Canonical recurrent neural circuits: A unified sampling machine for static and dynamic inference

Eryn Sale[1,2]
eryn.sale@utsouthwestern.edu

Wen-Hao Zhang[1,2*]
wenhao.zhang@utsouthwestern.edu

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center
[2]O'Donnell Brain Institute, UT Southwestern Medical Center

Abstract

The brain lives in an ever-changing world and needs to infer the dynamic evolution of latent states from noisy sensory inputs. Exploring how canonical recurrent neural circuits in the brain realize dynamic inference is a fundamental question in neuroscience. Nearly all existing studies on dynamic inference focus on deterministic algorithms, whereas cortical circuits are intrinsically stochastic, with accumulating evidence suggesting that they employ stochastic Bayesian sampling algorithms. Nevertheless, nearly all circuit sampling studies focused on static inference with fixed posterior over time instead of dynamic inference, leaving a gap between circuit sampling and dynamic inference. To bridge this gap, we study the sampling-based dynamic inference in a canonical recurrent circuit model with excitatory (E) neurons and two types of inhibitory interneurons: parvalbumin (PV) and somatostatin (SOM) neurons. We find that the canonical circuit unifies Langevin and Hamiltonian sampling to infer either static or dynamic latent states with various moving speeds. Remarkably, switching sampling algorithms and adjusting model's internal latent moving speed can be realized by modulating the gain of SOM neurons without changing synaptic weights. Moreover, when the circuit employs Hamiltonian sampling, its sampling trajectories oscillate around the true latent moving state, resembling the decoded spatial trajectories from hippocampal theta sequences. Our work provides overarching connections between the canonical circuit with diverse interneurons and sampling-based dynamic inference, deepening our understanding of the circuit implementation of Bayesian sampling.

1 Introduction

The brain is bombarded with a continuous stream of sensory inputs that convey noisy and ambiguous information about the external world. Since the world is dynamic and can change instantaneously, the brain must seamlessly track the dynamic evolution of the world. In statistics, inferring changing latent states is formulated as the inference of a hidden Markov model (HMM), called *dynamic inference*. How the canonical recurrent neural circuits in the brain infers dynamic latent states from noisy observations is a fundamental question in neuroscience [1]. Previous studies have investigated how neural circuits implement dynamic inference through *deterministic* algorithms [2–10], where the circuit is either deterministic or its internal noise is non-essential in their computational theory.

Accumulating evidence, however, suggests that the neural circuits in the brain employ *stochastic* sampling-based algorithms to perform inference [11–19]. The stochastic nature of sampling closely matches the stochastic neural circuit dynamics, characterized by the large, structured neuronal response variability [15–17, 20–22]. Despite this, nearly all existing studies on neural circuit

*Corresponding author.

Preprint.

sampling focused on *static inference* [11, 13–17, 19], where the latent state posterior is static over time. There is a gap in our understanding about neural circuit sampling algorithms and dynamic inference. Therefore, we seek to unify neural circuit sampling and dynamic inference by investigating how the canonical circuit can implement sampling-based dynamic inference. Moreover, since the brain performs both static and dynamic inference contingent on the performed task, we further investigate how the same recurrent neural circuit can flexibly implement and switch between both inference modes without modifying its synaptic weights.

In the brain, the canonical cortical circuit is a fundamental computational building block repeated throughout the cerebral cortex which consists of excitatory (E) neurons and diverse classes of inhibitory (I) interneurons [23–27], including parvalbumin (PV), somatostatin (SOM), and vasoactive intestinal peptide (VIP) neurons. Different interneuron classes have unique intrinsic electrical properties and form specific connectivity patterns (Fig. 1A). The rich classes of interneurons not only keep the stability of the circuit, but may also modulate the computations in E neurons, e.g., switching the circuit's sampling algorithms [19]. It is reasonable to hypothesize that the diverse interneurons may enable neural circuits to employ sampling algorithms to implement both dynamic and static inference and flexibly switch between them.

To investigate how the same canonical recurrent circuit can realize both static and dynamic inference, we focus on a canonical recurrent neural circuit model with E neurons and two types of interneurons: PV and SOM neurons. The E neurons are modeled as a rate-based continuous attractor network (CAN), a recurrent circuit model widely-used in neuroscience to explain the processing of continuous stimulus features [28–31]. PV neurons provide divisive, unstructured global inhibition to maintain the circuit's stability, while SOM neurons contribute subtractive, structured local inhibition to E neurons.

After theoretical analyses of the nonlinear circuit dynamics, we find the circuit implements a mixture of Langevin and Hamiltonian *sequential* sampling to implement both dynamic and static inference of the latent stimulus. We analytically identify how the internal model of the latent stimulus transition is stored in the circuits (Fig. 2D,G, Fig. 3A). Remarkably, once circuit weights are set (but not fine tuned) as one of optimal configurations, the same circuit with fixed weights can flexibly sample both dynamic and static stimulus posteriors under different latent transition probabilities by only changing the **gain** of SOM neurons (Fig. 3G). Specifically, we find that the SOM's gain is composed of two parts, each serving a specific role: one is a speed dependent gain whose increment enlarges the moving speed of latent stimulus in circuit's internal model, the other is an algorithmic "switching" gain that determines the proportion of blending Langevin and Hamiltonian sampling. Moreover, once the circuit can do dynamic inference, it is automatically *backwards compatible* with simpler static sampling, without adjusting its weights. This non-trivial property arises because the parameter space for dynamic inference is a subset of static inference. When increasing the Hamiltonian sequential sampling component by increasing SOM's gain and E weights, the circuit's sampling trajectory will oscillate around the true latent (moving) stimulus (Fig. 3F), resembling the decoded spatial trajectories from hippocampal theta sequences [32], further supporting the biological plausibility of the sequential sampling algorithms employed in the circuit.

Significance. This study reveals for the first time how the canonical circuit can flexibly implement sequential sampling for both static and dynamic inference via the gain modulation of SOM neurons, unifying the sampling algorithm, dynamic inference, and its own heterogeneous structure.

2 The canonical cortical circuit model

We consider a canonical circuit model, regarded as a hypercolumn of neurons in the cortex (Fig. 1A). The circuit consists of E neurons, two classes of interneurons (PV and SOM), and is modeled based on a continuous attractor network (CAN); a widely used circuit model to explain continuous stimulus processing in the brain [16, 28, 33–36]. The circuit model reproduces neuronal tunings and the modulation of E neurons' tunings by PV and SOM (Fig. S1B in Supp. Info.).

