

**ORIGINAL ARTICLE****Journal Section**

# Nonlinear Evolution via Spatially-Dependent Linear Dynamics for Electrophysiology and Calcium Data

Daniel Hernandez<sup>1\*</sup> | Antonio K. Moretti<sup>2\*</sup> | Shreya Saxena<sup>1</sup> | Ziqiang Wei<sup>3†</sup> | John Cunningham<sup>1</sup> | Liam

Paninski<sup>1</sup>

<sup>1</sup>Department of Statistics, Columbia University, USA

<sup>2</sup>Department of Computer Science, Columbia University, USA

<sup>3</sup>Janelia Research Campus, HHMI, USA;

**Correspondence**

Daniel Hernandez, Google, NYC  
Email: dhernandezdiaz@google.com

**Funding information**

D.H was funded by NIH R01NS100066 and Simons Foundation 54296 grants. A. M. received funding from the NIH/NCI grant U54CA209997 and two NIH shared instrumentation grants, S10 OD012351 and S10 OD021764. S.S. was supported by the Swiss National Science Foundation P400P2\_186759 and NIH 5U19NS104649. Z.W. was funded by the Howard Hughes Medical Institute and by the Simons foundation collaboration on the global brain SCGB 542943SPI.

Latent variable models have been widely applied for the analysis of time series resulting from experimental neuroscience techniques. In these datasets, observations are relatively smooth and possibly nonlinear. We present Variational Inference for Nonlinear Dynamics (VIND), a variational inference framework that is able to uncover nonlinear, smooth latent dynamics from sequential data. The framework is a direct extension of PfLDS; including a structured approximate posterior describing spatially-dependent linear dynamics, as well as an algorithm that relies on the fixed-point iteration method to achieve convergence. We apply VIND to electrophysiology, single-cell voltage and widefield imaging datasets with state-of-the-art results in reconstruction error. In single-cell voltage data, VIND finds a 5D latent space, with variables akin to those of Hodgkin-Huxley-like models. VIND's learned dynamics are further quantified by predicting future neural activity. VIND excels in this task, in some cases substantially outperforming current methods.

**KEYWORDS**

variational inference, nonlinear dynamics, single-cell voltage data

## 1 | INTRODUCTION

In recent years, advances on neural data acquisition have made it possible to record the simultaneous sequential activity of up to thousands of neurons [1]. The analysis of these datasets often focuses on dimensionality reduction techniques that encode the activity of the population in a lower dimensional latent trajectory [2]. At the other extreme, there is a large body of detailed electrophysiological data coming from voltage measurements in single cells [3]. In this setting it is understood that the underlying dynamics are in fact highly nonlinear and multidimensional, though the experimenter only has access to a one-dimensional (1D) observation. From such 1D recordings, the task is to approximately recover the complete latent space paths and dynamics.

A host of sophisticated techniques has been proposed for the analysis of complex sequential data that is not well described by linear transitions and observations [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. Here we present Variational Inference for Nonlinear Dynamics (VIND), a method particularly well-suited for the inference of latent dynamics from relatively smooth time series, such as those typical of neuroscience experiments. The main contribution of VIND is an algorithm that allows variational inference (VI) from structured, intractable approximations to the posterior distribution. In particular, VIND can handle variational posteriors that (i) represent nonlinear evolution in the latent space, and (ii) disentangle the latent dynamics (transition) from the data encoding (recognition). Crucially, the VIND approximate posterior shares the exact nonlinear structure of latent dynamics evolution with the model for data generation. This makes the VIND approximation potentially more powerful than models in which the choice of approximate posterior is made solely on grounds of tractability.

In this work we focus on a VIND variant in which the latent dynamics is represented as a Locally Linear Dynamical System (LLDS). The running time of LLDS/VIND is linear in the number of time points in a trial. We are especially interested in determining LLDS/VIND's ability to find hidden dynamics. After training, can the VIND-trained model generate neural data that is indistinguishable from the original observations, if provided with a suitable starting point?

In the second half of this work we apply VIND to four datasets, one synthetic and the remaining three using experimental data from multi-electrode electrophysiology, single-cell voltage recordings, and dorsal cortex calcium imaging respectively. We show that VIND excels in generative tasks, in some cases outperforming established methods by orders of magnitude in the predictive mean squared error (MSE).

## 2 | BACKGROUND

For a set of temporally ordered, correlated, noisy observations  $\mathbf{X} \equiv \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ ,  $\mathbf{x}_t \in \mathbb{R}^{d_X}$ , a latent variable model proposes an additional, time-ordered set of random variables  $\mathbf{Z} \equiv \{\mathbf{z}_1, \dots, \mathbf{z}_T\}$ ,  $\mathbf{z}_T \in \mathbb{R}^{d_Z}$  that is hidden from view. The hidden state  $\mathbf{z}_t$  is endowed with a stochastic dynamics:  $\mathbf{z}_{t+1} \sim p(\mathbf{z}_{t+1} | \mathbf{z}_{1:t})$  by which it evolves. The observations  $\mathbf{x}_t$  are generated by drawing samples from a  $\mathbf{z}_t$ -dependent probability distribution.

**Variational Inference.** A naive objective for such a model is the marginal log-likelihood  $\log p(\mathbf{X})$ , with the latent variables integrated out of the joint. However, marginalization with respect to  $\mathbf{Z}$  is intractable when the likelihood and prior are non-conjugate and the dynamics are nonlinear [19]. VI overcomes this problem by approximating the posterior  $p(\mathbf{Z} | \mathbf{X})$  with a distribution  $q(\mathbf{Z} | \mathbf{X})$ , the Recognition Model (RM), from a tractable class. The objective becomes the celebrated ELBO, a lower bound to  $\log p(\mathbf{X})$  [20]:

$$\log p(\mathbf{X}) \geq \mathcal{L}_{\text{ELBO}}(\mathbf{X}) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z} | \mathbf{X})]. \quad (1)$$

Auto Encoding Variational Bayes [21] (AEVB) trains both  $q_\phi(\mathbf{Z}|\mathbf{X})$  and  $p_\theta(\mathbf{X}, \mathbf{Z})$  simultaneously. Monte Carlo samples from the recognition distribution are used to compute the expectation in Eq. (1).

