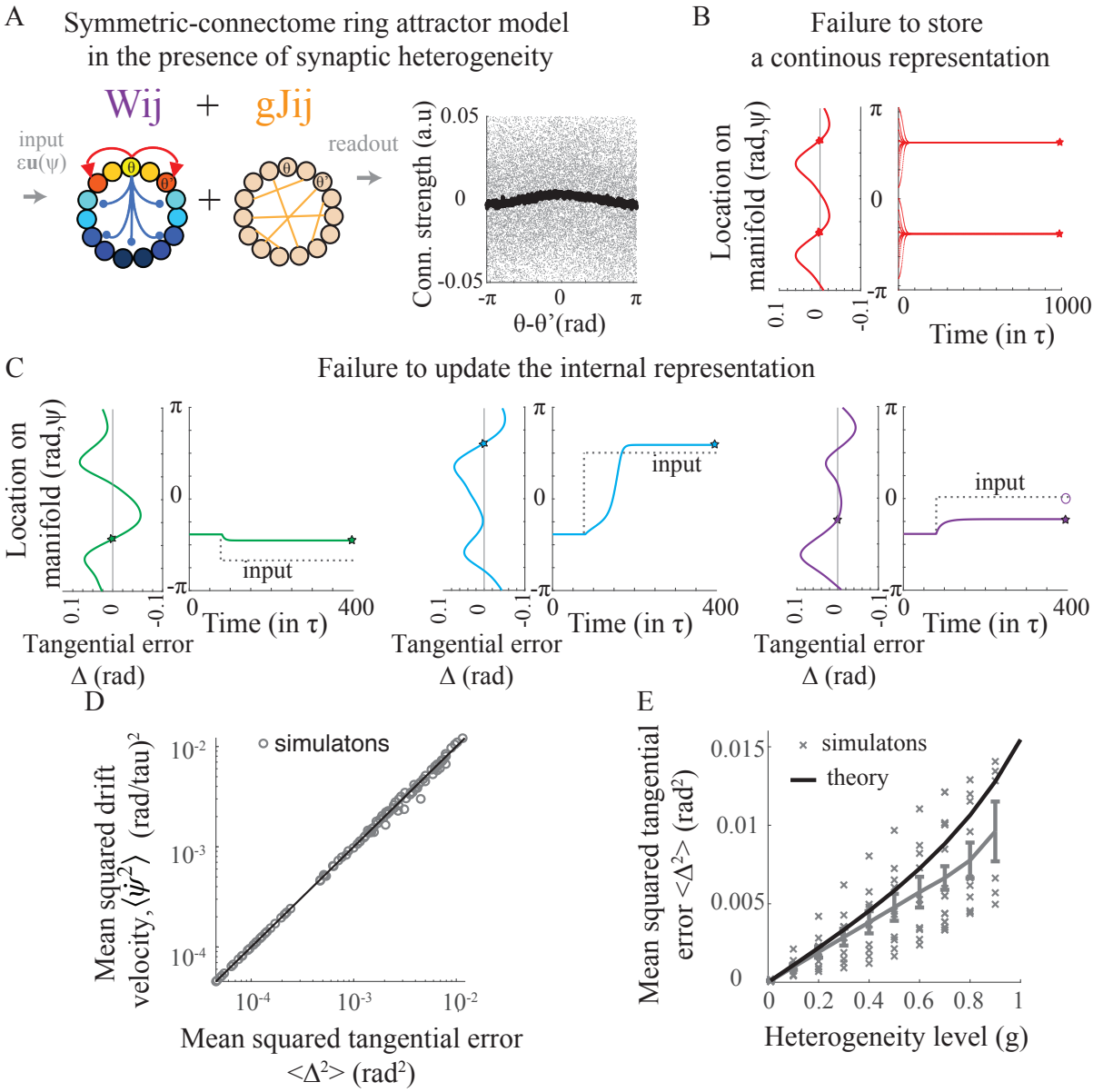

Figure S 1: Failure modes of symmetric-connectome ring attractor network in the presence of synaptic heterogeneity. **A.** Right: Cartoon of symmetric-connectome ring model in the presence of synaptic heterogeneity (g=1). The structured part is as in 1Ai and is unlearned. Left: Connectivity strength vs. distance in POs. **B.** Tangential error vs bump's location (left) and the bumps location vs. time (right) for the network in (A). Due to the heterogeneity the bump drifts toward one of the two stable fixed points in a few time steps. $N = 1000, g = 1, \epsilon = 0$. **C.** Response of the network in (A) to external input. Same as (B), but with $\epsilon = 0.04$).**D.** Mean squared drift velocity, averaged over initial conditions of 160 uniformly sampled points from the manifold, vs. Mean squared tangential error (see Eq.(4)). Each of 100 points corresponds to a combination of hyperparameters g, A times five random seeds (Methods). **E.** Mean squared tangential error vs. heterogeneity level. Theory: Eq. (107), developed for small g, fits well the simulations for small g and start to deviate from the simulations for $g \approx 0.4$; $N = 4000$.

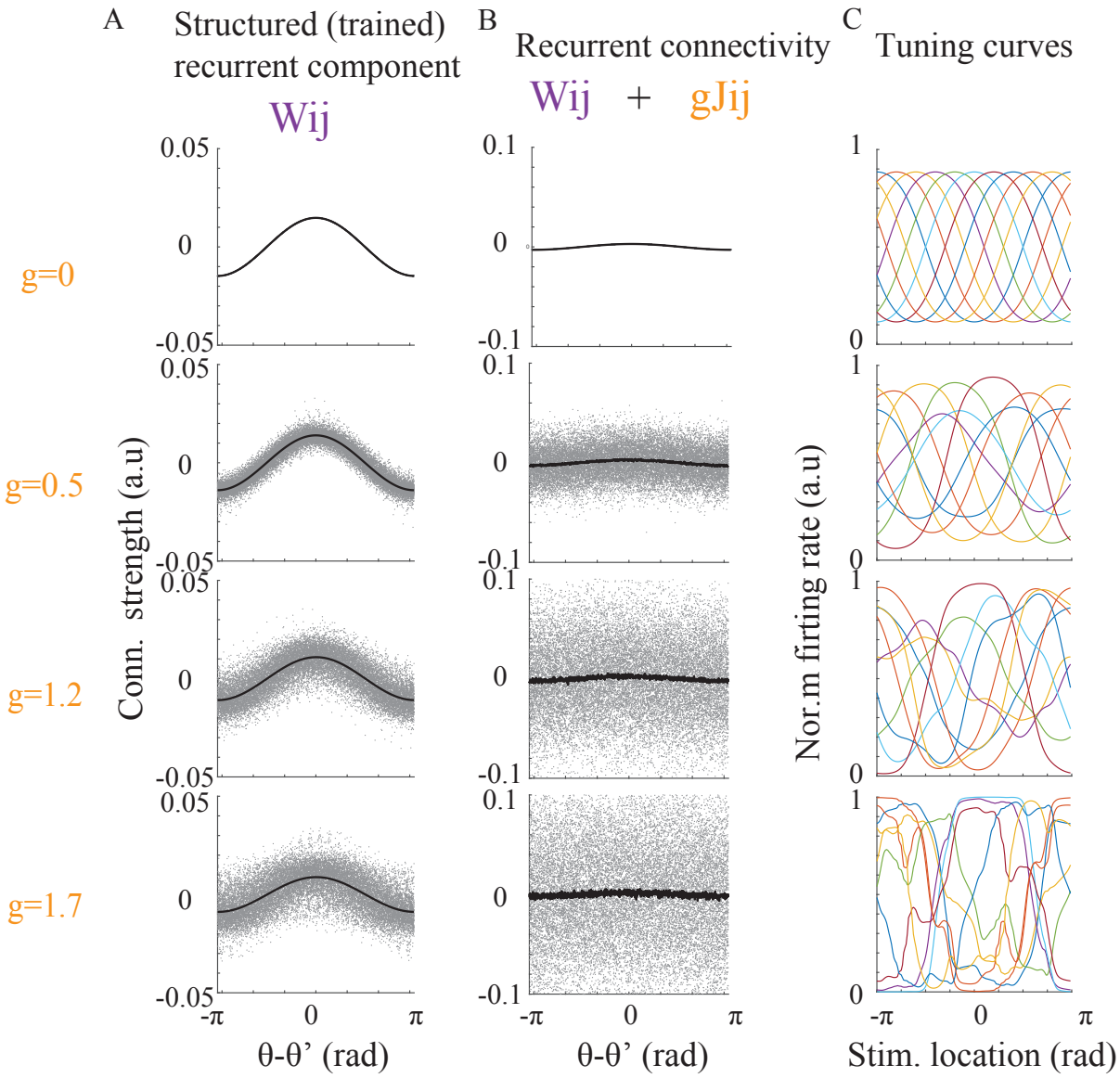

Figure S 2: Recurrent connectivity and tuning curves in trained ring attractor networks . **A.** The structured component of the trained network. **B.** The full recurrent connectivity (not the scale difference with (A).Black: average over all neurons with a bin $(\theta, \theta + \delta\theta)$ **C.** Tuning curves vs. stimulus location.