E neurons. Each E neuron is selective for a 1D stimulus feature z: the orientation or location. θ_j denotes the preferred feature of the j-th E neuron. The preferred feature of all ,N_E, E neurons, $\{\theta_j\}_{j=1}^{N_E}$, uniformly covers the whole space of z (Fig. 1B). Mathematically, in the continuous limit ($\theta_j \to \theta$) corresponding to an infinite number of neurons, the E dynamics is [34, 37],

$$\tau \dot{\mathbf{u}}_E(\theta, t) = -\mathbf{u}_E(\theta, t) + \rho \sum\nolimits_{X \in \{E,F,S\}} (\mathbf{W}_{EX} * \mathbf{r}_X)(\theta, t) + \sqrt{\tau \mathsf{F}[\mathbf{u}_E(\theta, t)]_+} \xi(\theta, t), \quad (1)$$

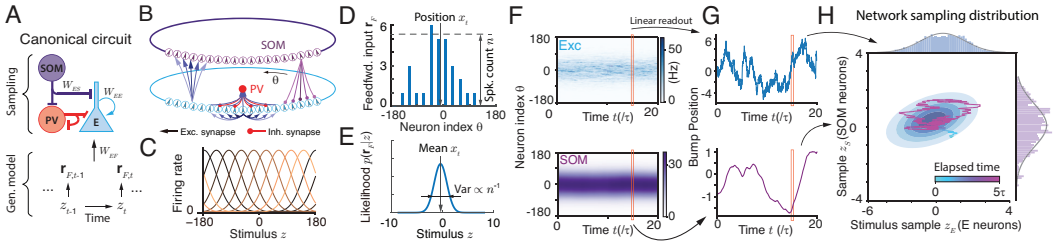

Figure 1: The canonical recurrent circuit model and Bayesian sampling. (A) The canonical neural circuit with E and three types of I neurons receives stochastic feedforward inputs evoked by changing latent stimulus. (B) The recurrent circuit model considered in the present study consists of E and two types of interneurons (PV and SOM). (C) Tuning curves of E neurons. (D-E) A schematic of spiking feedforward input received by E neurons (D) that conveys the whole stimulus likelihood (E). Position and spike count of feedforward input determine the likelihood mean and variance respectively. (F-G) E and SOM population bump responses (F) and instantaneous bump positions (G) regarded as the stimulus sample, z_E, and auxiliary variable, z_S, respectively. (H) Hamiltonian sampling in the circuit for static stimulus.

where $\mathbf{u}_E(\theta, t)$ and $\mathbf{r}_E(\theta, t)$ represent the synaptic inputs and firing rates of neurons preferring $z = \theta$, respectively. X denotes neuronal types with E, F and S standing for E neurons, sensory feedforward inputs, and the SOM neurons, respectively. τ is the time constant, $\rho = N/2\pi$ is the neuronal density covering the stimulus feature space, and $[x]_+ = \max(x, 0)$ is negative rectification. E neurons receive internal Poisson variability with Fano factor F, mimicking stochastic spike generation. The symbol $*$ denotes the convolution $\mathbf{W}(\theta) * \mathbf{r}(\theta) = \int \mathbf{W}(\theta - \theta')\mathbf{r}(\theta')d\theta'$.

Recurrent weight kernel. $\mathbf{W}_{YX}(\theta)$ is the recurrent connection kernel from neurons of type X to those of type Y, modeled as Gaussian functions (Fig. S1A),

$$\mathbf{W}_{YX}(\theta) = w_{YX}\left(\sqrt{2\pi}a_{XY}\right)^{-1}\exp(-\theta^2/2a_{XY}^2), \tag{2}$$

where w_{YX} controls the peak strength and a_{XY} is the connection width from neuron type X to Y.

PV neurons. PV neuron firing is driven by E neurons. For simplicity, they are not selective to stimulus, an extreme case of their weak tuning in reality [23, 38, 39]. PV provides divisive, unstructured global inhibition to E neurons to stabilize the circuit. Their effect is modeled as divisive normalization, a canonical operation (activation function) observed in neural circuits [24, 40, 41],

$$\mathbf{r}_E(\theta, t) = [\mathbf{u}_E(\theta, t)]_+^2/(1 + \rho w_{EP} r_P), \quad r_P = \int [\mathbf{u}_E(\theta', t)]_+^2 d\theta', \tag{3}$$

where r_p (scalar) corresponds to the mean firing rate of the population of PV neurons and w_{EP} (a positive scalar) characterizes the global inhibition strength from PV neurons to E neurons.

SOM neurons. In contrast to PV, SOM neurons have tuning and provide subtractive, local structured inhibition to E neurons [39]. The SOM's dynamics is

$$\tau\dot{\mathbf{u}}_S(\theta, t) = -\mathbf{u}_S(\theta, t) + \rho(\mathbf{W}_{SE} * \mathbf{r}_E)(\theta, t); \quad \mathbf{r}_S(\theta, t) = g_S \cdot [\mathbf{u}_S(\theta, t)]_+. \tag{4}$$

where g_S is the "gain" of SOM neurons and can be modulated (see Discussion). For simplicity, we do not consider mutual inhibition between SOM neurons, which negligibly affects our results. Consistent with neuroanatomy, our SOM neurons do not receive direct feedforward input [25, 27], which is necessary to realize Hamiltonian sampling as suggested in a recent study [19].

Stochastic sensory feedforward input. The feedforward input $\mathbf{r}_F(\theta, t)$ (Eq. 1) is stochastically evoked from the latent stimulus z_t (Fig. 1D-E), modeled as conditionally independent Poisson spikes with Gaussian tuning given z_t, the conventional setting in neural coding studies [17, 42],

$$\mathbf{r}_F(\theta, t) \sim \text{Poisson}[\lambda_F(\theta|z_t)], \quad \lambda_F(\theta|z_t) = R_F \exp[-(\theta - z_t)^2/2a^2], \tag{5}$$

where $\lambda_F(\theta|z_t)$ is the mean firing rate of neuron θ given z_t. $\mathbf{r}_F$ is approximated as a continuous Gaussian random variable with multiplicative noise.

3 Neurally plausible sampling algorithms for dynamic and static inference

Performing inference over the latent stimulus feature z in the neural circuit requires us to define a generative model. Only then can the posterior of z_t be defined and used as the computational objective of the circuit. We will start by assuming a generative model and then show how is can be represented in the experiment-supported neural circuit model via mathematical analysis.

3.1 The subjective internal generative model in the canonical circuit

The internal model of latent stimulus transition. We assume the latent stimulus transition stored in the canonical circuit has the following form, similar to previous studies (e.g., [5, 10])

$$p(z_{t+1}|z_t) = \mathcal{N}\big(z_{t+1}|z_t + v\delta t, \Lambda_z^{-1}\delta t\big), \tag{6}$$

where δt is the time bin, v is the moving speed of latent stimulus, and Λ_z is the precision of the transition. When $v = 0$ and $\Lambda_z \to \infty$, the latent stimulus is static over time, and the stimulus inference degenerates into extensively studied static inference (e.g., [15–17, 19, 22]).