**Structured generative models.** We consider the joint density  $p(\mathbf{X}, \mathbf{Z})$ . Our focus is on factorizations of the form:

$$p(\mathbf{X}, \mathbf{Z}) \equiv p_{\phi, \theta}(\mathbf{X}, \mathbf{Z}) = c_{\phi, \theta} \cdot H_\phi(\mathbf{Z}) \prod_{t=0}^T f_\theta(\mathbf{x}_t | \mathbf{z}_t), \quad (2)$$

where the distribution parameters have been written explicitly. The unnormalized densities  $f_\theta$  stand for an observation model that, for the purposes of this work, can be either Gaussian,  $\mathbf{x}_t | \mathbf{z}_t \sim \mathcal{N}(m_\theta(\mathbf{z}_t), \Sigma)$ , or Poisson,  $\mathbf{x}_t | \mathbf{z}_t \sim \text{Poisson}(\lambda_\theta(\mathbf{z}_t))$ . The respective mean,  $m_\theta(\mathbf{z}_t)$ , and rate,  $\lambda_\theta(\mathbf{z}_t)$ , are nonlinear functions of the latent state  $\mathbf{z}_t$ , that we represent as neural networks. The standard deviation  $\Sigma$  of the Gaussian observation model is taken to be  $\mathbf{z}_t$ -independent.  $c_{\phi, \theta}$  is a normalization constant.  $H_\phi$  is the latent evolution term in  $\mathbf{Z}$ -space with a Markov Chain structure [9, 10, 11, 4, 6]:

$$H_\phi(\mathbf{Z}) = h_0(\mathbf{z}_0) \prod_{t=1}^T h_\phi(\mathbf{z}_t | \mathbf{z}_{t-1}), \quad (3)$$

$$\mathbf{z}_0 \sim \mathcal{N}(a_0, \Gamma_0), \quad (4)$$

$$\mathbf{z}_t | \mathbf{z}_{t-1} \sim \mathcal{N}(a_\phi(\mathbf{z}_{t-1}), \Gamma), \quad (5)$$

where  $a_\phi(\mathbf{z})$  is a nonlinear function, usually parameterized by a neural network, and  $\Gamma$  is a trainable parameter.

From Eq. (2), the posterior distribution of the Generative Model (GM) can be factorized as

$$p_{\phi, \theta}(\mathbf{Z}|\mathbf{X}) = \frac{c_{\phi, \theta} \prod f_\theta(\mathbf{x}_t | \mathbf{z}_t)}{p_{\phi, \theta}(\mathbf{X})} \cdot H_\phi(\mathbf{Z}). \quad (6)$$

### 3 | VARIATIONAL INFERENCE FOR NONLINEAR DYNAMICS (VIND)

**Approximate posterior.** Successful VI relies on the choice of the approximation  $q(\mathbf{Z}|\mathbf{X})$ . This choice is constrained by two desirable features that stand in tension: expressiveness and tractability. Specifically, we are interested in representing nonlinear flow in the latent space. Taking Eq. (6) as a guidance, we therefore propose to include the GM evolution term  $H_\phi(\mathbf{Z})$  into the variational posterior. That is, we first consider a posterior that factorizes as:

$$Q_{\phi, \varphi}(\mathbf{Z}|\mathbf{X}) = \kappa_{\phi, \varphi}(\mathbf{X}) G_\varphi(\mathbf{X}, \mathbf{Z}) H_\phi(\mathbf{Z}). \quad (7)$$

The distinguishing feature of VIND is this reusage of the generative evolution term in the Recognition Model.

By design, the factor  $G_\varphi$  in Eq. (7) contains all the dependence on the observations  $\mathbf{X}$ . For definiteness, the case

$$G_\varphi(\mathbf{X}, \mathbf{Z}) = \prod_{t=0}^T g_\varphi(\mathbf{z}_t | \mathbf{x}_t), \quad (8)$$

$$\mathbf{z}_t | \mathbf{x}_t \sim \mathcal{N}(\mu_\varphi(\mathbf{x}_t), \sigma_\varphi(\mathbf{x}_t)), \quad (9)$$

is considered in this work, where  $\mu_\varphi(\mathbf{x})$  and  $\sigma_\varphi(\mathbf{x})$  are nonlinear maps. In Eq. (7),  $\kappa_{\phi, \varphi}$  is a normalization constant. We note that, regardless of the specific form of  $G_\varphi$ ,  $\kappa_{\phi, \varphi}$  cannot be computed in closed form. In particular, the non-Gaussian term  $h(\mathbf{z}_T | \mathbf{z}_{T-1})$ , after integration with respect to  $\mathbf{z}_T$ , yields an intractable  $\mathbf{z}_{T-1}$ -dependent factor, see App. A. As a

consequence of the shared evolution, VI cannot be formulated directly in terms of  $Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$ .

VIND represents a way out of this conundrum that, effectively, allows for the use of an intractable, unnormalized  $Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$  as the Recognition Model in VI. In what follows, we refer to  $Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$  as the *parent* distribution. VIND's idea is to compute a Gaussian approximation  $q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$  to the parent; this *child* distribution then being used as the actual variational posterior in Eq. (1). The inference problem becomes tractable since the child is normal. Importantly, the parameters in  $q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$ , with respect to which we optimize, are inherited from the parent. After training, they can be replaced back into  $Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$  obtaining, in particular, the nonlinear dynamics  $a_\phi(\mathbf{z})$  for the latent space.

Concretely, let the variational posterior  $q_{\phi,\varphi}$  be a Laplace approximation to  $Q_{\phi,\varphi}$ ,

$$q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X}) = \mathcal{N}(\mathbf{P}_{\phi,\varphi}(\mathbf{X}), \mathbf{C}_{\phi,\varphi}^{-1}(\mathbf{X})) . \quad (10)$$

The mean  $\mathbf{P}_{\phi,\varphi}$  in Eq. (10) is the solution to the following equation in  $\mathbf{Z}$ ,

$$\frac{\partial}{\partial \mathbf{Z}} \log Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X}) = 0 , \quad (11)$$

and the precision is given by

$$[\mathbf{C}_{\phi,\varphi}(\mathbf{X})]_{mn} = \frac{\partial^2}{\partial \mathbf{Z}_m \partial \mathbf{Z}_n} \log Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X}) \Big|_{\mathbf{Z}=\mathbf{P}_{\phi,\varphi}(\mathbf{X})} \equiv [s_{\phi,\varphi}(\mathbf{P}_{\phi,\varphi}(\mathbf{X}), \mathbf{X})]_{mn} , \quad (12)$$

where Eq. (12) defines  $s_{\phi,\varphi}$ . The ELBO in Eq. (1) can then be computed with respect to  $q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$  in Eq. (10).