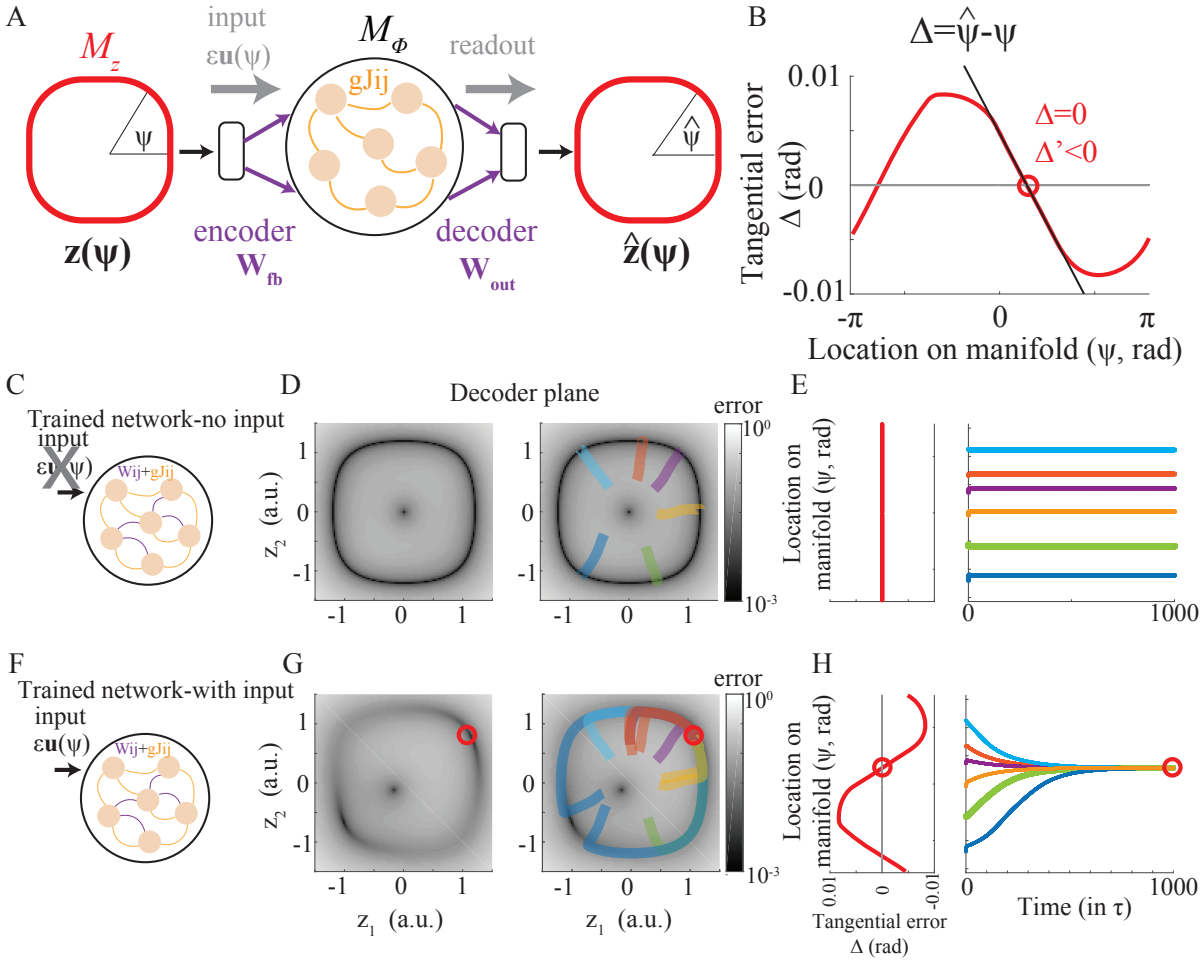

Figure S 3: A Recurrent Auto-Encoder framework (RAE) for analysing manifold attractors. A. Cartoon of the RAE. The structured recurrent loop (purple lines in Fig1C) is opened and decomposed into an encoder and decoder. RAE is driven by fixed stimulus on the 2D plane $\boldsymbol{z}$ and the corresponding steady state output $\hat{z}$ is decoded. **B.** The tangential error, between encoded and decoded angles $\Delta = \hat{\psi} - \psi$. Fixed point of the dynamics are points in which $\Delta = 0$ and stable fixed points (red circles) with $\Delta' < 0$ (Eq.(4)). **C.-H.** Implications of decoding error on network dynamics. **C.-E.** The trained network in the absence of inputs.**D.** The reconstruction error of the RAE. Left: error is shown in 2D z plane. Right: error is superimposed with with 6 randomly initiated trajectories of the full system. Dynamics converge rapidly to the closest point on the manifold in which the error is negligible. Note the log scale of the error . **E.** Left: Tangential error Δ vs. location on the manifold. Right: Trajectories in (D) plotted against time. **F.-H.** Same network as in as **C-E**, but with a weak external input ($\epsilon = 0.01$) at $\pi/6$ rad. Here, convergence to the manifold is followed by drift to a single stable fixed point, in which the tangential error vanishes **H**. Decoded angular feature $\hat{\psi}$ is recovered via $\hat{\psi} = tan^{-1}(\hat{z}_2/\hat{z}_1)$. The reconstruction error, $\mathcal{L}$ is defined by $\mathcal{L} = |\hat{z} - z|$

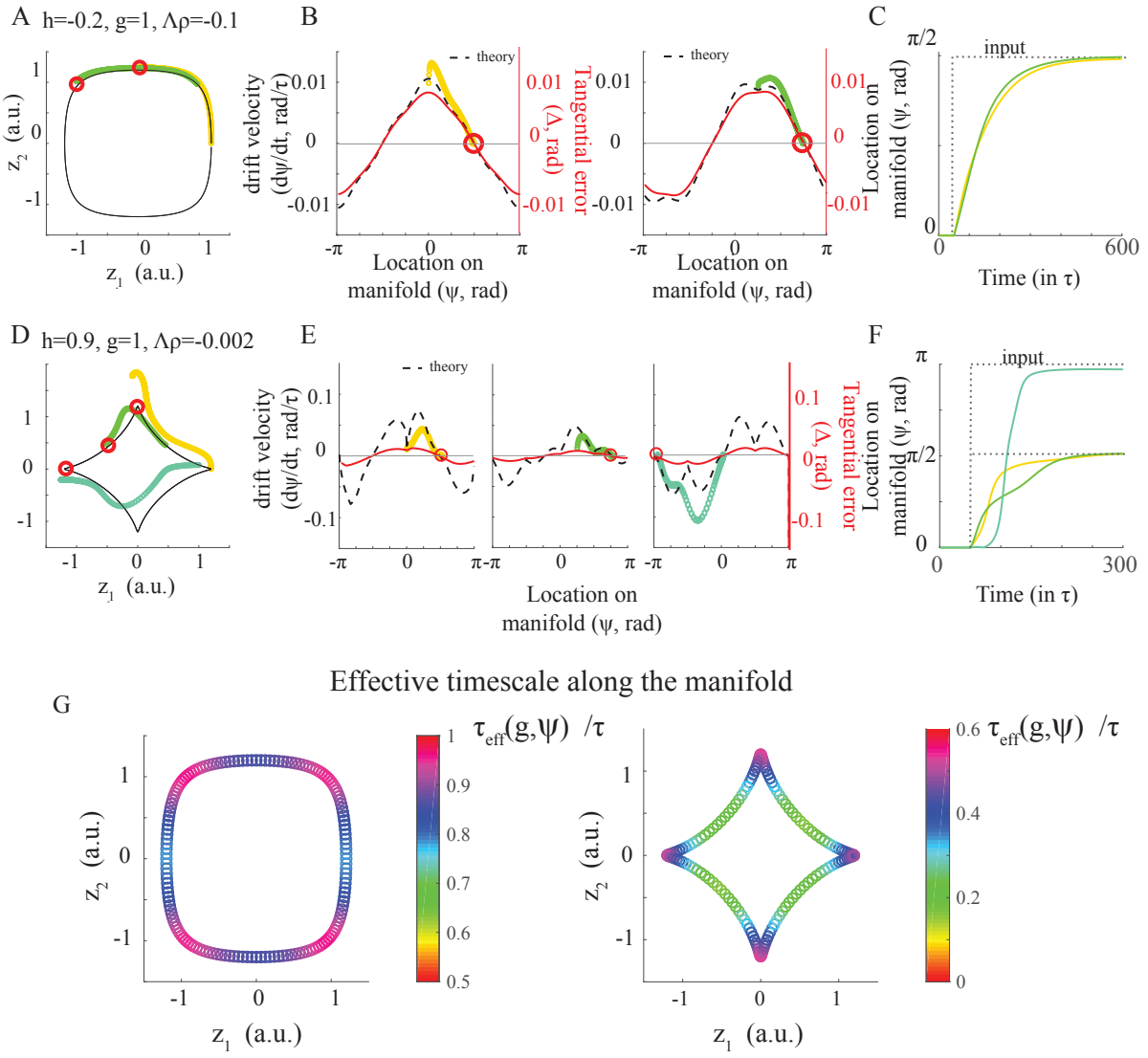

Figure S 4: Response to inputs in manifolds for which amplitude direction is not strongly attractive **A.-C.** Example manifold where amplitude direction approaches instability (maximal eigenvalue along all points on the manifold in amplitude direction is $\Lambda = -0.1$). Note deviations between theory and simulations due to the lack in timescale separation between the amplitude and manifold direction. **D.-F.** Example manifold where amplitude direction is marginal (maximal eigenvalue along all points on the manifold in amplitude direction is $\Lambda = -0.002$). Note that transitions are no longer along the manifold. **G.** Effective timescale along the manifold.

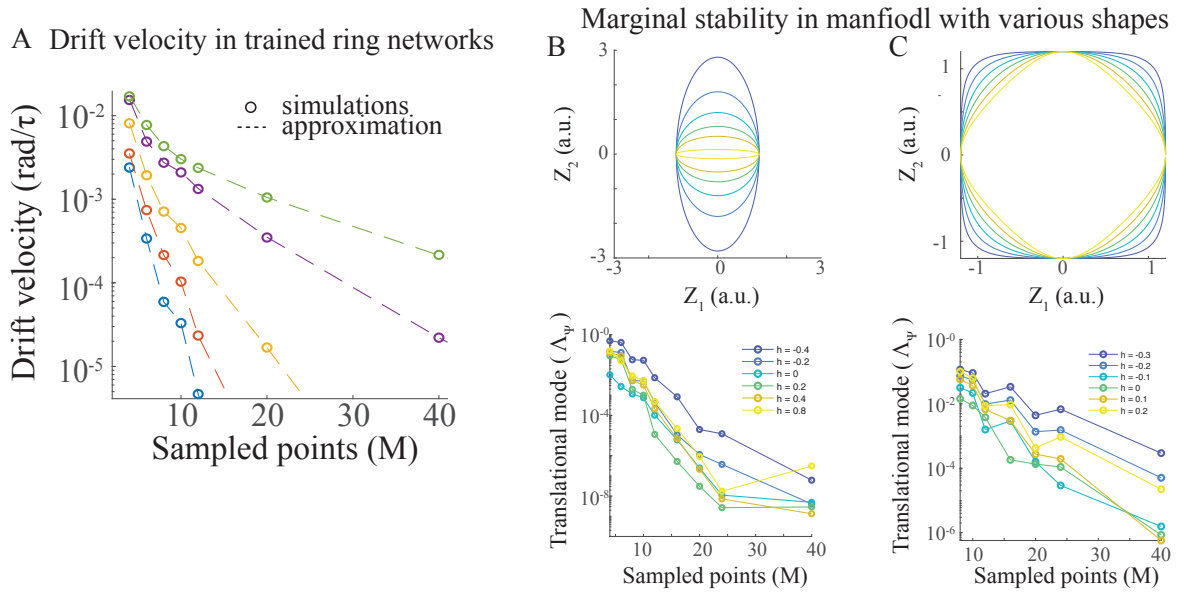

Figure S 5: Drift velocity and marginal directions in manifold attractors A. Drift velocity vs. number of sampled points in trained ring manifold. Same data as in Fig.4F. **B-C.** Build-up of manifold attractor beyond ring geometry. **B** Top: manifold shape projected on the decoder plane. Bottom: Stability along manifold direction for the manifolds in (A). Manifold parameterization: $\boldsymbol{z} = A[\cos(\psi), \frac{(1-h)}{1+h}\sin(\psi)]$ **C.** Same as (B), but for a different manifold, defined by $\boldsymbol{z} = \frac{A}{1+h/3}[\cos(\psi) + h/3\cos(3\psi), \sin(\psi) - h/3\sin(\psi)]$.