Stimulus likelihood $f(z)$. The stochastic feedforward input from the stimulus feature z (Eq. 5) naturally specifies the stimulus likelihood calculated as a Gaussian (see SI. Sec. 2.1),

$$\begin{aligned} &f(z_t) \propto p[\mathbf{r}_F(t)|z_t] = \textstyle\prod_\theta \text{Poisson}[\lambda_F(\theta|z)] \propto \mathcal{N}\big(z_t|x_t, \Lambda_F^{-1}\big), \\ &\text{where } \ x_t = \textstyle\sum_j \mathbf{r}_F(\theta_j)\theta_j / \sum_j \mathbf{r}_F(\theta_j), \ \ \Lambda_F = a^{-2}\sum_j \mathbf{r}_F(\theta_j, t) = \sqrt{2\pi}\rho a^{-1} R_F. \end{aligned} \tag{7}$$

The Gaussian stimulus likelihood results from the Gaussian profile of feedforward input tuning $\lambda_F(\theta|z_t)$ (Eq. 7) [42]. Its mean x_t and precision Λ_F can be read from $\mathbf{r}_F$ via population vector [43, 44], Geometrically, x_t is regarded as $\mathbf{r}_F$'s location in the stimulus feature space, and Λ_F is proportional to the feedforward input strength R_F (Fig. 1D-E). A single snapshot of $\mathbf{r}_F(t)$ parametrically conveys the whole stimulus likelihood $\mathcal{L}(z_t)$ [42].

3.2 Sequential sampling for dynamic inference

The inference of dynamic latent stimuli has been well established in statistics, and the instantaneous posterior $p(z_t|\mathbf{r}_F(1:t)) \equiv \pi_t(z_t)$ given all the history of feedforward inputs $\mathbf{r}_F(1:t)$ up to time t can be iteratively computed as the recursive Bayesian filtering [45],

$$\pi_{t+1}(z_{t+1}) \propto f(z_{t+1}) \int p(z_{t+1}|z_t)\pi_t(z_t)dz_t. \tag{8}$$

We consider the sequential sampling to implement the recursive Bayesian filtering, which can be implemented by the proposed stochastic neural circuit as shown below and has never been explored in previous work. Sequential sampling utilizes stochastic sampling to approximate the integration (marginalization) in Eq. (8), the most challenging computation in Bayesian filtering, which clearly distinguishes our work from the circuit algorithms in previous studies (see Discussion):

$$\begin{aligned} \pi_{t+1}(z_{t+1}) &\propto f(z_{t+1}) \cdot \big[\tfrac{1}{L}\textstyle\sum_{l=1}^{L} p\big(z_{t+1}|\tilde{z}_t^{(l)}\big)\big], \quad \tilde{z}_t^{(l)} \sim \pi_t(z_t), \\ &\approx f(z_{t+1}) \cdot p\left(z_{t+1}|\tilde{z}_t\right), \quad \text{(only draw one sample, } L = 1). \end{aligned} \tag{9}$$

When only one sample is drawn at each time step, there is no need to average over probabilities (Eq. 9). Furthermore, it keeps the posterior $\pi_t(z_t)$ closed as a Gaussian distribution, which, otherwise, will be a Gaussian mixture and needs extra approximations in the circuit as in previous studies [3, 4].

$$\pi_t(z_t) = \mathcal{N}(z_t|\mu_t, \Omega_t^{-1}), \text{ with } \Omega_t = \Lambda_F + \Lambda_z\delta t^{-1}, \ \mu_t = \Omega_t^{-1}[\Lambda_F x_t + \Lambda_z(\tilde{z}_t\delta t^{-1} + v)]. \tag{10}$$

It is worth noting that drawing one sample in each time step is a conceptual way of understanding the sampling process in the discrete dynamics. In the continuous case ($\delta t \to 0$), the number of samples that can be drawn in a single "time step" is related to the relative time scales between the latent transition and sampling dynamics. Hence, this approximation works well when z changes slowly.

3.3 Neurally plausible sequential sampling dynamics

There are multiple ways of generating random samples, $\tilde{z}_t \sim \pi_t(z_t)$ (Eq. 9). One naive technique, $\tilde{z}_t = \mu_t + \Omega_t^{-1/2}\xi_t$, is not the sampling algorithm employed by the circuit with temporal memories.

As will be shown later, we find that the canonical circuit can employ both Langevin and Hamiltonian sequential sampling. To help understand the sampling algorithms, we present the sampling dynamics, before showing how they are implemented in the circuit dynamics.

Langevin sequential sampling. It uses the last sample $\tilde{z}_{t-1}$ as the initial condition to run one step of the Langevin dynamics, performing stochastic gradient on $\pi_t(z)$ [46–48],

$$\tilde{z}_t = \tilde{z}_{t-1} + \left(\tau_L^{-1}\delta t\right)\nabla_z \ln \pi_t(\tilde{z}_{t-1}) + \left(2\tau_L^{-1}\delta t\right)^{1/2}\eta_t, \tag{11}$$

with the equilibrium distribution $\pi_t(z)$. In sequential sampling, it is not necessary to draw many samples to reach equilibrium. Instead, one sample is drawn from $\pi_t(z)$ (running Eq. 11 for one step), and then the next sample $\tilde{z}_{t+1}$ is drawn from the next posterior $\pi_{t+1}(z)$. This corresponds to a non-equilibrium Langevin dynamics. The efficiency of the non-equilibrium sequential sampling depends on the time scale of the sampling, τ_L, and the change of posterior $\pi_t(z)$.

Hamiltonian sequential sampling defines a Hamiltonian function $H(z, p)$ with a momentum p,

$$H_t(z, p) = -\ln \pi_t(z) + K(p), \quad K(p) = m^{-1}p^2/2. \tag{12}$$

$K(p)$ is called kinetic energy with m analogous to the mass in physics. We consider the Hamiltonian sampling dynamics with friction γ that dampen momentum [48, 49],

$$\frac{d}{dt}\begin{bmatrix}\tilde{z}_t \\ p_t\end{bmatrix} = -\begin{bmatrix}0 & -\tau_H \\ \tau_H & \gamma\end{bmatrix}\begin{bmatrix}\nabla_z \ln \pi_t(\tilde{z}_{t-1}) \\ m^{-1}p\end{bmatrix} + \sqrt{2}\begin{bmatrix}0 & 0 \\ 0 & \gamma^{1/2}\end{bmatrix}\boldsymbol{\xi}_t. \tag{13}$$

Adaptive sampling step size. Although the step size (τ_L in Eq. 11) leaves the equilibrium distribution invariant, it is critical in non-equilibrium sequential sampling with Langevin/Hamiltonian dynamics when only a few samples are drawn from each instantaneous posterior $\pi_t(z)$. The step size determines a trade-off between sampling efficiency and accuracy. For example, a small step size slows down sampling, causing the samples to fail to track fast moving latent stimuli. In theory, *Riemann manifold* sampling provides an elegant framework to use $\pi_t(z)$'s curvature, measured by Fisher information $G_t(z)$, to adaptively adjust the step size [47, 50, 51],

$$\tau_L \propto G_t(z); \quad m \propto G_t(z); \quad \text{where } G_t(z) = -\mathbb{E}_{\pi_t(z)}\left[\nabla_z^2 \ln \pi_t(z)\right] = \Omega_t. \tag{14}$$

Later we will show the circuit can adaptively adjust its sampling step size based on the transition speed and precision of the latent stimulus.