**Fixed-point iteration.** A closed form solution for Eq. (11) is not possible in general. However, for a large class of distributions, and in particular for any  $Q_{\phi,\varphi}$  such that  $\log Q_{\phi,\varphi}$  includes terms quadratic in  $\mathbf{Z}$ , it is possible to rewrite Eq. (11) in the form

$$\mathbf{Z} = r_{\phi,\varphi}(\mathbf{Z}, \mathbf{X}) , \quad (13)$$

where  $r_{\phi,\varphi}$  is a nonlinear function that depends on the trainable parameters in  $Q$ , the data  $\mathbf{X}$  and the choice of the nonlinearities in  $H$ . In this form, Eq. (13) can be solved numerically by making use of the FPI method. That is, a numerical solution for Eq. (13) is found by choosing an initial point  $\mathbf{P}^{(0)}$  and iterating

$$\mathbf{P}^{(n)} = r_{\phi,\varphi}(\mathbf{P}^{(n-1)}, \mathbf{X}) , \quad (14)$$

The VIND method assumes that this FPI converges. In practice, this assumption is guaranteed throughout training by appropriate choices of hyperparameters and network architectures (see supplementary material). In the following section, we will choose a specific form for  $H$  as locally linear dynamics.

**Locally Linear Dynamics.** In the experiments conducted in this paper, the nonlinear dynamics is specified as  $a_\phi(\mathbf{z}) = A_\phi(\mathbf{z})\mathbf{z}$ , where  $A_\phi(\mathbf{z})$  is a state-space dependent  $d_z \times d_z$  matrix. We call this evolution rule, a Locally Linear Dynamical System, and the resulting inference algorithm LLDS/VIND. To derive the latter, consider a parent distribution  $Q_{\phi,\varphi}$ , as defined in Eq. (7). The mean  $\mu_\varphi$  and the standard deviation  $\sigma_\varphi$  in Eq. (9) are represented as deep neural networks:

$$\mu_\varphi = \text{NN}_{\varphi_\mu}(\mathbf{x}_t) , \quad \sigma_\varphi = \text{NN}_{\varphi_\sigma}(\mathbf{x}_t) . \quad (15)$$

**Algorithm 1** Learning VIND: At every epoch  $\mathbf{P}_i^{(\text{ep})}$  is the numerical estimate of the hidden path corresponding to example  $i$ , while  $\mathbf{P}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i)$  is the  $\phi, \varphi$ -dependent posterior mean.

---

```

Initialize  $\phi, \varphi, \theta, \text{Nfpis} = 3$ 
1: for all  $i$  do
2:   Initialize  $\mathbf{P}_i^{(\text{ep})} \leftarrow \mathbf{P}_i^{(0)}$ 
3: end for
4:  $\text{ep} \leftarrow 1; n \leftarrow 0,$ 
5:  $\mathbf{P}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i) \leftarrow \mathbf{P}_i^{(\text{ep}-1)},$ 
6:  $\mathbf{C}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i) \leftarrow s_{\phi,\varphi}(\mathbf{P}_i^{(\text{ep}-1)}, \mathbf{X}_i)$ 
7: while not converged do
   # Sample from  $q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$ 
8:    $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{P}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i), (\mathbf{C}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i))^{-1})$ 
   # Perform gradient descent on  $\sum_i \mathcal{L}_{\text{ELBO}}(\mathbf{X}_i, \mathbf{Z}_i)$ 
9:   ADAM update  $\phi, \varphi, \theta$ 
   # Update  $\mathbf{P}$  and carry the FPI
10:   $\mathbf{P}_i^{(\text{ep})} \leftarrow \mathbf{P}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i)|_{\phi,\varphi}$ 
11:  while  $n \leq \text{Nfpis}$  do
12:     $\mathbf{P}_i^{(\text{ep})} \leftarrow r_{\phi,\varphi}(\mathbf{P}_i^{(\text{ep})}, \mathbf{X})$ 
13:     $n \leftarrow n + 1$ 
14:  end while
   # Initialize next epoch
15:   $\text{ep} \leftarrow \text{ep} + 1; n \leftarrow 0,$ 
16:   $\mathbf{P}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i) \leftarrow \mathbf{P}_i^{(\text{ep}-1)},$ 
17:   $\mathbf{C}_{\phi,\varphi}^{(\text{ep})}(\mathbf{X}_i) \leftarrow s_{\phi,\varphi}(\mathbf{P}_i^{(\text{ep}-1)}, \mathbf{X}_i)$ 
18: end while

```

---

The remaining ingredient of  $Q_{\phi,\varphi}$  is the shared evolution law  $H_\phi$ , Eq. (3). We write the  $h_\phi$  factors that determine the latent evolution model as

$$h_\phi(\mathbf{z}_{t+1}|\mathbf{z}_t) = \exp\left\{-\frac{1}{2}(\mathbf{z}_{t+1} - A_\phi(\mathbf{z}_t)\mathbf{z}_t)^T \Gamma (\mathbf{z}_{t+1} - A_\phi(\mathbf{z}_t)\mathbf{z}_t)\right\}, \quad (16)$$

where  $\Gamma$  is a constant precision matrix, and  $A_\phi(\mathbf{z}_t)$  is specified as

$$A_\phi(\mathbf{z}_t) = \mathbb{A} + \alpha \cdot B_\phi(\mathbf{z}_t) \quad (17)$$

where  $\mathbb{A}$  is a state-space-independent linear transformation initialized to the identity,  $B_\phi(\mathbf{z}_t) = \text{NN}_{\phi_B}(\mathbf{z}_t)$ , and  $\alpha$  is a tunable hyperparameter of the model.

The parameter  $\alpha$  in Eq. (17) parameterizes the deviation of the latent dynamics from linear. For  $\alpha = 0$ , LLDS/VIND reduces, both the statistical model and the algorithm, to GfLDS/PfLDS, [4, 6] which provides a baseline for our model. Also,  $\alpha$  quantifies how much the FPI cost deviates from a quadratic form in the variables  $\mathbf{Z}$ , which is an important consideration for convergence analysis. The results of the paper were obtained by setting  $\alpha = 10^{-2}$ .

Eq. (16) corresponds to the stochastic dynamics of LLDS/VIND:

$$\mathbf{z}_{t+1} \sim \mathbf{A}(\mathbf{z}_t)\mathbf{z}_t + \text{noise} . \quad (18)$$

LLDS/VIND has some desirable features:

1. The limit of linear evolution is easily taken as  $\mathbf{A}_\phi(\mathbf{z}_t) \rightarrow \text{const.}$ .
2.  $\max_{\mathbf{z}} \|\mathbf{A}_\phi(\mathbf{z}_t) - \mathbb{I}\|$  is a simple measure of the smoothness of the latent trajectories.