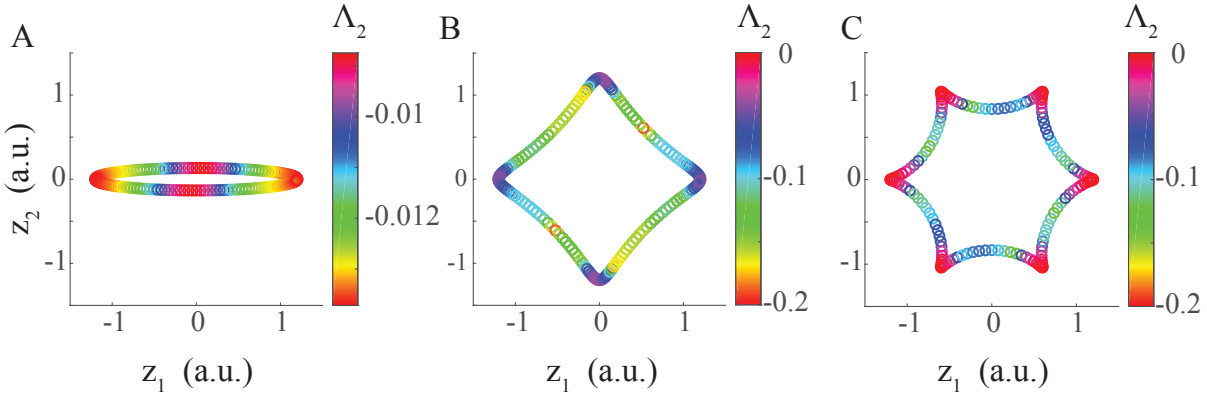

Figure S 6: Marginal stability in the amplitude direction tends to occur at reflection points A. The second largest eigenvalue of the stability matrix of the RNN, Eq.(13), plotted along the ellipse manifold. $a = 2, h = 0.8, g = 0.6$ **B.** Same as (A), but with $a = 4, h = 0.6, g = 1$. **C.** Same as (A), but with $a = 6, h = 0.9, g = 1.2$.

Tangential error in case of external input In the case of $0 < \epsilon \ll 1$ we find that it is sufficient to approximate the tangential error by

$$\Delta(\psi) \approx atan(\frac{f_2 + \epsilon\tilde{u}_2}{f_1 + \epsilon\tilde{u}_1}) - atan(f_2/f_1)$$

In other words, there is no need to calculate the error using the auxiliary RAE system and it can be devised solely from the manifold in the decoder plane, $\boldsymbol{f}$ and the the input $\tilde{\boldsymbol{u}}$. Specifically, for $\epsilon \ll 1$ we get $\Delta(\psi) \approx -\frac{\epsilon}{A}\sin(\psi - \psi_1)$.

Selectivity index and preferred direction Given the tuning curve of.a neuron i, $r(\theta_i, \psi)$, we calculate the first Fourier component, $r_1^k e^{i\Psi_1^k} = \int d\psi r(\theta_k, \psi)e^{i\psi}$, with Ψ_1^i being an estimate for the preferred direction of the neuron and r_1^k the selectivity index. Selectivity index of zero corresponds to a flat tuning curve and is closely related to a circular variance of 1.

Effective timescale With the gain of the RAE in Eq.(80), tangential component is obtained using Eq.(22) and the effective time scale τ_{eff} then follows:

$$\tau_{eff} = -\frac{d}{ds}G_{OL}^{\parallel}(s)|_{s=0} \tag{8}$$

Specifically, in the case of a trained ring manifold the effective timescale is given by Eq.(69). In comparison with simulations in Fig.3C we divided the drift velocity at $\psi = \pi/4$ for each level of heterogeneity by the drift velocity in the absence of heterogeneity.

Calculating spectrum of stability matrix In the case of a ring manifold or at reflection points we color the outliers of the spectrum of the $N \times N$ stability matrix according to their relation with either the bump's amplitude (circles in Fig.6B,C and green points in Fig.7A-Diii), or manifold direction (red points in Fig.7A-Diii). Specifically, we calculate the spectrum of $\boldsymbol{H} = -\boldsymbol{I} + (g\boldsymbol{J} + \boldsymbol{W}_\mu^{fb}\boldsymbol{W}_\mu^{outT})\phi'(x)$, for which $\mu = 1$ corresponds to the amplitude of the bump and $\mu = 2$ for the transnational direction.

Parameters for figures Unless written otherwise, the parameters are $A = 1.2, g = 1, N = 1000, M = 40, a = 2, h = 0$.

- Figure 1: A. $g = 0, A = 1.2$. C-D.$g = 1.2, A = 1.2$
- Figure 2: A. Black:$h = 0, g = 1$. Blue: $a = 2, h = 0.2, g = 1$. The predefined trained manifold in E-G is given by Eq.(7), with $a = 2, h = 0.2, g = 1$. In A $N = 16000$ and B-I $N = 4000$.

$$\text{Norm. pop. rate}(\psi) = \frac{\sqrt{\int d\theta\phi^2(x(\theta,\psi))}}{\sqrt{\int d\theta d\psi\phi^2(x(\theta,\psi))}} \tag{9}$$

- Figure 3: $\epsilon = 0.01$ A-D. $h = 0$. E-G. $a = 2, h = 0.2, g = 1$. C. N=4000.
- Figure 4: A. $N = 16000$. E-F, $N = 4000$.
- Figure 6: A. N=4000, a=4,h=-0.5,g=1. B. N=4000, a=2, g=0.2. C. N=4000, a=4, g=1.

- Figure 7: A. h=0, g=0.8. B. a=2, h=0.5, g=0.2. C. a=2, h=-0.5, g=0.2. D. a=4, h=-0.4, A=0.9. g=0.2.
- Figure S1: A-C. Symmetric-connectome ring attractor network ($W_{ij} \propto \cos(\theta_i - \theta_j)$, see Methods) in the presence of a random connectivity with $g = 1$. C. $\epsilon = 0.04$. D. Hyper-parameters for the scatter plot of drift speed vs. error $A \in \{0.5, 1.0, 1.2, 1.5, 2.0\}, g \in \{0.01, 0.1, 0.3, 0.5\}$, times 5 random seeds per setting.
- Figure S2: N=1000.
- Figure S3: $a = 4, h = -0.15, g = 1, A = 1.2$
- Figure S4: A-C,G left $a = 4, h = -0.2, g = 1, A = 1.2$ D-F,G. right$a = 4, h = 0.9, g = 1, A = 1.2$

5.1 The network model

We consider a network of N units following the rate dynamics:

$$\tau \frac{\boldsymbol{x}(t)}{dt} = -\boldsymbol{x}(t) + (\boldsymbol{W} + g\boldsymbol{J})\phi(\boldsymbol{x}(t)) + \epsilon \boldsymbol{u}(t) \tag{10}$$

with the neuronal state $\boldsymbol{x} \in \mathbb{R}^N$ representing the neuronal input and the neuronal rate given by $\phi(x_i(t))$, with a symmetric activation function, $\phi(x) = -\phi(-x)$. The recurrent connectivity consists of two components: The random heterogeneous part is represented by i.i.d Gaussian weights $J_{ij} \sim \mathcal{N}(0, 1/N)$ times strength parameter g [Sompolinsky et al., 1988]. The other recurrent component is a rank-2 structured part, $\boldsymbol{W} = \boldsymbol{W}_{fb}\boldsymbol{W}_{out}^T$, with $\boldsymbol{W}_{fb}, \boldsymbol{W}_{out} \in \mathbb{R}^{N\times 2}$. The external input is $\boldsymbol{u}(t) = \boldsymbol{W}_{in}\tilde{\boldsymbol{u}}(t)$, with $\tilde{\boldsymbol{u}} \in \mathbb{R}^2$, and where for simplicity we choose $\boldsymbol{W}_{in} = \boldsymbol{W}_{fb}$. The decoder is given by $z = \boldsymbol{W}_{out}\phi(\boldsymbol{x})$, for which the angular feature is calculated according to $\psi = atan(z_2/z_1)$. The training goal is to obtain $\boldsymbol{W}_{out}$ such that:

$$\boldsymbol{z}(\psi) = \boldsymbol{f}(\psi) \tag{11}$$

with target $f : [0, 2\pi) \to \mathbb{R}^2$ being a curve in 2D. For example, the particular case of a ring manifold is given by $\boldsymbol{f}(\psi) = A[\cos(\psi), \sin(\psi)]$. Finally, following the symmetric-connectome ring model of [Ben-Yishai et al., 1995], we choose $W_{1i}^{fb} = \cos(\theta_i)$ and $W_{2i}^{fb} = \sin(\theta_i)$, however other choices such as in [Mastrogiuseppe and Ostojic, 2018] give similar results (see also Section 5.9).