4 Theoretical analysis of the circuit dynamics

We performed theoretical analyses of the canonical circuit dynamics to reveal how it implements Langevin and Hamiltonian sequential sampling with adaptive step size for dynamic and static inference. Our theoretical analyses are based on the widely-used theory of continuous attractor networks that provide *analytical* solutions of the nonlinear circuit dynamics [52]. This approach is based on perturbative analysis and solving for the eigenvectors of the relaxation dynamics of the perturbation. The theoretical analyses are briefly described below, with details in SI. Sec. 3.

Attractors. Given a feedforward input $\mathbf{r}_F$ (Eq. 5), it can be shown that the synaptic input $\mathbf{u}_E(\theta)$ and firing rate $\mathbf{r}_E(\theta)$ of E neurons both have Gaussian bump attractor states (Fig. 1F; SI. Sec. 3.1) ,

$$\bar{\mathbf{u}}_E(\theta) = \bar{U}_E \exp[-(\theta - \bar{z}_E)^2/4a^2], \;\; \bar{\mathbf{r}}_E(\theta) = \bar{R}_E \exp[-(\theta - \bar{z}_E)^2/2a^2]. \tag{15}$$

The Gaussian bump is located at $\bar{z}_E = x$ on the stimulus feature manifold, with x is feedforward input's location (Eq. 7). $\bar{U}_E$ and $\bar{R}_E$ are the "height" of the population synaptic input and firing rate respectively, whose analytical solutions are in (SI. Sec. 3.2). Similar attractor states exist for SOM neurons' dynamics(details at SI. Sec. 3.1).

Perturbative analysis. We perturb the circuit dynamics from its attractors, $\mathbf{u}_X(\theta, t) = \bar{\mathbf{u}}_X(\theta) + \delta\mathbf{u}_X(\theta, t)$ ($X = \{E, S\}$), derive the relaxation dynamics of the perturbation $\delta\mathbf{u}_X(\theta, t)$, and analytically calculate the eigenvectors of the perturbation dynamics (SI. Sec. 3.3, [52]). The eigenvector with the largest eigenvalue corresponds to the change on the stimulus feature manifold z_X (bump position), [35, 52, 53], i.e., $\phi_1(\theta|z_X) \propto \nabla_z \bar{\mathbf{u}}_X(\theta) \propto (\theta - z_X)\exp[-(\theta - z_X)^2/4a^2]$.

Dimensionality reduction for stimulus sampling dynamics. We project the E and SOM dynamics (Eqs. 1,4) onto their respective eigenvectors about the stimulus feature manifold, $\langle \phi_1(\theta), g(\theta)\rangle =$

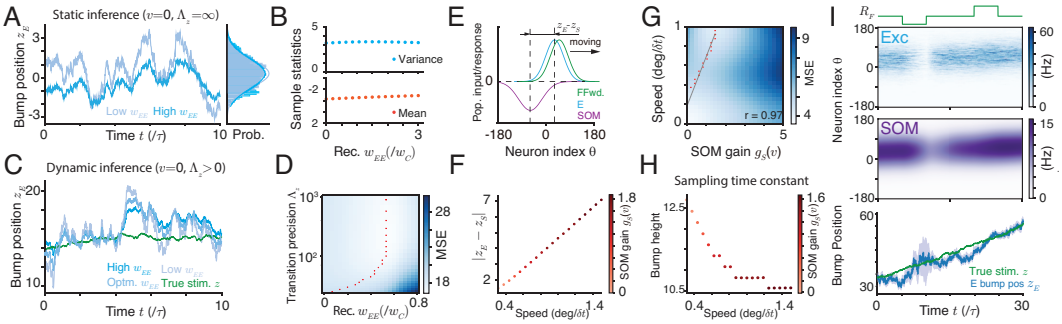

Figure 2: Langevin sampling in the circuit for both static and dynamic inference. (A-B) The recurrent E weights don't change the equilibrium sampling distribution in static inference while affecting the temporal sampling dynamics. (C) The recurrent weight determines a trade-off between exploitation and exploration, and is critical for non-equilibrium sequential sampling of dynamic inference. (D) The optimal recurrent weight for dynamic inference increases with the latent transition precision, reflected as the mean square error between the E bump position and true latent stimulus (E) The internal speed generation in the circuit comes from the spatial offset of population responses between SOM and E neurons. (F) The internal speed increases with the mean separation between E and SOM bump positions that is controlled by SOM gain. (G) Higher speed of the latent stimulus requires higher SOM gain in the circuit. (H) The bump height monotonically decreases with the speed as a result of increased gain, yielding a larger sampling step size. (I) The circuit with fixed weights can flexibly sample latent moving stimulus with varying input intensity that controls likelihood precision.

$\int \phi_1(\theta) g(\theta) d\theta$, with $g(\theta)$ representing the left or right hand side of Eqs. (1, 4). It yields the sampling dynamics of the E and SOM neurons on the stimulus feature manifold (see details in SI. Sec. 3.3)

$$\text{E neurons:} \quad \dot{z}_E \approx \tau_E^{-1}[U_{EF}(x_t - z_E) + U_{ES}(z_S - z_E)] + \sigma_E \tau_E^{-1/2} \xi_t, \tag{16a}$$

$$\text{SOM:} \quad \dot{z}_S \approx (\tau_S)^{-1}[U_{SE}(z_E - z_S)] \tag{16b}$$

where σ_E is a constant invariant with network activities. The sampling time constant $\tau_X = \tau U_X$ is proportional to the bump height U_X ($X \in \{E, S\}$), which is solved as (SI. 3.2),

$$(a).\ U_E \approx U_{EE} + U_{EF} + U_{ES}, \quad (b).\ U_S \approx U_{SE}, \quad \text{with } U_{XY} = \rho w_{XY} R_Y / \sqrt{2} \tag{17}$$

where U_{XY} denotes the population input magnitude from neuron type Y to X (see Sec. 3.3).