Using Eqs. (3) and (9), we obtain for the loglikelihood of the parent:

$$\log Q_{\phi,\varphi} = \log C_{\phi,\varphi} - \frac{1}{2} [(\mathbf{Z} - \mathbf{M}_\varphi)^T \mathbf{\Lambda}_\varphi (\mathbf{Z} - \mathbf{M}_\varphi) + \mathbf{Z}^T \mathbf{S}_\phi(\mathbf{Z}) \mathbf{Z}] \quad (19)$$

where  $\mathbf{M}_\varphi = \{\mu_\varphi(\mathbf{x}_1), \dots, \mu_\varphi(\mathbf{x}_T)\}$ ,  $\mathbf{\Lambda}_\varphi$  is a block-diagonal precision matrix,

$$\mathbf{\Lambda}_\varphi = \text{diag}\{\sigma(\mathbf{x}_1), \dots, \sigma_\varphi(\mathbf{x}_T)\} , \quad (20)$$

and  $\mathbf{S}_\phi(\mathbf{Z})$  is a state-space-dependent, block-tridiagonal covariance whose  $d_Z \times d_Z$  blocks are given by:

$$[\mathbf{S}_\phi(\mathbf{Z})]_{t,\tau} = \begin{cases} \mathbf{A}_t^T \mathbf{\Gamma} \mathbf{A}_t & \text{for } \tau = t \\ -\mathbf{\Gamma} \mathbf{A}_t & \text{for } \tau = t + 1 \\ -\mathbf{A}_t^T \mathbf{\Gamma} & \text{for } \tau = t - 1 \\ 0 & \text{otherwise} \end{cases} \quad (21)$$

Here  $\mathbf{A}_t \equiv \mathbf{A}_\phi(\mathbf{z}_t)$ .

Taking the gradients, Eq. (11), we obtain the LLDS/VIND FPI equation for the posterior mean, Eq. (13), with

$$\mathbf{r}_{\phi,\varphi}(\mathbf{Z}, \mathbf{X}) = [\mathbf{\Lambda}_\varphi + \mathbf{S}_\phi(\mathbf{Z})]^{-1} \cdot \mathbf{Y}(\mathbf{Z}) \quad (22)$$

$$\mathbf{Y}(\mathbf{Z}) = \mathbf{\Lambda}_\varphi \mathbf{M}_\varphi - \frac{1}{2} \mathbf{Z}^T \frac{\partial \mathbf{S}_\phi(\mathbf{Z})}{\partial \mathbf{Z}} \mathbf{Z} . \quad (23)$$

Note that the value of the constant  $C_{\phi,\varphi}$  is not required for the FPI step nor for the gradient descent step, thus intractability is evaded. The time complexity of VIND is  $O(T)$ . In particular the matrix  $\mathbf{\Lambda}_\varphi + \mathbf{S}_\phi(\mathbf{Z})$  can be inverted in linear time due to it being block-tridiagonal.

**VIND's algorithm.** VIND's complete algorithm includes two steps per epoch that are carried out in alternation, see Algorithm. 0. The first step is a FPI that, for the current values of the parameters  $\phi, \varphi$ , determines the mean and variance of a Laplace approximation to the parent. The second is a regular ADAM gradient descent update [22] with respect to the ELBO objective. As it is customary, in order to estimate the gradients, the so called "reparameterization trick" is used. Samples are extracted from the child distribution  $q_{\phi,\varphi}$  via:

$$\mathbf{Z}_i = \mathbf{P}_{\phi,\varphi}(\mathbf{X}_i) + [\mathbf{C}_{\phi,\varphi}(\mathbf{X}_i)]^{-1/2} \epsilon \quad (24)$$

where  $\epsilon$  is a standard normal sample, [21, 23]. Upon convergence, the set of parameters  $\phi, \varphi, \theta$  that maximizes the ELBO

can be plugged into  $a_\phi(\mathbf{z})$ , to obtain a dynamical rule that interpolates between the different latent trajectories inferred from the data trials.

During training, the mean  $\mathbf{P}_{\phi,\varphi}$  represents the best current estimate of the latent trajectory. The FPI step in Eq. (14) for LLDS/VIND mixes all the components in  $\mathbf{P}_{\phi,\varphi}$ . In particular, the  $t$ -th component of  $\mathbf{P}_{\phi,\varphi}^{(n)}$  depends in general on all the time steps, both past and future, in  $\mathbf{P}_{\phi,\varphi}^{(n-1)}$  via the inverse covariance in Eq. (23). At every training epoch, the best estimate for the path at a specific time point  $t$  contains information from the complete data. VIND's algorithm is in this sense a smoother.

The FPI procedure gives a closed expression for  $\mathbf{P}_{\phi,\varphi}^{(n)}$ , and derivatives can be taken with respect to the model parameter. In practice we found that  $n = 2$  is enough for good convergence results. The normalization constant for  $q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$ , the determinant of the precision matrix, can be computed in closed form making use of the fact that the precision matrix is block-tridiagonal [24].

## 4 | RELATION TO PREVIOUS WORK

The problem of inference for sequential data has been treated extensively in the literature. The GfLDS and PfLDS models introduced in [4, 6] are particular cases of VIND in which the dynamics in the latent space is linear and time-invariant, i.e.  $\mathbf{z}_t|\mathbf{z}_{t-1} \sim \mathcal{N}(\mathbf{A}\mathbf{z}_{t-1}, \mathbf{Q})$ . In the jargon used in this paper, this corresponds to the situation in which the parent distribution is Gaussian, and therefore equal to its own Laplace approximation. Eq. (11) can be solved analytically in this case and no FPI step is needed.

Several methods model the latent dynamics as a Gaussian Process (GP) [25, 15, 26], differing in the output and observation nonlinearities they assume. Gaussian Process Factor Analysis (GPFA) [25] assumes linear, time-invariant dynamics as well as a linear observation model, i.e.  $\mathbf{x}_t|\mathbf{z}_t \sim \mathcal{N}(\mathbf{C}\mathbf{z}_t + \mathbf{d}, \mathbf{R})$ , for some  $\mathbf{C}$ ,  $\mathbf{d}$ , and  $\mathbf{R}$ . In [26] a GP is used instead as a mapping function with Poisson observations.

AESMC [27], FIVO [28], VSMC [29] and SVO [30, 31, 32] are methods for model inference and learning that maximize a lower bound to the marginal log likelihood, which is in turn approximated using Sequential Monte Carlo. The model learned by VIND is explicitly compared to results obtained by these models in Sec. 5.