5.2 Dynamics in the vicinity of a manifold attractor

For a one-dimensional subset $\mathcal{M}_\Phi \subset \mathbb{R}^N$ to be a continuous attractor it is required that in the vicinity of any point $\phi(\boldsymbol{x}) \in \mathcal{M}_\Phi$, local dynamics will be convergent in $N - 1$ off-manifold dimensions and remain marginally stable in the one remaining direction associated with translations on the manifold (Fig 5A). In the case of the dynamics of Eq.(10), this implies that linearized system:

$$\tau \frac{d\boldsymbol{\delta x}(t)}{dt} = \boldsymbol{H}\delta\boldsymbol{x}(t) + \epsilon \boldsymbol{u}(t) \tag{12}$$

with the stability matrix

$$\boldsymbol{H} = -\boldsymbol{I} + (\boldsymbol{W} + g\boldsymbol{J})diag(\phi'(\boldsymbol{x})) \tag{13}$$

will have a spectrum $\{\lambda_i\}_{i=1}^N$ in which $\lambda_1 = 0$ and $\Re(\lambda_j) < 0\ \forall j \neq 1$. Extending the approach of [Rivkind and Barak, 2017] we argue that training the structured rank-two connectivity matrix affects only a small number of dynamical modes while the rest of the spectrum consists of a bulk of eigenvalues that are associated with random connectivity and are not affected by the structured component (red and green circles in Fig. 5C,E,6). The latter are confined to a circle of radius (purple in Figs. 5C, 7, [Ahmadian et al., 2015]):

$$\Lambda_{pop} = g\sqrt{\langle \phi'(x_i)^2 \rangle_i} \tag{14}$$

with x_i calculated self-consistently (see below) and $\langle . \rangle_i$ average over all neurons.

5.3 Recurrent Autoencoder setting

To analyze the dynamics of continuous attractor networks we consider the dynamics of an auxiliary setting, which we call Recurrent Autoencoder (RAE) (Fig.S3A):

$$\begin{aligned} \tau\frac{d\boldsymbol{x}(t)}{dt} &= -\boldsymbol{x}(t) + g\boldsymbol{J}\phi(\boldsymbol{x}(t)) + \boldsymbol{W}_{fb}\boldsymbol{z}(\psi) + \epsilon\boldsymbol{u}(t) \\ \hat{\boldsymbol{z}}(\psi, t) &= \boldsymbol{W}_{out}^T\phi(\boldsymbol{x}(t)) \end{aligned} \tag{15}$$

where we unfold the structured component of the recurrent connectivity from Eq.(1) (Fig.1Ci), and test what would be the *decoded* output of the RAE, $\hat{\boldsymbol{z}}$, when externally enforcing its input through the *encoder*, $\boldsymbol{W}_{fb}$, to be $\boldsymbol{z}$.

The dynamics of the RNN (Eq.(1)) is governed by the discrepancy between the input and the output of the auxiliary system, $\mathcal{L} = ||\hat{\boldsymbol{z}} - \boldsymbol{z}||^2$. Namely, given a point $\boldsymbol{x}$ on the attractor, and its corresponding low-D projection $\boldsymbol{z}$, the dynamics of Eq.(1) should regenerate the same point persistently. In the RAE setting, this regeneration coincide with having $\hat{\boldsymbol{z}} = \boldsymbol{z}$. Conversely, a point $\boldsymbol{x}$ in the vicinity of the attractor would not be regenerated perfectly, resulting in a flow in the dynamics of Eq.(1).

As a result of the difference between dynamics along the on- and off- manifold directions, any neural trajectory in the vicinity of the manifold will first converge in the N-1 dimensions orthogonal to the 1D manifold, followed by slower dynamics along the manifold itself. Consequently, it is useful to consider the tangential error of the RAE. This error is given by projecting the 2D error vector, $\hat{\boldsymbol{z}} - \boldsymbol{z}$, on the tangent to the manifold at point ψ and is approximated by:

$$\Delta(\psi) \approx \hat{\psi}(\psi) - \psi \tag{16}$$

It measures the difference between the angle of the reconstructed point of the RAE, $\hat{\psi} = tan^{-1}(\hat{z}_2/\hat{z}_1)$ and the input angle ψ (Fig.S3A,G).

5.4 The gain of the RAE

Linearizing Eq.(15) around a putative fixed point, $\boldsymbol{x}(\psi)$ yields:

$$(1+s\tau)\boldsymbol{X}(\psi,s) = g\boldsymbol{J}diag(\phi'(\boldsymbol{x}(\psi)))\boldsymbol{X}(\psi,s)+\boldsymbol{W}_{fb}\boldsymbol{Z}(\psi,s) \quad (17)$$

$$\hat{\boldsymbol{Z}}(\psi,s) = \boldsymbol{W}_{out}^{T}\phi'(\boldsymbol{x})\boldsymbol{X}(\psi,s) \quad (18)$$

where $\boldsymbol{X}(\psi,s)$ is the Laplace transform of the linearized state. The gain of the (open loop) RAE is defined as:

$$\hat{\boldsymbol{Z}}(s) = \boldsymbol{G}_{OL}(s)\boldsymbol{Z}(s) \quad (19)$$

and can be computed as:

$$G_{OL}(s) = \boldsymbol{W}_{out}^{T}diag(\phi'(\boldsymbol{x}))[(1+s\tau)\boldsymbol{I}-g\boldsymbol{J}diag(\phi'(\boldsymbol{x}))]^{-1}\boldsymbol{W}_{fb} \quad (20)$$

By closing the loop, i.e. by setting $\boldsymbol{Z}(\psi,s)=\hat{\boldsymbol{Z}}(\psi,s)$ in Eq. (17), we obtain the gain of the fully recurrent network:

$$\boldsymbol{G}(s) = (\boldsymbol{I}-\boldsymbol{G}_{OL}(s))^{-1}\boldsymbol{G}_{OL}(s) \quad (21)$$

$\boldsymbol{G}(s)$ is a 2×2 matrix of N degree polynomial ratios and the poles of its determinant correspond to a subset of eigenvalues of the stability matrix of Eq.(13). In large N limit this degree is vastly reduced as the majority of linear dynamical modes become not observable via readout $\boldsymbol{Z}$, and it is only a small number of modes that persist in (19) and hence in (21). Specifically, these are the eigenvalues that appear due to the structured component of the connectivity and the bulk of remaining eigenvalues are not affected and obey circular law (Fig.5A and Fig.7).

A marginal stability emerges when $\det(G)$ has a pole at $s=0$. Equivalently, it follows from Eq.(21) that $\boldsymbol{G}_{OL}(s=0)$ must have an eigenvalue *one*, with the eigenvector being the tangent vector, $\hat{\boldsymbol{t}} = \frac{dz_\psi}{d\psi}||\frac{dz_\psi}{d\psi}||^{-1}$. To analyze the dynamics it is convenient to consider the gain of the RAE in the coordinates of the tangent and normal directions, with the normal direction, $\hat{\boldsymbol{n}}$, defined by a clockwise rotation of $\hat{\boldsymbol{t}}$ by $\pi/2$. We define the gain in the tangent direction to the manifold as

$$G_{OL}^{\parallel}(s) = \hat{\boldsymbol{t}}\boldsymbol{G}_{OL}(s)\hat{\boldsymbol{t}} \quad (22)$$

and similarly in the gain in the normal direction as $G_{OL}^{\perp}(s)=\hat{\boldsymbol{n}}\boldsymbol{G}_{OL}(s)\hat{\boldsymbol{n}}$. The conditions for a stable manifold are then that $G_{OL}^{\perp}(s)$ obey stability conditions of a scalar feedback system, (e.g. Nyquist criterion [Nyquist, 1932]), while $G_{OL}^{\parallel}(s)$ is required to obey marginality:

$$G_{OL}^{\parallel}(s=0)=1 \quad (23)$$

In the sequel we argue that the local dynamics along the manifold in this case can be approximated by first order differential equation, even for large g (see Section 5.7.3). Consequently it must have a form:

$$G_{OL}^{\parallel}(s) \approx \frac{1}{1+\tau_{eff}s} \quad (24)$$

Finally, we note that small and slow translations along the manifold does not induce displacement at the normal direction, implying $\hat{\boldsymbol{n}}G_{OL}(s=0)\hat{\boldsymbol{t}}=0$. The second cross term $\hat{\boldsymbol{t}}G_{OL}(s=0)\hat{\boldsymbol{n}}$ is non-vanishing in general case, however, by symmetry considerations, it vanishes if the manifold features reflection symmetry around $\hat{\boldsymbol{n}}$. In particular, this condition holds at any point for ring geometry and for reflection points in the parameterized manifolds of Fig.6-7.

5.5 Drift and connection to RAE framework

We assume that the dynamics in the N-1 directions orthogonal to the 1D manifold, $\mathcal{M}_\phi$, is convergent, and that the tangential error following training is small (or conversely $\epsilon \ll 1$ in the case of external input), such that translations on the manifold are slower than the dynamics in the N-1 orthogonal directions. Under this separation of timescale assumption, we can link the drift along the manifold to the tangential error, as shown in the main text (Eq.(4)).

We start by considering the linearized RAE in the transnational direction. Adding a constant error $\Delta(\psi,t)=\Delta(\psi)$ yields:

$$\hat{\psi}(s)=G_{OL}^{\parallel}(s)\psi(s)+\delta(s)\Delta(\psi) \tag{25}$$

where δ is Dirac delta function. Closing loop $\hat{\psi}(s)=\psi(s)$:

$$\delta(s)\Delta(\psi)=(1-G_{OL}^{\parallel}(s))\psi(s) \tag{26}$$

Assuming we can write the gain in the transnational direction as in Eq.(24) we get:

$$\psi(s)=\left(1+(\tau_{eff}s)^{-1}\right)\delta(s)\Delta(\psi) \tag{27}$$

Recalling that the factor s^{-1} in Laplace domain translates into integral in the time domain, we get:

$$\psi(t)=(1+\tau_{eff}^{-1}t)\Delta(\psi) \tag{28}$$

The drift along the manifold then follows:

$$\dot{\psi}=\Delta(\psi)\tau_{eff}^{-1} \tag{29}$$

recapitulating Eq.(4) in the main text. Linerazing Eq.(29) around a fixed point:

$$\dot{\psi}=\frac{d\Delta(\psi)}{d\psi}\psi\tau_{eff}^{-1} \tag{30}$$

and:

$$\Lambda_\psi=\frac{d\Delta(\psi)}{d\psi}\tau_{eff}^{-1} \tag{31}$$

with stability achieved for $\Lambda_\psi<0$, and marginality for $\Lambda_\psi=0$. Here we assumed (without loss of generality) that the fixed point is located at $\psi=0$.