5 Langevin sequential sampling for dynamic and static inference in the circuit

We investigated how the circuit implements Langevin sequential sampling with adaptive step size by aligning the projected dynamics (Eq. 16a) with the Langevin dynamics (Eq. 11). For ease of understanding, we started from a simplified case where the latent stimulus z_t had zero speed ($v = 0$, Eq. 6), but with transition noise ($\Lambda_z \neq 0$). We then extended our solution to the full case with $v \neq 0$.

5.1 Sampling for latent stimuli with Brownian motion ($v = 0$, $\Lambda_z \neq 0$)

We compared Langevin sampling (substituting $\pi_t(z)$ with $v = 0$ (Eq. 10) and step size (Eq. 14) into Eq. (11)) with the circuit sampling dynamics without SOM inhibition ($U_{ES} = 0$, Eq. 16a).

	Sampling dynamics	Sampling time constant
E neurons	$\dot{z}_E \approx \tau_E^{-1}[U_{EF}(x_t - z_E)] + \sigma_E \tau_E^{-1/2} \xi_t$	$\tau_E = \tau U_E = \tau(U_{EE} + U_{EF})$
Langevin	$\dot{z}_t = \tau_L^{-1}[\Lambda_F(x_t - z)] + \left(2\tau_L^{-1}\right)^{1/2} \eta_t$	$\tau_L \propto \Omega_t = \Lambda_z \delta t^{-1} + \Lambda_F$

Conditions for Langevin sequential sampling in the circuit. The coefficients in the z_E dynamics should be consistent with the Langevin dynamics, explored below.

1) No SOM inhibition ($g_S = 0$). The Langevin sampling only has one drift term of the likelihood gradient conveyed by the feedforward input in the circuit, $\Lambda_F(x_t - z)$ (Eq. 7). Hence there is no need for SOM inhibition, which can be realized by switching off the SOM gain $g_S = 0$ (Eq. 4).
2) w_{EF} can be adjusted to sample instantaneous posteriors $\pi_t(z)$. w_{EF} determines the feedforward input strength, $U_{EF} \propto w_{EF} R_F$, and affects the sampling variance.

3) w_{EE} is constrained by the adaptive sampling time constant τ_E. τ_E is critical for non-equilibrium sequential sampling. Since $U_{EF} \propto w_{EF} R_F \propto w_{EF} \Lambda_F$ (Eqs. 7 and 17), the w_{EE} can be set to make $U_{EE} \propto \Lambda_z$ and hence $\tau_E \propto \tau_L$.

Combined, the circuit parameters for Langevin sequential sampling are (see SI. Sec. 4.1.1),

$$(a).\ w_{EF} = \sqrt{\pi}\sigma_E^2/a = (2\sqrt{3})^3\mathsf{F};\ \ (b).\ w_{EE} = a w_{EF}(\sqrt{2\pi}\rho\delta t R_E)^{-1}\Lambda_z;\ \ (c).\ g_S = 0. \quad (18)$$

Once w_{EF} is set to align the sampling variance with the posterior (σ_E is a constant value), the circuit can sample the $\pi_t(z)$ of different input intensities without changing synaptic weights (Fig. 2I); however, with greater uncertainty for lower feedforward input. Fig. 2C shows the bump position trajectory compared with optimal or suboptimal recurrent E weight w_{EE}, suggesting the recurrent E weight is critical in determining the sampling step size for non-equilibrium sequential sampling. Each transition precision has an optimal recurrent E weight to minimize the mean square error between the circuit's sample trajectory and the true latent stimulus, and the recurrent weight w_{EE} increases with the transition precision Λ_z (Fig. 2D), confirming our theory (Eq. 18b).

5.2 Backwards compatibility for static sampling

Remarkably, the circuit with parameters set for sequential sampling of latent stimuli with latent transition noise ($v = 0$ and $\Lambda_z^{-1} \neq 0$, Eq. 18) can automatically sample static posteriors without readjusting any circuit parameters (Fig. 2A). Here, the static posterior ($v = \Lambda_z^{-1} = 0$) is a special case of dynamic inference, and it degenerates into the likelihood, $\pi_t(z) = \mathcal{N}(z|x, \Lambda_F^{-1})$ (Eq. 7). The automatic backwards compatibility arises since the sampling time constant τ_E, controlled by recurrent E weight w_{EE}, is a *free parameter* for static sampling (Eq. 16a), i.e., the equilibrium distribution doesn't depends on w_{EE} (Fig. 2B), despite its critical role in sequential sampling (Eq. 18b, Fig. 2D). In equilibrium static sampling, the convergence time is not an issue and the sampling dynamics can run until equilibrium.

5.3 Circuit sampling of latent stimuli with various speeds by modulating SOM gain

With non-zero latent transition speed v, the Langevin sampling has an extra speed v-drift term , $\dot{z}_t = \tau_L^{-1}[\Lambda_F(x_t - z) + \Lambda_z v] + (2\tau_L^{-1})^{1/2}\eta_t$. Since $\Lambda_F(x_t - z)$ comes from the feedforward input (Eq. 7), it is straightforward to reason that the inhibitory feedback $U_{ES}(z_S - z_E)$ in the circuit (Eq. 16a) provides the speed-drift term $\Lambda_z v$. We next analyze this in the circuit dynamics.

1) To infer a moving latent stimulus with speed v by the circuit, the E neurons' bump position z_E should move with the same average speed: $\langle \dot{z}_E \rangle = \langle \dot{x}_t \rangle = v$ where $\langle \cdot \rangle$ denoting the average across trials. A similar condition holds for SOM neurons: $\langle \dot{z}_S \rangle = v$ (Eq. 16b).
2) To minimize systematic bias, the average separation between z_E and the input feature x_t should be zero: $\langle x_t - z_E \rangle = 0$, otherwise $\langle z_E(t) \rangle$ is offset from the true latent z_t.

All of this implies that the circuit should *internally* generate a moving neural sequence with the same speed as the latent stimulus, rather than being passively driven by the input feature x_t. To study the internal speed generation mechanism, we decompose $z_E = \langle z_E \rangle + \delta z_E$, where $\langle z_E \rangle$ is the mean of Eq. (16a) capturing the internally generated movement while the residue δz_E captures sampling fluctuation. Substituting the decomposition back to Eq. (16a) yields the dynamics of $\langle z_E \rangle$ and δz_E,

$$(a).\ \langle \dot{z}_E \rangle = \tau_E^{-1} U_{ES} \langle z_S - z_E \rangle = v;\ \ (b).\ \dot{\delta z}_E \approx \tau_E^{-1} U_{EF}(\delta x_t - \delta z_E) + \sigma_E \tau_E^{-1/2} \eta_t, \quad (19)$$

Interestingly, the dynamics of residue δz_E correspond to sampling a latent stimulus with zero speed as presented in Sec. 5.1, in that the $\langle \delta x_t \rangle = \langle x_t - \langle x_t \rangle \rangle = 0$.