In [10], Deep Kalman Filters (DKF) were proposed to handle variational posterior distributions that describes nonlinear evolution in the latent space. Their approximate posterior, analogous to the parent distribution in this paper, is plugged directly into the ELBO. This imposes some restrictions in the form the posterior can take - for instance, it must be Gaussian conditioned on the observations. VIND can handle factorizations of the parent distribution that are not restricted in this way, an example being LLDS/VIND, which has the form in Eq. (7). VIND's ability to handle unnormalizable parent distributions is due to the fact that VIND's actual approximate posterior is always strictly normal. The same authors built upon their idea in [11], where a variational posterior was proposed that partially uses the conditional structure implied by the generative model. In this paper, a similar prescription is used by assuming that  $Q_{\phi,\varphi}$  and  $p_{\phi,\varphi}$  share exactly the same factorization for the latent evolution.

The authors of [9] combine probabilistic graphical models with message passing in an approach based on conjugate priors. The approximate posterior distributions considered in that work are restricted by the conjugacy requirements, in particular, the evolution term must belong to the exponential family. VIND's parent distribution is not subject to this requirement. However, since VIND's actual approximate posterior is still Gaussian, it may be possible to combine the two methods into one that can handle both nonlinear evolution and discrete latent variables.

In [5], Gaussian noise is added to the deterministic evolution rule of an RNN in the context of a variational autoencoder, termed VRNN. Similarly to LLDS/VIND, these authors share the evolution factorization between the generative

model and the approximate posterior and, indeed, the only difference between the structure of their model and that of LLDS/VIND is that the evolution there is expressed as an RNN instead of as an LLDS. However, their inference algorithm only uses past data to estimate the hidden state at any given time. VIND's algorithm, based on the FPI, uses information both from the past and from the future to estimate the latent paths. In [33] a non-parametric approach was taken to determine the best latent dimension in an LDS. It would be interesting to apply those same methods to VIND. Finally, in [13, 14] a sophisticated, bidirectional, Deep Learning-based RNN architecture called LFADS was proposed with neuroscience applications in mind. For both LFADS and DKF, we found difficult to modify their code to compute the quantities that are used in this paper to evaluate the quality of training. However, given the expressive power of these works, we expect them to perform comparably to VIND in the tasks considered in the next section.

## 5 | RESULTS

We demonstrate the capabilities and performance of LLDS/VIND by applying it to four datasets. The first dataset consists of synthetically generated, 10-dimensional noisy observations on top of a 3D latent sequence whose evolution is dictated by an Euler discretization of the Lorenz system. This dataset is the simplest and cleanly illustrates VIND's ability to infer the underlying nonlinear dynamics. Secondly, VIND is applied to a multi-electrode neural recording from a mouse performing a delayed-discrimination task. LLDS/VIND is run with both Gaussian and Poisson observation models. It is found that while a Gaussian observation model is superior for the explaining the variance in the data, the Poisson model performs better when it comes to interpolation of the dynamics.

The third dataset consists of a 1D voltage measurement from single-cell recordings. The problem in this case is not dimensionality reduction but rather to determine the nonlinear underlying dynamics (dimensionality expansion). Interestingly, the number of latent dimensions at which the accuracy of the VIND-extracted dynamics stabilizes coincides with the expectation from theoretical models of spiking neurons. Finally, we apply VIND to the difficult task of uncovering hidden dynamics in a dataset coming from dorsal cortex calcium imaging. We find that it is possible to model the data using a surprisingly low number of latent dimensions and show how to use VIND to reconstruct the dynamics of one side of the brain from the other.

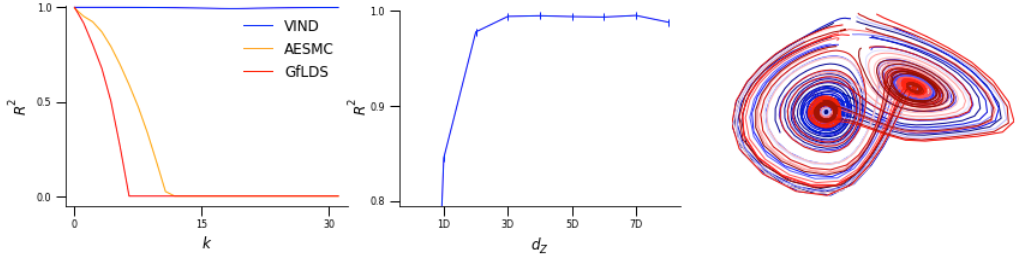
Given an inferred starting point in state-space, the quality of the dynamics learned by LLDS/VIND can be ascertained by evolving the system  $k$  steps into the future *without* any input data. To clarify terminology, this is not strict prediction in the sense of pure extrapolation, since we use information about all  $\mathbf{x}_t$ , both in the past and in the future, to infer the starting point. In order to avoid doubt, we use the term *forward interpolate*. Forward interpolation essentially tests the extent to which the dynamics are accurately learned. We take VIND's capability for forward interpolation as the main measure of the fit's success. As we will show, this task remains highly challenging for simpler smoothing priors like the latent LDS, and it is one of the key strengths of VIND.

To make this analysis quantitative, we compute the  $k$ -step mean squared error ( $\text{MSE}_k$ ) on test data, and its normalized version, the  $R_k^2$ , defined as

$$\text{MSE}_k = \sum_{t=0}^{T-k} (\mathbf{x}_{t+k} - \hat{\mathbf{x}}_{t+k})^2, \quad R_k^2 = 1 - \frac{\text{MSE}_k}{\sum_{t=0}^{T-k} (\mathbf{x}_{t+k} - \bar{\mathbf{x}})^2} \quad (25)$$

where  $\bar{\mathbf{x}}$  is the data average for this trial and  $\hat{\mathbf{x}}_{t+k}$  is the prediction at time  $t + k$ . The latter is obtained by i) using the full data  $\mathbf{X}$  to obtain the best estimate for  $\mathbf{z}_t$ , ii) using  $k$  times the LLDS/VIND evolution equation  $\mathbf{z}_{t+1} = A_\varphi(\mathbf{z})\mathbf{z}_t$ , or  $\mathbf{z}_{t+1} = A\mathbf{z}_t$  for the LDSs, to find the latent state  $k$  time steps in the future, and iii) using the generative network to compute the forward-interpolated observation. Note that in particular,  $k = 0$  corresponds to the standard  $R^2$ . The more





**FIGURE 1** Comparison of results for the Lorenz dataset ( $d_z = 3$ ) between GfLDS and VIND: (left)  $R_k^2$  comparison; (center)  $R_{10}^2$  as a function of dimension of the latent space; (right) VIND's inferred validation trajectories for this dataset.

general  $R_k^2$  ensures that VIND yields more than just a good autoencoder. We will be comparing results obtained with LLDS/VIND to several models, namely, GfLDS, PflDS, AESMC and GPFA (see Sec. 4 for details).