5.6 Trained manifold attractors

We train the network by sampling $M \ll N$ points of a pre-defined manifold, $z(\psi_m) \in \mathcal{M}_2$, with $m = 1...M$, $\psi_m = 2\pi\frac{m}{M}$ and run the dynamics in an open loop RAE setting (Eq.(15),Fig.S3) with $\epsilon = 0$ until the recurrent dynamics converges [Jaeger, 2001, Sussillo and Abbott, 2009]. We obtain M states $\{\phi(x_m)\}_{0\leq m\leq M-1}$ and train the decoder by minimizing the reconstruction error:

$$\mathcal{L}(\boldsymbol{W}_{out}) = |\hat{\boldsymbol{z}}(\psi, A) - \boldsymbol{z}(\psi, A)|_2^2 \tag{32}$$

with $\hat{\boldsymbol{z}} = \boldsymbol{W}_{out}^T\phi(\boldsymbol{x})$. The least square (LS) solution yields:

$$\boldsymbol{W}_{out} = \boldsymbol{\Phi}\boldsymbol{C}^{-1}\bar{\boldsymbol{z}} \tag{33}$$

where here $\bar{\boldsymbol{z}} \in \mathbb{R}^{M\times 2}(\bar{\boldsymbol{z}} = \boldsymbol{z}(\psi_m))$, the fixed point solution of Eq.(15) are

$$\boldsymbol{\Phi}_{im} = \phi(x(\theta_i, \psi_m)) \tag{34}$$

and the correlation between the rates is:

$$\boldsymbol{C} = \boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{35}$$

A point on the manifold can thus be written using the correlation matrix:

$$\hat{\boldsymbol{z}}(A, \psi) = \sum_{m=0}^{M-1} \boldsymbol{z}(A, \psi_m)C(A; \psi_m, \psi) \tag{36}$$

We next consider the singular value decomposition (SVD) of $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi} = \boldsymbol{v}\boldsymbol{D}^{1/2}\boldsymbol{\eta} \tag{37}$$

where $\boldsymbol{\eta} \in \mathcal{R}^{M\times M}$ and $\boldsymbol{v} \in \mathcal{R}^{N\times N}$ are the right and left singular vectors. The matrix $\boldsymbol{D}^{1/2} \in \mathcal{R}^{N\times M}$ is the singular value (SV) matrix with $\sqrt{C_m}$, $m = 1..M$, being the SVs and C_m being the elements of the spectrum of the correlation matrix of Eq.(35). The decoder is then:

$$\boldsymbol{W}_{out} = \boldsymbol{v}\boldsymbol{D}^{-1/2}\boldsymbol{\eta}\boldsymbol{f} \tag{38}$$

Due to stability requirements (Fig.5A and Section5.2), the above training procedure does not guarantee the emergence of a manifold attractor. First, a translational mode needs to emerge from sampling only a finite set of M points (red arrow in Fig.5A). Second, amplitude direction needs to be stabilized (green arrow in Fig.5A). Finally, spontaneous activity must be suppressed [Rajan et al., 2010, Mastrogiuseppe and Ostojic, 2018] (purple arrow in Fig.5A). To assess that, we developed a mean field approach, calculated the steady state mean field solution of Eq.(10), and analyzed its stability.

5.7 Heterogeneous trained ring attractor model

To train a ring manifold on top of heterogeneous connectivity ($g > 0$ in Eq. (1)) we apply the least square learning rule from Eq. (33) to M samples $z_{0..M-1}$ from the ring curve $\mathcal{M}_z$:

$$\boldsymbol{z}(\psi) = A[\cos(\psi), \sin(\psi)] \tag{39}$$

We will argue in the sequel (Eq. (45)) that in the limit of large N and for any M the correlation matrix is circulant. Consequently $\boldsymbol{z}(\psi)$ is an eigenvector of $\boldsymbol{C}$ and the LS solution is of a particularly simple form:

$$\boldsymbol{W}_{out} = \frac{1}{C_1}\boldsymbol{\Phi}\bar{\boldsymbol{z}} \tag{40}$$

Remarkably, as depicted in Fig.1E and as quantified by Eq.(53), circulant property of correlation matrix does not imply symmetry between individual neuronal representations. Consequently, the structured part of the recurrent interactions, $\boldsymbol{W}$, is not symmetric any more: while feedback weights are assumed to be $\boldsymbol{W}_i^{fb} = [\cos(\theta_i), \sin(\theta_i)]$ as in the symmetric-connectome model, this is not the case for the learned readout weights: $\boldsymbol{W}_{,i}^{out} \neq w[\cos(\theta_i), \sin(\theta_i)]$ for any factor w.

To assess dynamical properties of the putative manifold, we now turn to developing a mean field estimate for the open loop RAE gain (19).

5.7.1 Mean field solution for the neural representation of the manifold

To obtain the mean field solution in real space we decompose the steady state solution of Eq.(15) into its deterministic and stochastic parts [Rajan et al., 2010, Rivkind and Barak, 2017]:

$$x_i(\psi_m) = x_i^0(\psi_m) + x_i^1(\psi_m) \tag{41}$$

where the deterministic part, independent of disordered connectivity J, is

$$x_i^0(\psi_m) = A\cos(\theta_i - \psi_m) \tag{42}$$

and the quenched disorder is given by:

$$x_i^1(\psi_m) = g\sum_j J_{ij}\phi_j(\psi_m) \tag{43}$$

In the large N limit, x_i^1 is replaced by a Gaussian r.v., σy, with y being a mean zero and unit Gaussian r.v. and the variance, σ^2, that needs to be evaluated self-consistently:

$$\sigma^2 = g^2\int\frac{d\theta}{2\pi}\int Dy\phi^2(\sigma y + A\cos(\theta)) \tag{44}$$

with $Dy = \frac{e^{-y^2/2}dy}{2\pi}$. Importantly, due to the rotational symmetry in the deterministic part of Eq.(42), σ is independent of θ and ψ_m in the large N limit. Moreover, since we assumed that $\phi(-x) = -\phi(x)$, there is no bias term in Eq.(44).

Correlation between the inputs to the neurons on the manifold $c(\psi_m, \psi_{m'}) = \langle x_i(\psi_m)x_i(\psi_{m'})\rangle$ can be computed self consistently as well. This is done by projecting Eq.(43) for ψ_m onto the same equation but with $\psi_{m'}$. In the specific case of a ring geometry, the latter only depends on the difference

$\psi_m - \psi_{m'}$ so to simplify notations we assume $\psi_{m'} = 0$ and obtain:

$$c(\psi_m) = g^2 C(\psi_m) = g^2 \int \frac{d\theta}{2\pi} \int Dy \Big(\int Dy_1 \phi \left(\sqrt{\sigma^2 - |c(\psi_m)|} y_1 + \sqrt{|c(\psi_m)|} y + A\cos(\theta) \right) \int Dy_2 \phi \left(\sqrt{\sigma^2 - |c(\psi_m)|} y_2 + sign(c(\psi_m)) \sqrt{|c(\psi_m)|} y + A\cos(\theta - \psi_m) \right) \Big) \tag{45}$$

such that $c(0) = \sigma^2$ and where we denote by capital latter the the correlations among the rates of the neural state:

$$C(\psi_m) = \langle \phi(x_i(\psi_m)) \phi(x_i(\psi_0)) \rangle \tag{46}$$

5.7.2 Representation in Fourier space

Instead of mean field estimate in real space, we can also write the fixed point of the dynamics by the singular value decomposition of Eq.(37). This turns out to be handy when considering the stability analysis, as we can write the decoder using the SVD. Using Eq.(37), the quenched disorder is

$$x^1_{im} = x^1_i(\psi_m) \equiv \sum_{n=1}^{M} g C_n^{1/2} a_{ni} \eta_n(\psi_m) \tag{47}$$

where $a_{ni} = \sum_j J_{ij} v_{nj}$ are Gaussian r.v. with zero mean and unit variance (and we assume $\frac{1}{N}\sum v^2_{nj} = 1$) and the spectrum of the correlation function is:

$$C_n = \sum_m \cos(n\psi_m) C(\psi_m) \tag{48}$$

In the case of the ring attractor, the correlation matrix is circulant (Fig.2A). As a result, the SVs are simply the spatial Fourier modes. In this case we denote the coefficients of even (resp. odd) modes corresponding to cos (resp. sin) functions by a (resp. b), as opposed to a case of general principal component decomposition where we do not make such a distinction (see Section 5.8). Eq.(47) thus yields:

$$x^1(\theta, \psi_m) \equiv \sum_{n=1}^{M} \sqrt{c_n} (a_n(\theta) \cos n\psi_m + b_n(\theta) \sin n\psi_m) \tag{49}$$

where we took the limit of $N \to \infty$ such that $a_{ni}, b_{ni} = a_n(\theta), b_n(\theta)$ and where we define

$$c_m = 2 \int \frac{d\psi}{2\pi} \cos(m\psi) c(\psi_m)$$

We can thus write the fixed point solution using the Gaussian r.v. a_n, b_n:

$$\begin{aligned} c(\psi_m) &= g^2 \int \frac{d\theta}{2\pi} \Pi_{n=1}^{M} \int Da_n \int Db_n \phi\Big(\sum_{n=1}^{M} \sqrt{c_n}(a_n(\theta)\cos n\psi_m + b_n(\theta)\sin n\psi_m) + A\cos(\theta - \psi_m)\Big) \\ &\times \phi\Big(\sum_{n=1}^{M} \sqrt{c_n} a_n(\theta) + A\cos\theta\Big) \end{aligned} \tag{50}$$

and as before, with $\sigma^2 = c(0)$. Thus the statistics of the neuronal activity is fully specified in large N limit. In principle, to estimate Eq.(50), one would need to integrate over a large number, $2M$, of random

variables a_n, b_n. In practice, however, one can get a very good approximation for the correlation and the variance by using a rather small number of random variables. This is because as the spectrum of the correlation function decreases rapidly when n is increased, the amplitude of the higher frequency components in the quenched disorder is increasingly small. We thus write the approximate MF soultion in Fourier space by taking the approximation

$$\hat{x}^1(\theta, \psi_m) \equiv \sum_{n=1}^{K} \sqrt{c_n}(a_n(\theta)\cos n\psi_m + b_n(\theta)\sin n\psi_m) \tag{51}$$

and where we defined the cut-off frequency, K ($K < M$).