Internal speed modulates SOM gain g_S. Combining Eqs. (19a and 17), the circuit's internal speed v is proportional to E and SOM's bump position separation $\langle z_E - z_s \rangle$ (Fig. 2E-F; see SI. Sec. 4.1.2)

$$v = \tau^{-1}\langle z_E - z_s \rangle \approx \left(4a^2\tau^{-2}\ln g_s + \text{const.}\right)^{1/2} \quad (20)$$

$\langle z_E - z_s \rangle$, however, is a circuit response rather than an adjustable parameter, so we seek a circuit parameter to modulate the internal speed v. We are interested in changing the internal speed by modulating the SOM's gain g_S (Eq. 4), rather than changing synaptic weights, although both can modulate speed. This is because the neural gain can be quickly modulated at cognitively relevant time scales (tens to hundreds of milliseconds). In contrast, changing synaptic weights requires synaptic plasticity, which is too slow to coordinate with fast cognitive computation.

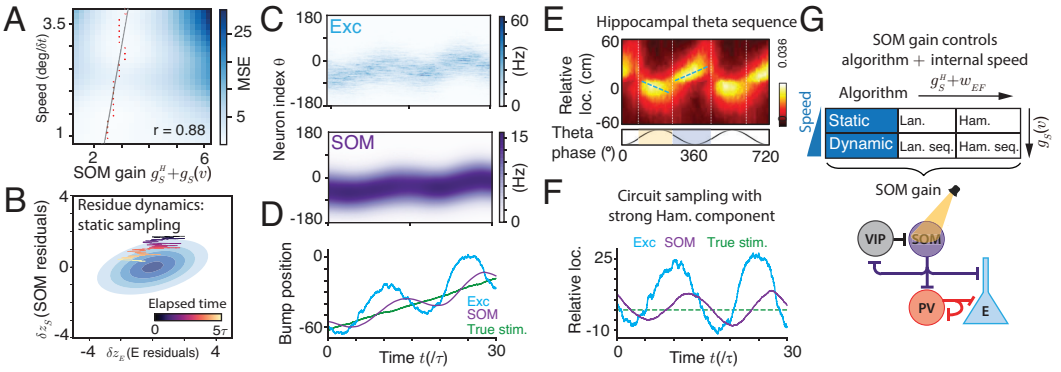

Figure 3: Hamiltonian sequential sampling in the circuit. (A) Additional SOM gain and feedforward weight are needed to upgrade the Langevin sequential sampling into Hamiltonian. (B) The trajectory of the residual dynamics of E and SOM samples resembles sampling a latent stimulus with zero speed. (C-D) Hamiltonian sequential sampling induces oscillations in the E and SOM samples (bump positions) around the true moving latent stimulus. (C): spatiotemporal responses; (D): decoded sampling trajectories. (E-F) The Hamiltonian sequence sampling is characterized by relative oscillations of samples around the latent moving stimulus (E), and resembles the spatial trajectories decoded during hippocampal theta sequences (F) when a rodent is actively exploring. (F) is adapted from [32]. (G) The SOM gain changes the internal speed of latent stimulus, and switches the sequential sampling algorithms with adjusted with feedforward weights together.

Considering the weak limit of feedforward input intensity R_F, solving Eqs. (16a - 17) obtains a neat relation between v and the SOM gain g_S (Eq. 20, details in SI. Sec. 4.1.2). The internal speed increases with the SOM gain g_S, a fact confirmed by numerical results (Fig. 2G). In practice, g_S may be modulated by VIP neurons (not included in the present circuit model), conveying the self-motion speed to modulate the internal stimulus speeds (see Discussion).

Conditions of Langevin sequential sampling for nonzero speed. Based on the condition for implementing circuit Langevin sampling for a zero speed (Eq. 18), we only need to adjust the SOM gain g_S to make the circuit's internal speed equal to the latent speed (Fig. 2G).

Faster sampling associated with faster speed. The circuit has an additional desirable feature, in that it automatically uses a larger sampling step size to sample a faster latent stimulus. This is because the sampling time constant $\tau_E = \tau U_E$ (inversely proportional to the step size) decreases with the SOM gain g_S, (Fig. 2H), a feature favored for higher internal speed v. By contrast, the speed-dependent sampling step size is not present in the Riemann manifold sampling. The Fisher information of the instantaneous posterior Ω_t does not rely on latent stimulus speed (Eq. 10).

6 Hamiltonian sequential sampling in the circuit

We further investigate how the circuit implements Hamiltonian sequential sampling (Eq. 13) with SOM gain modulation. Inspired by Langevin sequential sampling in the circuit, we know that inferring a moving latent stimulus requires a deterministic internal speed, realized by the speed-dependent SOM gain g_S (Eq. 20, Fig. 2G). We also know that the residue dynamics δz_E correspond to Langevin sampling of a latent stimulus with zero speed (Eq. 19b). Intuitively, upgrading the Langevin to Hamiltonian sampling corresponds to upgrading the Langevin dynamics of δz_E (Eq. 19) into Hamiltonian dynamics (Eq. 13). As will be shown, Hamiltonian sequential sampling can be easily realized by increasing **the SOM gain** g_S and **feedforward weight** w_{EF} based on the circuit weights for Langevin sequential sampling in the last section.

We can decompose the bump position dynamics in a similar way, with $z_X = \langle z_X \rangle + \delta z_X$ ($X \in \{E, S\}$) with $\langle z_X \rangle$ capturing deterministic speed. We obtain the residue dynamics of δz_E and δz_S [2],

$$\dot{\delta z}_E = \tau_E^{-1}[U_{EF}(\delta x_t - \delta z_E) + U_{ES}(\delta z_S - \delta z_E)] + \sigma_E \tau_E^{-1/2} \eta_t \tag{21a}$$

$$\dot{\delta z}_S = \tau_S^{-1} U_{SE}(\delta z_E - \delta z_S), \tag{21b}$$

[2]The $U_{ES}(\delta z_S - \delta z_E)$ is negligible in Langevin sampling (Eq. 19) but is essential for Hamiltonian sampling.