## 5.1 | Lorenz system

The Lorenz system is a classical nonlinear differential equation in 3 independent variables.

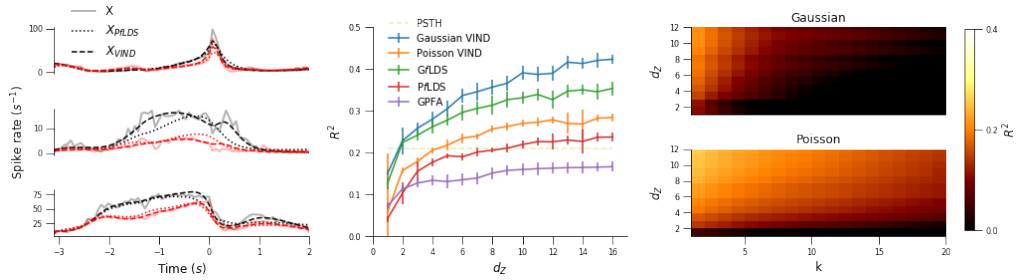
$$\begin{aligned}\dot{z}_1 &= \sigma(z_2 - z_1), \\ \dot{z}_2 &= z_1(\rho - z_3) - z_2, \\ \dot{z}_3 &= z_1 z_2 - \beta z_3.\end{aligned}\tag{26}$$

This is a well studied system with chaotic solutions that serves to cleanly demonstrate VIND's capabilities for inferring nonlinear dynamics.

We generated numerical solutions of the Lorenz system from randomly generated initial conditions, with no noise. Taking  $\sigma = 10, \rho = 28, \beta = 8/3$  Gaussian 10D observations were generated with the mean specified by a  $z$ -dependent neural network. The complete synthetic data consisted of 100 trials, each comprising 250 time-steps, of which 66% was used for training and the remaining were evenly split for test and validation.

In this case, we expect the  $k R_k^2$  to deteriorate very slowly. The results of the fit to this data are shown in Fig. 1. The left panel shows the  $R_k^2$  comparison for VIND and GfLDS fits, with  $d_z = 3$ . Strikingly, for this dataset, VIND's performance does not substantially deteriorate over a 30-step forward interpolation. We show in the left panel comparison with our implementation of the GfLDS and AESMC algorithms. The center panel illustrates VIND's capability to infer properties of the underlying dynamics: VIND hits peak performance at  $d_z = 3$ , the true dimensionality of this system. In the rightmost panel, all the paths inferred by VIND have been put together, showing the famous butterfly pattern.

The obtained latent trajectories are topologically similar to the Lorenz attractor but do not reproduce it exactly. This is expected since VIND's decoder can, in principle, learn to undo any smooth transformation applied to the true Lorenz trajectories. Thus, the same set of observations can be described by different sets of latent paths connected by smooth transformations.



**FIGURE 2** Electrophysiology data. (left) Sample cell spike rates,  $t = 0$  signals the start of the response epoch (center) Performance of explained variance ( $R^2$ ) using different setups of VIND and other models. (right) Performance of forward interpolation ( $R_k^2$ ) using two setups of VIND models.

## 5.2 | Electrophysiology

VIND was used to analyze neural data collected from mice performing a delayed discrimination task in a simultaneous recording session (multi-unit electrophysiology) [34, 35]. In this task, the animals were trained to discriminate the location of a pole using whiskers. The pole was presented at  $t = -1.3$ s, and an auditory go cue at  $t = 0$  signaled the beginning of the response epoch. During response, the mice reported the perceived pole position by licking one of two lick ports. Neurons in this task exhibit complex dynamics across behavioral epochs; some neurons show ramping and persistent activity from sample to delay, which relates to the preparation of the choice at response [34, 35, 36], while some other neurons show the peaking activity in response to the behavioral epochs, see Fig. 2, left.

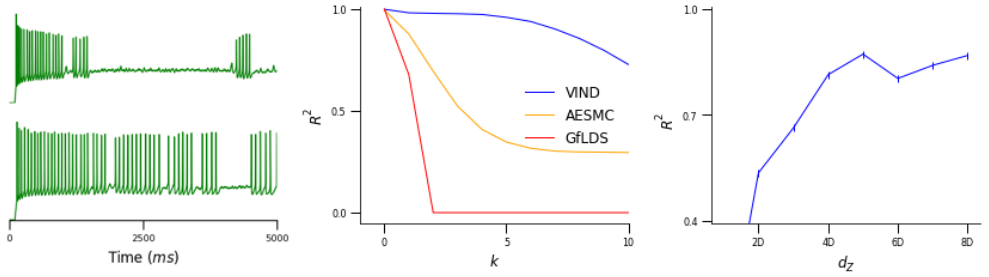
We asked whether VIND can capture the variety of neural dynamics using a few latent observations. The data was fitted for  $d_z = 5$ , using a Poisson observation model. The fit not only reproduces the neural observation, but also provides insights to the dynamics in the latent space and. Specifically, the latent paths separate cleanly by trial type, and the different epochs of the experiment can be seen.

Subsequently, a 10-fold cross-validation method was used to decide the performance of fit using VIND's Gaussian and Poisson observation models with up to 12 dimensions in the latent space, regardless of trial type. The  $R^2$  was computed to determine the performance of VIND as compared to other models. For VIND, both Poisson and Gaussian observation models were used. These are compared to a Peristimulus Time Histogram (PSTH), a GPFA model [25], as well as GfLDS and PLDS [4, 6]. The results are shown in the center panel in Fig. 2. We found that nonlinear Gaussian VIND performs the best regarding explained variance of the data.