At this point we can also estimate the diversity of tuning curves. Tuning curves are calculated when changing ψ per neuron:

$$r(\theta_i, \psi) = \phi(A\cos(\theta_i - \psi) + \hat{x}^1(\theta_i, \psi))$$

We can thus use the Gaussian statistics of $x^1(\theta_i, \psi)$ to generate tuning curves. Again, in principle this would require 2M random variables per one neuron, but in practice $K = 5$ is enough. We define the selectivity index of a neuron i as:

$$SI_k = r_1^k = \left| \int d\psi e^{i\psi} \phi(x_k(\psi))) \right| \tag{52}$$

and we can further calculate the SD of the selectivity index, yielding:

$$\begin{aligned} SD_{SI}^2 = \langle SI_k^2 \rangle_k - \langle SI_k \rangle_k^2 &= \int \frac{d\theta}{2\pi} \left| \int Da \int Db \int d\psi e^{i\psi} \phi(A\cos(\theta - \psi) + \hat{x}^1(a, b, \psi)) \right|^2 \\ &- \left\{ \int \frac{d\theta}{2\pi} \left| \int Da \int Db \int d\psi e^{i\psi} \phi(A\cos(\theta - \psi) + \hat{x}^1(a, b, \psi)) \right| \right\}^2 \end{aligned} \tag{53}$$

5.7.3 Dynamics in the vicinity of the attractor - mean field solution

Equipped with the mean field solution for the neuronal state on the attractor, we are now set to explore the dynamics in its vicinity by evaluating Eq.(19) and (21). Rather than applying Eq.(20) directly, we estimate the elements $G_{OL}^{\mu\nu}$ of 2x2 matrix $\boldsymbol{G}_{OL}$ by the mean field approximations of $\hat{\boldsymbol{Z}}_\mu$ (Eq.(18)) driven by the appropriate input $\boldsymbol{Z_\nu}(s) = \hat{\boldsymbol{e}}_\nu$ (with $\hat{\boldsymbol{e}}_\nu$ a unity vector id the direction ν). Recalling that according to Eq.(33), the linear decoder $\boldsymbol{W}_{out}$ is spanned by rate-states $\phi(x_i(\psi_m))$:

$$\boldsymbol{W}_\mu^{out} = \sum_{m=0}^{M-1} q_{\mu m} \phi(\boldsymbol{x}(\psi_m)) \tag{54}$$

where $\nu, \mu \in \{1, 2\}$ denote directions in the decoder plane, $\boldsymbol{z}$. The terms of the RAE gain matrix (19) are then:

$$G_{OL}^{\mu\nu}(s) = \sum_{m=0}^{M-1} q_{\mu m} \langle \phi(x_i(\psi_m))\phi'(x_i(\psi_0))X_{i\nu}(\psi_0, s) \rangle \tag{55}$$

Similarly to Eq.(41), we decompose the linearized response:

$$X_{i\nu}(\psi_0, s) = X^0_{i\nu}(\psi_0, s) + X^1_{i\nu}(\psi_0, s) \tag{56}$$

$$(1+s)X^0_{i\nu}(\psi_0, s) = W^{fb}_{i\nu} \tag{57}$$

$$(1+s)X^1_{i\nu}(\psi_0, s) = g\sum J_{ij}\phi'(x_i(\psi_0))X_{j\nu}(\psi_0, s) \tag{58}$$

multiplying Eq.(58) by Eq.(43) and project on the right singular vectors, η_{nm}, we obtain:

$$(1+s)\sum_{m=0}^{M-1}\eta_{nm}\langle x^1_i(\psi_m)X^1_{i\nu}(\psi_0)\rangle = g^2\sum_{m=0}^{M-1}\eta_{nm}\langle\phi(x_i(\psi_m))\phi'(x_i(\psi_0))X_{i\nu}(\psi_0)\rangle \tag{59}$$

The RHS of Eq.(59) is closely related to the gain of the RAE. In fact, due to circulant property of $\boldsymbol{C}$ in the case of a ring geometry, the Fourier modes are the principal components of $\boldsymbol{\Phi}$, and q_{lm} takes a particularly simple form (Eq.(40)): $q_{\mu m} = \frac{1}{C_1}\eta_{\mu m}$ ($\mu \in \{1, 2\}$). It is thus convenient to solve the problem in the Fourier domain with respect to ψ. For the sake of convenience, we focus on the point $\psi = 0$. We express X in the corresponding spatial basis:

$$X_{i\nu}(s) = \frac{W^{fb}_{i\nu}}{1+s} + \sum_k \sqrt{c_k}(\alpha^\perp_k(s)a_{ki} + \alpha^\parallel_k(s)b_{ki}) \tag{60}$$

where we omit the argument ψ_0 to simplify notations and $\alpha^\perp(s)$ (resp. $\alpha^\parallel(s)$) are coordinates of $X(s)$ in the a_k (resp. b_k) basis. Here, $\mu, \nu = 1$ correspond to the direction normal to the manifold and $\mu, \nu = 2$ corresponds to the tangential direction. We start by analysing the tangential direction, and we use the symbol $\parallel$ as shorthand for indexes 2 and $2, 2$ (and similarly $\perp$, for 1 and $1, 1$). Projecting Eq.(49) on Eq.(60) and using Eq.(59) yields a self-consistent equations for the projections in the tangential direction:

$$(1+s)c_n\alpha^\parallel_n(s) = g^2\frac{\beta^{\parallel 0}_n}{1+s} + g^2\sum_{n'=1}^{K}\beta^{\parallel 1}_{nn'}\alpha^\parallel_{n'}(s) \tag{61}$$

with

$$\beta^{\parallel 0}_n = \prod_{k=1}^{K}\int Da_k Db_k \int\frac{d\psi}{\pi}\int\frac{d\theta}{2\pi}\sin(n\psi)\phi(A\cos(\theta-\psi)+\hat{x}^1(\psi))\phi'(A\cos(\theta)+\sum_{k=1}^{K}\sqrt{c_k}a_k)\sin(\theta) \tag{62}$$

$$\beta^{\parallel 1}_{nn'} = \prod^{K}\int Da_k Db_k \int\frac{d\psi}{\pi}\int\frac{d\theta}{2\pi}\sin(n\psi)\phi(A\cos(\theta-\psi)+\hat{x}^1(\psi))\phi'(A\cos(\theta)+\sum_{k=1}^{K}\sqrt{c_k}a_k)\sqrt{c_n}b_{n'} \tag{63}$$

where we took the limit $M \to \infty$ (assuming $M \ll N$) and $\hat{x}^1$ is given by Eq.(51), an approximation of Eq.(49) up to K-th order.

To obtain the gain of the RAE we note that $\alpha^\perp_1$ is closely related to the tangential gain. Indeed, from Eq.(61) and (59) the gain of the autoencoder in the manifold direction now follows:

$$G^\parallel_{OL}(s) = (1+s)\frac{A}{g^2C_1}c_1\alpha^\parallel_1(s) = A\alpha^\parallel_1(s)(1+s) \tag{64}$$

and $\alpha_1^{\|}$ is the first element of vector $\boldsymbol{\alpha}^{\|}$:

$$\boldsymbol{\alpha}^{\|}(s) = \frac{g^2}{1+s}\left((1+s)\boldsymbol{diag(c)} - g^2\boldsymbol{\beta}^{\|1}\right)^{-1}\boldsymbol{\beta}^{\|0} \tag{65}$$

How many terms K are needed to estimate $\boldsymbol{\alpha}^{\|}(s)$ and hence $G_{OL}^{\|}(s)$? Using Stein's lemma, we write Eq.(63) as

$$\beta_{nn'}^{\|1} = c_{n'}\langle \sin(n\psi)\sin(n'\psi)\phi'(x_i(\psi))\phi'(x_i(0))\rangle \tag{66}$$

this implies an exponential decay with frequency difference $n-n'$. Furthermore, it can be shown that for $s=0$ the solution of (65) is given by $\alpha_n^{\|} = n$. Such a linear growth of elements of $\alpha_n^{\|}$ can not overcome decay of (66) and taking $K=1$ is enough. Numerical simulations also point that first order approximation $K=1$ is accurate for all the relevant ranges of s (not shown).