To understand the Hamiltonian sampling dynamics embedded in the circuit, we further decompose the E neurons' residue dynamics (Eq. 16a) into two parts: Langevin and Hamiltonian,

$$\tau_E \dot{\delta z}_E = \underbrace{[U_{ES}^H(\delta z_S - \delta z_E) + U_{EF}^H(\delta x_t - \delta z_E)]}_{\text{Momentum } p, \text{ (Hamiltonian)}} + \underbrace{[U_{EF}^L(\delta x_t - \delta z_E) + \sigma_E \sqrt{\tau_E}\xi_t]}_{\text{Langevin}},$$

U_{EF}^L is the feedforward input magnitude in Langevin sequential sampling (the same as in Eq. 19). U_{ES}^H is the extra SOM inhibition to the E neurons, overlaid with the U_{ES} in $\langle z_E \rangle$ dynamics for internal speed generation in Eq. (19). Transforming the above $(\delta z_E, \delta z_S)$ into the $(\delta z_E, p)$ dynamics and organizing it in the form of Hamiltonian sampling (Eq. 13) (see SI. Sec. 4.2),

$$\begin{bmatrix} \dot{\delta z}_E \\ \dot{p} \end{bmatrix} = - \begin{bmatrix} U_{EF}^L(\tau_E \Lambda_F)^{-1} & -\beta_E \Lambda_F^{-1} \\ \beta_E \Lambda_F^{-1} & \tau_E \beta_p \beta_E \Lambda_F^{-1} \end{bmatrix} \begin{bmatrix} -\nabla_z \ln \pi_t(z_E) \\ (\tau_E \beta_E)^{-1} \Lambda_F \cdot p \end{bmatrix} + \begin{bmatrix} \sigma_E \tau_E^{-1/2} & 0 \\ 0 & \sigma_p \end{bmatrix} \boldsymbol{\xi}_t \quad (22)$$

where β_p, β_E and σ_p are functions of the coefficients in Eqs. (21a-21b) (details in SI Sec 4.2.1). Eq. (22) shows that the circuit implements mixed Langevin/Hamiltonian sampling rather than pure Hamiltonian sampling, and that the momentum p has a friction term that dampens the momentum.

Conditions for realizing Hamiltonian sampling. Realizing sampling in the circuit requires altering the ratio of the drift and diffusion coefficients (Eq. 22). Comparing Eq. (22) to Eqs. (11 and 13), we can find the requirements for realizing Hamiltonian sampling in the circuit,

$$(a).\ \text{Langevin: } U_{EF}^L(\tau_E \Lambda_F)^{-1} = \sigma_E^2 \tau_E^{-1}/2; \quad (b).\ \text{Hamiltonian friction: } \tau_E \beta_p \beta_E \Lambda_F^{-1} = \sigma_p^2/2,$$

Solving the two equations,

$$(a).\ w_{EF}^L = (2\sqrt{3})^3 \mathsf{F}; \quad (b).\ \big(U_S w_{ES}\big) \cdot g_S^H - R_F \cdot w_{EF}^H = F\big(w_{EF}^H / w_{EF}^L\big) U_E, \quad (23)$$

where $F(x)$ is a monotonically increasing function of x and increases with the 1st order of x (see SI Sec. 4.2.1). Condition (a) is already satisfied in the Langevin sampling (Eq. 18a). This suggests that for circuit Hamiltonian sampling, we only need to enlarge the SOM gain g_S and feedforward weight w_{EF}, and that the increase of the two is approximately linear (Eq. 23, Fig. S1D).

Hamiltonian sampling trajectory resembles hippocampal theta sequences. Hamiltonian sequential sampling naturally introduces oscillations in the sampling trajectory z_E (Fig. 3D). For dynamic inference of a moving latent stimulus, oscillations appear as the samples switch between leading and lagging the true latent stimulus (Fig. 3F). This is comparable to the hippocampus' internal spatial trajectories during theta oscillations when animals are actively exploring the environment [32]. This suggests the biological plausibility of circuit Hamiltonian sequential sampling.

7 Conclusion and Discussion

Our theoretical study claims that our canonical recurrent circuit implements non-equilibrium sequential sampling to infer either a dynamic or static latent stimulus. It reveals for the first time that the circuit flexibly samples a latent stimulus with different speeds and precisions, using either Langevin or Hamiltonian sampling. This flexibility is realized by modulating **SOM neurons' gain** (g_S in Eq. 4) which contains two parts: one of which increases with the latent stimulus speed, and the other which acts as a baseline that changes the mixing proportion of Langevin and Hamiltonian sampling (Fig. 3G). Notably, our model is biologically plausible as the sampling trajectories in the circuit with strong Hamiltonian sampling resemble hippocampal theta sequences (Fig. 3C-F). Our work sheds profound insight into understanding sequential sampling in canonical circuit.

What controls SOM's gain? The SOM receives VIP neurons' inhibition (Fig. 1A, not include in the model) whose activities can be modulated by self-motion signals [54–56], possibly by receiving the efference copy from the motor cortex. When we regard our canonical circuit as a sensory cortex or the hippocampus, the circuit can infer the noisy moving sensory input due to self motion.

Comparison with other work. Previous studies have investigated how recurrent neural circuits infer a dynamic latent stimulus [2–10], but all of them considered deterministic circuit algorithms rather than sampling. It remains unknown which algorithms are used by real neural circuits, however our work suggests that sampling simplifies the circuit implementation compared to deterministic algorithms. Depending on the concrete neural representation strategy, deterministic inference circuits require either approximation by interchanging the order of logarithm and summation [3, 4], or need complicated nonlinear functions to implement the marginalization [5–10], even if the generative

model is linear and exists an exact solution (Kalman filter). The above two challenges are avoided in our sampling circuit: the first by having one effective sample at a time (Eq. 9, $L = 1$), and the second by replacing complex nonlinear functions with stochastic sampling and linear interactions.

Limitations and extensions of the model. We only address the dynamic inference of a univariate latent stimulus. The multivariate posteriors can be sampled by a coupled circuit where each circuit is the same as the model in the present study and samples a corresponding univariate stimulus. This scheme works for static inference [17], and is straightforward to generalize for dynamic inference. In addition, we implicitly assume that the latent speed is provided to the circuit via SOM's gain, and it is interesting to consider how the circuit infers the speed by itself.

Acknowledgments

W.H.Z. is supported by the UT Southwestern Endowed Scholars program.

References

[1] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170, 2013.

[2] Si Wu, Danmei Chen, Mahesan Niranjan, and Shun-ichi Amari. Sequential bayesian decoding with a population of neurons. *Neural Computation*, 15(5):993–1012, 2003.

[3] Rajesh PN Rao. Bayesian computation in recurrent neural circuits. *Neural computation*, 16(1):1–38, 2004.

[4] Jeffrey M Beck and Alexandre Pouget. Exact inferences in a neural implementation of a hidden markov model. *Neural computation*, 19(5):1344–1361, 2007.

[5] Sophie Deneve, Jean-René Duhamel, and Alexandre Pouget. Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of kalman filters. *Journal of neuroscience*, 27(21):5744–5756, 2007.

[6] Robert Wilson and Leif Finkel. A neural implementation of the kalman filter. *Advances in neural information processing systems*, 22, 2009.

[7] Jean-Pascal Pfister, Peter Dayan, and Máté Lengyel. Know thy neighbour: A normative theory of synaptic depression. *Advances in neural information processing systems*, 22, 2009.

[8] Jean-Pascal Pfister, Peter Dayan, and Máté Lengyel. Synapses with short-term plasticity are optimal estimators of presynaptic membrane potentials. *Nature neuroscience*, 13(10):1271–1275, 2010.