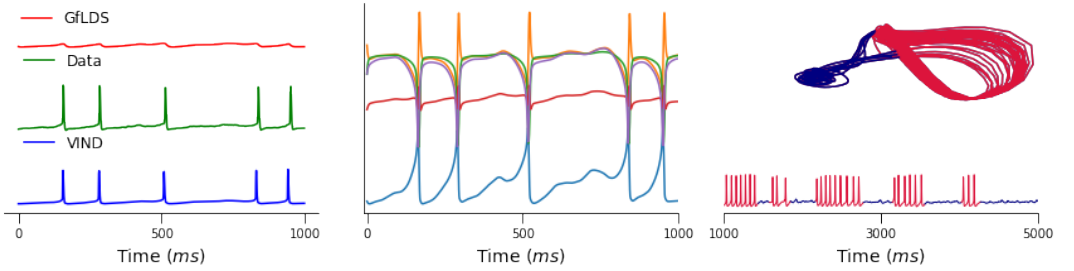
The VIND Poisson observation model gives a substantially better forward interpolation, signaling a dynamical system that more accurately represents the data evolution. This can be seen in the right panel in Fig. 2. These two results combined exemplify the VIND tradeoff between explained variance and forward interpolation capabilities. Using Poisson observations, VIND is less able to fit the higher frequency components of the data. The resulting dynamical system, however, is smoother and more appropriately captures the evolution of the system.

## 5.3 | Single Cell Voltage Data

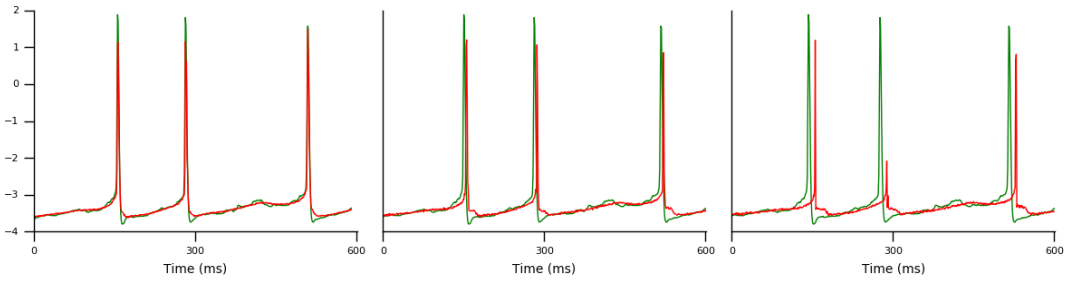
VIND's versatility to uncover underlying dynamics is demonstrated by applying it to 1D voltage electrophysiology data recorded from single cells. This is not a dimensionality reduction problem but rather one of recovering the latent phase



**FIGURE 3** Summary of the LLDS/VIND fit to the Allen dataset: (left) The dataset, neurons respond to an input current; (center) VIND vs GfLDS comparison for the best 5D fits; (right)  $R^2_{10}$  for different dimensions. The performance increases up to  $d_z = 5$  possibly indicating the hidden dimensionality of the system.



**FIGURE 4** Inferred sample paths: (left) Original data (green) versus the 10-step (2ms) forward interpolation given by VIND and by GfLDS; (center) Latent trajectories for a 5D VIND fit of this data, showing behavior similar to the Hodgkin-Huxley gating variables; (right) A 3D cross-section of the latent space showing the representation of the spikes as big cycles (red) and the transient periods (blue).



**FIGURE 5** Data (green) versus simulation of the observations (red) from the smoothed path: 10 steps ahead (left), 20 steps ahead (center), and 30 steps ahead (right). Some signs of deterioration of the prediction start to appear for the latter (failed spikes, late spiking times).

space from a single variable to identify the ‘true dimensionality’ of the system under study. The data is the publicly available Allen Brain Atlas dataset [3].

Intracellular voltage recordings from cells from the Primary Visual Cortex of the mouse, area layer 4 were selected. Trials with no spikes were removed, resulting in 44 trials from 7 different cells. The input for each of the remaining trials consists of a step-function with an amplitude between 80 and 151 pA. Observations were split into training (30 trials) and validation sets (14 trials). The data was then down-sampled from 50,000 time bins (sample rate of 50 kHz) to 5,000 in equal-time intervals, and subsequently normalized by dividing each trial by its maximal value.

LLDS/VIND was fit to this data for  $d_Z = 2, \dots, 8$ , repeated across 10 runs. The top three fits were averaged and the results are summarized in Fig. 3. The center panel displays the  $R_{10}^2$  values for each choice of latent dimensionality. The fits consistently improve up to  $d_Z = 5$ , after which there are diminishing returns. We note that single cell voltage data has traditionally been modeled using variants of the classical Hodgkin-Huxley neuron model ([37]), a set of nonlinear differential equations in 4 independent variables, plus an optional independent input current. It is interesting that 5 is exactly the minimal number of latent dimensions that provide a good VIND fit for this data. The right panel displays  $R_k^2$  with  $d_Z = 5$  for VIND, AESMC and for GfLDS. VIND outperforms GfLDS by an order of magnitude.

The forward-interpolated observations and sample paths for selected runs of VIND and GfLDS are shown in Fig. 4. The left panel represents the observations over a rolling window,  $k = 10$  time-points in advance for both VIND and GfLDS. The dynamics inferred by GfLDS is unable to capture the nonlinear behavior in both the hyperpolarization and depolarization epochs, a task at which VIND succeeds. The VIND latent trajectories are plotted in the center panel, with the latent dimensions exhibiting similar behavior to that of Hodgkin-Huxley gating variables. In state-space, spikes are represented by big cycles (red), while interspiking fluctuations correspond to separate regions of phase space (blue). This is shown in the right panel.

Fig. 5 shows simulated paths (forward interpolation with noise) versus the corresponding real data. The expected, progressive deterioration of the VIND prediction as  $k$  increases is of note. Fig. 6 shows several views of the same two latent paths corresponding to two different input currents showing VIND’s different placement of the paths for two different input currents.



**FIGURE 6** Different views of a 3D cross section of 5D latent paths for two different trials, showing how the paths occupy different regions of state-space depending on the value of the constant input current.