We can now recover the effective timescale τ_{eff} that governs the dynamics along the manifold (Eq.(4)). By assuming first order ansatz for the gain of the RAE we get:

$$G_{OL}^{\|}(s) = \frac{A\beta_1^{\|0}}{C_1(1+s) - \beta_{11}^{\|1}} \tag{67}$$

Reorganizing the above equation and recalling that the gain of the RAE at zero frequency is unity (see Eq.(23), yielding $\frac{1-\frac{1}{C_1}\beta_{11}^{\|1}}{\frac{A}{C_1}\beta_1^{\|0}} \approx 1$) gives:

$$G_{OL}^{\|}(s) \approx \frac{1}{1+\tau_{eff}s} \tag{68}$$

with the effective timescale given by:

$$\tau_{eff} = \tau(AC_1^{-1}\beta_1^{\|0})^{-1} \approx \tau(1-C_1^{-1}\beta_{11}^{\|1})^{-1} = \tau(1-\beta)^{-1} \tag{69}$$

and where using Eq.(66) we get Eq.(5) in the main text:

$$\beta = g^2\langle \sin(\psi)^2\phi'(x_i(\psi))\phi'(x_i(0))\rangle_{i,\psi}$$

Following the same analysis, similar equations control the gain in the amplitude direction, but with:

$$\beta_n^{\perp 0} = \prod_{k=1}^{K}\int Da_k Db_k \int\frac{d\psi}{\pi}\int\frac{d\theta}{2\pi}\cos(n\psi)\phi(A\cos(\theta-\psi)+\hat{x}^1(\psi))\phi'(A\cos(\theta)+\sum_k\sqrt{c_k}a_k)\cos(\theta) \tag{70}$$

$$\beta_{nn'}^{\perp 1} = \prod_{k=1}^{K}\int Da_k Db_k \int\frac{d\psi}{\pi}\int\frac{d\theta}{2\pi}\cos(n\psi)\phi(A\cos(\theta-\psi)+\hat{x}^1(\psi))\phi'(A\cos(\theta)+\sum_k\sqrt{c_k}a_k)\sqrt{c_{n'}}a_{n'} \tag{71}$$

Here, by using stein lemma we get

$$\beta_{nn'}^{\perp 1} = c_{n'}\left[\langle\cos(n\psi)\cos(n'\psi)\phi'(x_i(\psi))\phi'(x_i(0))\rangle + \langle\cos(n\psi)\phi(x_i(\psi))\phi''(x_i(0))\rangle\right] \tag{72}$$

the gain $G_{OL}^{\perp}$ is computed similarly to Equations (64), (65) with leading eigenvalue Λ_ρ of the full system corresponding to the leading pole of the closed loop gain $G^{\perp}(s) = G_{OL}^{\perp}(s)(1-G_{OL}^{\perp}(s))^{-1}$.

5.7.4 Finite sampling effect

As $M < N$, the tangential error is zero at the learned points. Nevertheless, as long as M is finite, the derivative of the tangential error is not zero. Due to the symmetry between the sampled points in the large N limit we can calculate the derivative of the tangential error at only one of the sampled points. Without loss of generality, we calculate the derivative at $\psi_0 = 0$. The angular tangential error at $\psi_0 = 0$ is given by $\frac{\hat{z}_2(\psi_0) - z_2(\psi_0)}{A}$. Substituting into (40) we obtain:

$$\Delta(\psi_0, M) = \frac{1}{C_1} \sum_m \sin(\psi_m) \langle \phi(\theta, \psi_m) \phi(\theta, \psi_0) \rangle - \psi_0 \tag{73}$$

yielding:

$$\Delta(\psi_0, M) = \frac{1}{C_1} \sum_{m=0}^{M-1} \sin(\psi_m) C(\psi_m - \psi_0) - \psi_0 \tag{74}$$

To obtain the derivative of the tangential error we define the continuous correlation function, $C(\psi)$. In this sense, the notation $C'(\psi_m)$ refers to $\frac{dC(\psi)}{d\psi}|_{\psi=\psi_m}$. We expand the (continuous) correlation function, which is an even function of ψ, in Fourier series as $C(\psi) = \sum_{k=0}^{\infty} C_k \cos(k\psi)$ and write the derivative of the tangential error at the sampled points:

$$\Delta'(\psi_0, M) = \frac{\sum_{m=0}^{M-1} \sin(\psi_m) C'(\psi_m)}{C_1} - 1 = \frac{\sum_{k=0}^{\infty} \sum_{m=0}^{M-1} k C_k \sin(\psi_m) \sin(k\psi_m)}{\sum_{k=0}^{\infty} \sum_{m=0}^{M-1} C_k \cos(\psi_m) \cos(k\psi_m)} - 1 \tag{75}$$

First, it is now obvious from RHS of (75) that $\lim_{M\to\infty} \Delta' = 0$. Furthermore, correction due to finite M is expressible via spectrum of C. Namely:

$$\Delta'(\psi_0, M) = \frac{1 - C_1^{-1} \sum_{k'=1}^{\infty} \left((Mk' - 1) C_{Mk'-1} - (Mk' + 1) C_{Mk'+1} \right)}{1 + C_1^{-1} \sum_{k'=1}^{\infty} \left(C_{Mk'-1} + C_{Mk'+1} \right)} - 1 \tag{76}$$

In case of a fast decaying spectrum of the correlation function, such that $C_k \gg C_{k'}$ for $k' > k$, as in cases shown in Fig. 4D, Eq.(76) can be approximated by:

$$\Delta'(\psi_0, M) \approx -(M-1) \frac{C_{M-1}}{C_1} \tag{77}$$

We can now plug this result, along with the effective time constant from (69) into (31) to obtain mean-field estimate for the maximal eigenvalue in the translational direction:

$$\Lambda_\psi \approx -(M-1) \frac{C_{M-1}}{C_1} \tau_{eff}^{-1} \tag{78}$$

and conclude that translational direction is recovered exponentially fast with the number of sampes M. Importantly, it must be acknowledged that the result relies on approximating $G_{OL}^{\parallel}$ by a first order in term. While this approximation matches the numerical simulations, it is not immediately clear why approximation remains valid for estimating exponentially small quantities such as (78), that may be potentially sensitive to small terms that were neglected in (24).

5.8 General 1D manifold attractor

Similarly to the case of the ring manifold, we write the steady state mean field solution and stability equations for the dynamics in Eq.(10) for a general 1D manifold. In contrast to the case of a ring manifold, in manifolds with general geometry the correlation matrix is no longer circulant and therefore the eigenvectors are not the Fourier modes. Still, similar analysis and mean-field equations control the stability.

For a general curve $f : [0, 2\pi) \to \mathbb{R}^2$ and $\boldsymbol{z}(\psi) = \boldsymbol{f}(\psi)$, we follow Section 5.7.3 and write the correlation function in the singular vectors domain:

$$\begin{aligned} c(\psi_m, \psi_0) &= g^2 \int \frac{d\theta}{2\pi} \Pi_{n=1}^{M} \int Da_n \phi(\sum_{n=1}^{M} \sqrt{c_n} a_n(\theta)\eta_n(\psi_m) + A\cos(\theta) f_1(\psi_m) + A\sin(\theta) f_2(\psi_m)) \\ &\times \phi(\sum_{n=1}^{M} \sqrt{c_n} a_n(\theta)\eta_n(\psi_0) + A\cos\theta) \end{aligned}$$

Next, stability is calculated based on Eq.(59), with the only difference that the decoder is spanned by more than two singular vectors. Recall that

$$W_{\mu i}^{out} = \sum_k f_\mu(\psi_m) C_k^{-1} \eta_k(\psi_m) \sum_{m'} \eta_k(\psi_{m'}) \phi_i(\psi_{m'}) \tag{79}$$

So that

$$G_{OL}^{\mu\nu}(s) = \sum_i W_{\mu i}^{out} \phi_i' X_i^\nu(s) = \tag{80}$$

$$\sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{m'=1}^{M} \sum_{i=1}^{N} f_{\mu m} \eta_{mk} C_k^{-1} \eta_{m'k} \phi_i(\psi_{m'}) \phi_i' X_{\nu i}(s) = \tag{81}$$

$$(1+s) \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_i f_{\mu m} \eta_{mk} \alpha_k^\nu(s) = \tag{82}$$

$$(1+s) \sum_{k=1}^{K} \tilde{q}_{\mu k} \alpha_k^\nu(s) \tag{83}$$

where we defined

$$\tilde{q}_{\mu m} = \sum_n f_{\mu n} \eta_{nm} \tag{84}$$

Following the derivation of Section 5.7.3, with

$$X_{i\nu}(s) = \frac{W_{i\nu}^{fb}}{1+s} + \sum_k \sqrt{c_k} (\alpha_k^\nu(s) a_{ki}) \tag{85}$$

we get to the same self-consistent equation for the order parameters $\alpha_n^\nu(s)$ as in Eq.(61), but with

$$\beta_n^{\nu 0} = \prod_{k=1}^{K} \int Da_k Db_k \int \frac{d\psi}{\pi} \int \frac{d\theta}{2\pi} \eta_n(\psi) \phi(\sum_{\mu'} W_{\mu'}^{fb}(\theta) f_{\mu'}(\psi) + \hat{x}^1(\psi)) \phi'(\sum_l W_{\mu'}^{fb}(\theta) f_{\mu'}(\psi_0) + \sum_{k=1}^{K} \sqrt{c_k} a_k) W_\nu^{fb}(\theta) \tag{86}$$

$$\beta^1_{nn'} = \prod_{k=1}^{K} \int Da_k Db_k \int \frac{d\psi}{\pi} \int \frac{d\theta}{2\pi} \eta_n(\psi)\phi(\sum_{\mu'} W^{fb}_{\mu'}(\theta) f_l(\psi) + \hat{x}^1(\psi))\phi'(\sum_{\mu'} W^{fb}_{\mu'}(\theta) f_{\mu'}(\psi_0) + \sum_{k=1}^{K} \sqrt{c_k} a_k)\sqrt{c_n} b_{n'} \tag{87}$$

with ψ_0 denoting the point at the manifold for which the gain is computed. Here, once again, we define the cut-off $K < M$. The coefficients $\beta^0_{nn'}$ depend on the input index ν, while coefficients $\beta^1_{nn'}$ do not. A notable difference from the ring geometry where gain matrix is the same for all points on the manifold, up to rotation transformation, is that in case of a general geometry the gain varies qualitatively with ψ. Destabilization may occur for some values of ψ, while other regions remain stable. Another difference with general geometry is that in case of manifolds with large deformations, taking a cutoff at K=1 is not enough and the anzats of Eq.(24) does not longer hold (Fig.6-7). Indeed, as the decoder is now spanned by a mixture of several singular vectors, different from the two leading singular vectors for the ring geometry, the stability is determined by several of the order parameters $\alpha^{\nu}_n(s)$.