[9] Balazs B Ujfalussy, Judit K Makara, Tiago Branco, and Mate Lengyel. Dendritic nonlinearities are tuned for efficient spike-based computations in cortical circuits. *Elife*, 4:e10056, 2015.

[10] Anna Kutschireiter, Melanie A Basnak, Rachel I Wilson, and Jan Drugowitsch. Bayesian inference in ring attractor networks. *Proceedings of the National Academy of Sciences*, 120(9):e2210622120, 2023.

[11] Patrik O Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. In *Advances in neural information processing systems*, pages 293–300, 2003.

[12] Lars Buesing, Johannes Bill, Bernhard Nessler, and Wolfgang Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology*, 7(11):e1002211, 2011.

[13] Laurence Aitchison and Máté Lengyel. The hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS computational biology*, 12(12), 2016.

[14] Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

[15] Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

[16] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, 2020.

[17] Wen-Hao Zhang, Si Wu, Krešimir Josić, and Brent Doiron. Sampling-based bayesian inference in recurrent circuits of stochastic spiking neurons. *Nature Communications*, 14(1):7074, 2023.

[18] Yu Terada and Taro Toyoizumi. Chaotic neural dynamics facilitate probabilistic computations through sampling. *Proceedings of the National Academy of Sciences*, 121(18):e2312992121, 2024.

[19] Eryn Sale and Wenhao Zhang. The bayesian sampling in a canonical recurrent circuit with a diversity of inhibitory interneurons. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[20] Michael N Shadlen and William T Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of neuroscience*, 18(10):3870–3896, 1998.

[21] Anne K Churchland, Roozbeh Kiani, Rishidev Chaudhuri, Xiao-Jing Wang, Alexandre Pouget, and Michael N Shadlen. Variance as a signature of neural computations during decision making. *Neuron*, 69(4):818–831, 2011.

[22] Paul Masset, Jacob Zavatone-Veth, J Patrick Connor, Venkatesh Murthy, and Cengiz Pehlevan. Natural gradient enables fast sampling in spiking neural networks. *Advances in neural information processing systems*, 35:22018–22034, 2022.

[23] Hillel Adesnik, William Bruns, Hiroki Taniguchi, Z Josh Huang, and Massimo Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–231, 2012.

[24] Cristopher M Niell. Cell types, circuits, and receptive fields in the mouse visual cortex. *Annual review of neuroscience*, 38:413–431, 2015.

[25] Gord Fishell and Adam Kepecs. Interneuron types as attractors and controllers. *Annual review of neuroscience*, 43:1–30, 2020.

[26] Cristopher M Niell and Massimo Scanziani. How cortical circuits implement cortical computations: mouse visual cortex as a model. *Annual Review of Neuroscience*, 44:517–546, 2021.

[27] Luke Campagnola, Stephanie C Seeman, Thomas Chartrand, Lisa Kim, Alex Hoggarth, Clare Gamlin, Shinya Ito, Jessica Trinh, Pasha Davoudian, Cristina Radaelli, et al. Local connectivity and synaptic dynamics in mouse and human neocortex. *Science*, 375(6585):eabj5861, 2022.

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

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

[30] James J Knierim and Kechen Zhang. Attractor dynamics of spatially correlated neural activity in the limbic system. *Annual review of neuroscience*, 35:267–285, 2012.

[31] Mikail Khona and Ila R. Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, December 2022.

[32] Mengni Wang, David J Foster, and Brad E Pfeiffer. Alternating sequences of future and past behavior encoded within hippocampal theta oscillations. *Science*, 370(6513):247–250, 2020.

[33] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417, 2015.

[34] Wen-Hao Zhang, Aihua Chen, Malte J Rasch, and Si Wu. Decentralized multisensory information integration in neural systems. *The Journal of Neuroscience*, 36(2):532–547, 2016.

[35] Si Wu, KY Michael Wong, CC Alan Fung, Yuanyuan Mi, and Wenhao Zhang. Continuous attractor neural networks: candidate of a canonical model for neural information representation. *F1000Research*, 5, 2016.

[36] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.

[37] Si Wu, Kosuke Hamaguchi, and Shun-ichi Amari. Dynamics and computation of continuous attractors. *Neural Computation*, 20(4):994–1025, 2008.

[38] Alexandra K. Moore and Michael Wehr. Parvalbumin-Expressing Inhibitory Interneurons in Auditory Cortex Are Well-Tuned for Frequency. *Journal of Neuroscience*, 33(34):13713–13723, August 2013.

[39] Nathan R. Wilson, Caroline A. Runyan, Forea L. Wang, and Mriganka Sur. Division and subtraction by distinct cortical inhibitory networks in vivo. *Nature*, 488(7411):343–348, August 2012.

[40] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[41] James E Cooke, Martin C Kahn, Edward O Mann, Andrew J King, Jan WH Schnupp, and Ben DB Willmore. Contrast gain control occurs independently of both parvalbumin-positive interneuron activity and shunting inhibition in auditory cortex. *Journal of Neurophysiology*, 123(4):1536–1551, 2020.

[42] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

[43] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[44] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience*, volume 806. Cambridge, MA: MIT Press, 2001.

[45] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[46] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[47] Francois Septier and Gareth W. Peters. Langevin and Hamiltonian based Sequential MCMC for Efficient Bayesian Filtering in High-dimensional Spaces. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):312–327, March 2016.

[48] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

[49] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

[50] Shun-Ichi Amari and Scott C Douglas. Why natural gradient? In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 1213–1216, 1998.

[51] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[52] C. C Alan Fung, K. Y. Michael Wong, and Si Wu. A moving bump in a continuous manifold: A comprehensive study of the tracking dynamics of continuous attractor neural networks. *Neural Computation*, 22(3):752–792, 2010.

[53] Wen-Hao Zhang and Si Wu. Neural information processing with feedback modulations. *Neural Computation*, 24(7):1695–1721, 2012.

[54] James Bigelow, Ryan J Morrill, Jefferson Dekloe, and Andrea R Hasenstaub. Movement and vip interneuron activation differentially modulate encoding in mouse auditory cortex. *ENeuro*, 6(5), 2019.

[55] Deepa L Ramamurthy, Andrew Chen, Jiayu Zhou, Chanbin Park, Patrick C Huang, Priyanka Bharghavan, Gayathri Krishna, Jinjian Liu, Kayla Casale, and Daniel E Feldman. Vip interneurons in sensory cortex encode sensory and action signals but not direct reward signals. *Current Biology*, 33(16):3398–3408, 2023.

[56] Julien Guy, Martin Möck, and Jochen F Staiger. Direction selectivity of inhibitory interneurons in mouse barrel cortex differs between interneuron subtypes. *Cell Reports*, 42(1), 2023.