Instead of calculating the RAE gain directly and obtain the closed loop gain, we take here a different approach and close the loop directly on the order parameters $\alpha_n(s)$. Which we now define to be:

$$X_i(s) = \sum_{\nu'} \frac{W^{fb}_{i\nu'}}{1+s} Z^{\nu'}_{OL}(s) + \sum_k \sqrt{c_k}(\alpha_k(s) a_{ki}) \tag{88}$$

Specifically, given our system with input $Z_{OL}(s)$, rather that with *unity* input that was assumed in an open loop system, we have α given by:

$$(1+s)c_n\alpha_n = g^2 \sum_m \beta^1_{nm}\alpha_m + g^2 \sum_{\nu'} \frac{\beta^{\nu' 0}_n}{1+s} Z^{\nu'}_{OL}(s) \tag{89}$$

and the readout $\hat{\boldsymbol{Z}}_{OL}(s)$ can be obtained from (83) as:

$$\hat{Z}^{\mu}_{OL}(s) = (1+s)\sum_{k=1}^{K} q_{\mu k}\alpha_k(s) \tag{90}$$

closing loop is done by requiring $Z_{OL}(s) = \hat{Z}_{OL}$ which turns (83) into:

$$(1+s)c_n\alpha_n = g^2 \sum_k \beta^1_{nk}\alpha_k + g^2 \sum_{k=1}^{K}\sum_{\nu'=1}^{2} \frac{\beta^{0,\nu'}_n}{1+s} \tilde{q}^{\nu'}_k (1+s)\alpha_k(s) \tag{91}$$

this can be re-written in a form:

$$s\boldsymbol{\alpha} = \boldsymbol{\mathcal{H}}\boldsymbol{\alpha} \tag{92}$$

where the $K \times K$ mean-field stability matrix is:

$$\mathcal{H}_{nm} = g^2 c_n^{-1}\left(\sum_{\nu'=1}^{2} \beta^{0,\nu'}_n \tilde{q}_{\nu',m} + \beta^1_{nm}\right) - \delta_{n,m} \tag{93}$$

K eigenvalues of the mean-field stability matrices $\boldsymbol{\mathcal{H}}$ corresponds to the change in $2K$ eigenvalues of the spectrum of the stability matrix $\boldsymbol{H}$ (Eq.(13)) as a result of adding the trained structured matrix $\boldsymbol{W}$ to the random matrix $g\boldsymbol{J}$.

In points with reflectional symmetry in the target function $\boldsymbol{f}$, further simplification is possible. The matrix $\boldsymbol{\mathcal{H}}$ in such cases is decomposable into blocks of odd and even principal components with no interaction in between them. The block of even principal components $\boldsymbol{\mathcal{H}}^{\parallel}$ will describe the stability in the normal (amplitude) direction while the block of odd modes $\boldsymbol{\mathcal{H}}^{\perp}$ will account for the tangential direction. Obviously this is the case for perfect ring geometry, for which such an even-odd dissection presented in Eq. (60) and the even and odd modes are given by $\cos(k\psi)$ and $\sin(k\psi)$ respectively. Furthermore, this is also the case for the minima and maxima of amplitude in the manifolds that were analysed here (Fig. 7).

5.8.1 Transfer function

To simplify the MF calculations we choose the error function. While this function is very similar to the commonly used tanh, in the case of integrations with Gaussian integrals it simplifies calculations and gives an analytical expression for quantities such as the correlation function. Specifically, we use:

$$\phi(x) = erf\left(\frac{x}{\sqrt{2}}\right) \tag{94}$$

where the error function is defined as $erf(y) = \frac{2}{\sqrt{\pi}}\int_0^y e^{-t^2}dt$. We then use the following identity:

$$\int Dz\phi(a+bz) = \phi(\frac{a}{\sqrt{1+b^2}})$$

in Eq.(45) and then we only need to numerically integrate over θ and y.

5.9 Symmetric-connectome ring model- revisit in RAE setting

For completeness we apply the RAE analysis for the classical, symmetric connectome ring attractor model [Ben-Yishai et al., 1995, Mastrogiuseppe and Ostojic, 2018]. This case corresponds in our framework to taking $\boldsymbol{W}_{out}^T\boldsymbol{W}_{fb} \propto \boldsymbol{I}$ and $g = 0$. Thus, when we unfold the structured component of the connectivity the RAE becomes a simple feedforward autoencoder. Furthermore, the strength of the structured recurrent loop is determined by $w = \mathrm{Tr}\ \boldsymbol{W}_{fb}^T\boldsymbol{W}_{out} = |\boldsymbol{W}_{\boldsymbol{fb}}^{\boldsymbol{l}}|_2|\boldsymbol{W}_{\boldsymbol{out}}^{\boldsymbol{l}}|_2$ with $l = 1, 2$ and it equals $w = J_2$ in the original model of [Ben-Yishai et al., 1995].

Following [Ben-Yishai et al., 1995], one example for such a choice is $W_{1i}^{fb} = \cos(\theta_i)$ and $W_{2i}^{fb} = \sin(\theta_i)$, with $\theta_i = 2\pi i/N$ (see [Mastrogiuseppe and Ostojic, 2018] for a different choice of orthogonal vectors). The decoder is then normalized such that $W_{1i}^{out} = \frac{2w}{N}\cos(\theta_i)$,$W_{2i}^{out} = \frac{2w}{N}\sin(\theta_i)$, yielding the cosine connectivity profile: $W_{ij} = \frac{2w}{N}\cos(\theta_i - \theta_j)$. The parameter w thus controls the amplitude of the

bump through the self-consistent equation:

$$m_1 = \frac{1}{N}\sum_i \cos(\theta_i)\phi(wm_1\cos(\theta_i)) \tag{95}$$

To follow the recurrent autoencoder framework, the representation in 2D plane of the RAE needs to be obtained. Specifically, we need to compute A of Eq.(11). On the one hand, since every point on the manifold attractor is a fixed point, we get that for $\psi = 0$ (as well as for any other $0 \leq \psi < 2\pi$)) $z_1(0) = \boldsymbol{W}^1_{out}\phi(\boldsymbol{x}(\psi = 0)) = \frac{w}{N}\sum_j \cos\theta_j \phi(W^{1j}_{fb} z_1(0)) = \frac{w}{N}\sum_j \cos\theta_j \phi(cos\theta_j z_1(0))$, and together with Eq.(95) yields $z_1(0) = wm_1$. On the other hand, according to notation of Eq.(11) $z_1(\psi = 0) = A$. Consequently:

$$A = wm_1 \tag{96}$$

We are now set to evaluate local dynamics and, consequently, the drift velocity: It follows from (17), that for unity input in the $\hat{\psi}$ direction $Z(s) = (0; 1)$ linearized state is given by

$$X_i = \frac{W^{fb}_{2i}}{1 + \tau s} \tag{97}$$

and the transnational gain is hence of a form:

$$G^{\parallel}_{OL}(\omega) = \sum_i W^{out}_{2i}\phi'(x_i)X_i = \frac{\sum_i W^{out}_{2i}W^{fb}_{2i}\phi'(AW^{fb}_{1i})}{1 + \tau s} \tag{98}$$

Approximating sum by integral, and integrating by parts we obtain:

$$\sum_i W^{out}_{2i}W^{fb}_{2i}\phi'(AW^{fb}_{1i}) = \int \frac{d\theta}{2\pi} w\sin^2\theta\phi'(A\cos\theta) = \int \frac{d\theta}{2\pi} wA^{-1}\phi(A\cos\theta)\cos\theta = A^{-1}wm_1 = 1 \tag{99}$$

that is, we recovered the general formula (24): albeit with $\tau_{eff} = \tau$.

After obtaining the gain of the autodecoder, we continue with the response of the symmetric-connectome network to external inputs. We assume that $\boldsymbol{W}_{in} = \boldsymbol{W}_{fb}$ and that input of a strength ϵ is introduced at direction ψ_1 to a system located at $\psi_0 = 0$. We then have $\Delta z \approx \epsilon(\cos(\psi_1)\sin(\psi_1))$. In case of $\psi_0 = 0$ we have $z_0 = (A, 0)$ and the tangential RE is $\Delta\psi \equiv \Delta = \frac{\epsilon}{A}\sin\psi_1$. By the virtue of (29) the rotation speed is:

$$\dot{\psi} = \frac{\epsilon}{\tau w m_1}\sin\psi_1 \tag{100}$$

Analysing the gain in the amplitude direction yields:

$$G^{\perp}_{OL}(\omega) = \frac{a}{1 + i\tau\omega} \tag{101}$$

with $a = w\langle\phi'(x)\rangle - 1$ and $\lambda = \tau^{-1}(a - 1) = \tau^{-1}(w\langle\phi'(x)\rangle - 2)$.

To conclude, in the classical symmetric-connectome ring attractor model the linearized dynamics around the manifold is given by a simple, first order, ODE.

5.10 Adding noise to unlearned ring

Given structured connectivity J_2 the amplitude m_2 can be computed self-consistently as:

$$m_2 = \int \frac{d\theta}{2\pi} \cos\theta\theta\phi(A\cos\theta) \tag{102}$$

with $A = J_2 m_2$. If synaptic noise is added then along with m_2, the "membrane potential" noise σ^2 needs to be found self consistently:

$$m_2 = \int \tfrac{d\theta}{2\pi} \cos\theta\phi(A\cos\theta + \sigma z) \tag{103}$$

$$\sigma^2 = g^2 \int Dz \int \tfrac{d\theta}{2\pi} \phi^2(A\cos\theta + \sigma z) \tag{104}$$

to estimate the mean squared drift velocity we can use equation (29) and note that in case of small g we have $\tau_{eff} \approx \tau$. It remains to estimate the mean squared error. This can be done by explicitly computing variance of readout:

$$\langle (z_{ol}^{sin})^2 \rangle = \frac{J_2^2}{N} \int \frac{d\theta}{2\pi} \sin^2\theta \left(\int \phi^2(A\cos\theta + \sigma z) Dz - \left(\int \phi(A\cos\theta + \sigma z) Dz \right)^2 \right) \tag{105}$$

In case of small g and consequently small σ we can linearize the above variance and use:

$$\langle (z_{ol}^{sin})^2 \rangle \approx \frac{J_2^2 \sigma^2}{N} \int \frac{d\theta}{2\pi} \sin^2\theta \phi'^2(A\cos\theta) \tag{106}$$

to obtain Δ we normalize error z by amplitude A and get:

$$\langle \Delta^2 \rangle = m_2^{-1} \langle (z_{ol}^{sin})^2 \rangle \approx m_2^{-1} N^{-1} \sigma^2 \int \frac{d\theta}{2\pi} \sin^2\theta \phi'^2(A\cos\theta) \tag{107}$$

which can be compared to [Itskov et al., 2011].