## 5.4 | Spontaneous Activity in Widefield Imaging Data

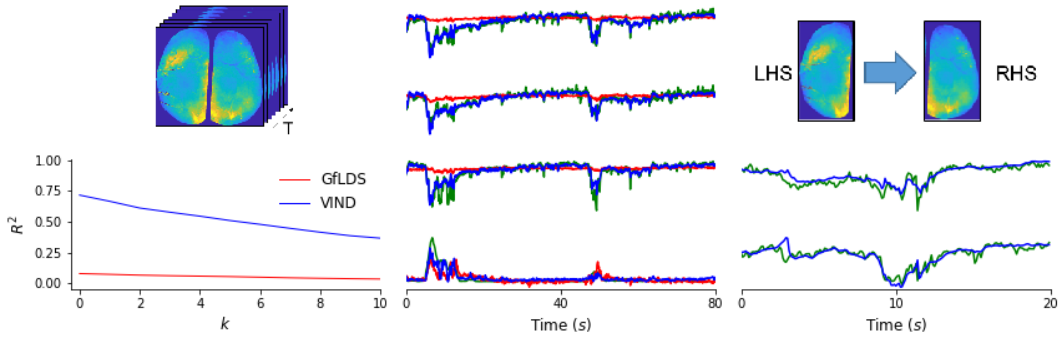
The unsupervised modeling of spontaneous brain activity is inherently challenging due to the lack of task structure. Here, we study the temporal dynamics of widefield optical mapping (WFOM) data and simultaneous behavior recorded from an awake head-fixed mouse during spontaneous activity [38]. This data was recorded and corrected for hemodynamics in the Laboratory for Functional Optical Imaging at Columbia University. An example frame of the data is shown in Fig. 7 (top-left). The preprocessing of the WFOM cortical data leads to reduced-dimension, denoised cortical activity. Details are provided in the supplementary material.

The temporal activity of the cortex and the movement speed (jointly called  $X$ ) are simultaneously modeled using both GfLDS and VIND, with the results for validation data on one mouse shown in Fig. 7, where  $d_X = 148$ , and  $d_Z = 9$ . The  $k$ -step forward interpolation is shown in Fig. 7 (bottom-left), with varying  $k$ , for both VIND and GfLDS. 4 of the 148 dimensions of  $X$  and  $\hat{X}$  on validation data are shown in Fig. 7 (center). VIND is seen to outperform GfLDS, capturing the fine-tuned dynamics in  $X$ , thus also leading to better interpolations. We highlight VIND's capability to roughly capture the dynamics of the whole superficial dorsal cortex using a 9-D latent vector and the corresponding evolution and generative network.

Next, a VIND model was fit to the brain dynamics of only the left hand side (LHS) of the brain, after similar preprocessing of the data. Here,  $d_{X_{LHS}} = 60$ ,  $d_{Z_{LHS}} = 9$ . A separate neural network was fit from the latents learned on the left hand side ( $Z_{LHS}$ ) to the temporal dynamics of the right hand side (RHS) of the brain ( $X_{RHS}$ ;  $d_{X_{RHS}} = 66$ ), with an MSE loss function. The goal was to infer dynamics from one half of the brain to the other. Fig. 7 (right) shows 5 out of 66 reconstructions of the temporal dynamics of the RHS in held-out data (variance weighted average  $R^2 = 0.49$  for entire data). For comparison, we ran a baseline CCA analysis which yielded an  $R^2$  of 0.45. This shows that the latent variables learned by VIND on one half of the brain are useful to coarsely reconstruct the temporal dynamics of the other half.

## 6 | DISCUSSION

In this work we introduced VIND, a novel variational inference framework for nonlinear latent dynamics that is able to handle intractable distributions. We successfully implemented the method for the specific case of Locally Linear Dynamical Systems, which allows for a fast inference algorithm (linear in  $T$ ). When applied to real data, VIND consistently outperforms other methods, in particular methods that rely on an approximate posterior representing linear dynamics and nonlinear, filtering SMC methods. Furthermore, VIND's fits yield insights about the dynamics of these systems.



**FIGURE 7** Widefield Imaging Data: (top-left) An example frame of the data. The temporal dynamics and behavior signal are characterized by  $X$  after preprocessing, which are simultaneously modeled using both GfLDS and VIND. (bottom-left) Variance weighted average  $R^2$  values for  $k$ -step forward interpolation, with  $d_Z = 9$ . (center) An example fit of  $X$  using VIND on held out data. Only 4 of the signals in the 148-dimensional  $X$  signal are shown here. (right) A different VIND model was fit to the temporal dynamics of only the left hand side (LHS) of the brain ( $X_{LHS}$ ). The latents ( $Z_{LHS}$ ) are used to reconstruct the temporal dynamics of the right hand side (RHS) of the video ( $X_{RHS}$ ). Fits are shown on 4 of the 66-dimensional  $X_{RHS}$  in held out data.

Highlights are the ability to identify the transition points and distinguish among trial types in the electrophysiology task, the dimensionality suggested by VIND's fits for the single-cell voltage data, and the ability of the latents learned from one half of the brain to reconstruct activity from the other half in widefield imaging data. Moreover, VIND can be naturally extended to handle labelled data and data with inputs. This is work in progress.

LLDS/VIND is written in tensorflow and the source code is publicly available.

## REFERENCES

- [1] Paninski L, Cunningham J. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *bioRxiv* 2017;p. 196949.
- [2] Cunningham JP, Yu BM. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* 2014 08;17:1500 EP –.
- [3] Jones AR, Overly CC, Sunkin SM. The Allen Brain Atlas: 5 years and beyond. *Nature Reviews Neuroscience* 2009 10;10:821 EP –.
- [4] Archer E, Park IM, Buesing L, Cunningham J, Paninski L. Black box variational inference for state space models. *arXiv: 151107367* 2015;.
- [5] Chung J, Kastner K, Dinh L, Goel K, Courville AC, Bengio Y. A Recurrent Latent Variable Model for Sequential Data. *CoRR* 2015;abs/1506.02216.
- [6] Gao Y, Archer E, Paninski L, Cunningham JP. Linear dynamical neural population models through nonlinear embedding. *NIPS* 2016 2016;.
- [7] Gao Y, Busing L, Shenoy KV, Cunningham JP. High-dimensional neural spike train analysis with generalized count linear dynamical systems. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems 28* Curran Associates, Inc.; 2015.p. 2044–2052.

- [8] Hernandez D, Paninski L, Cunningham J. Variational Inference for Nonlinear Dynamics. TSW, NIPS 2017 2017;.
- [9] Johnson MJ, Duvenaud D, Wiltchko AB, Datta SR, Adams RP. Composing graphical models with neural networks for structured representations and fast inference. arXiv: 160306277 2016;.
- [10] Krishnan RG, Shalit U, Sontag D. Deep Kalman Filters. arXiv: 151105121 2015;.
- [11] Krishnan RG, Shalit U, Sontag D. Structured Inference Networks for Nonlinear State Space Models. arXiv: 160909869 2016;.
- [12] Linderman S, Johnson M, Miller A, Adams R, Blei D, Paninski L. Bayesian learning and inference in recurrent switching linear dynamical systems. In: Artificial Intelligence and Statistics; 2017. p. 914–922.
- [13] Pandarinath C, O'Shea DJ, Collins J, Jozefowicz R, Stavisky SD, Kao JC, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. Nature Methods 2018;15(10):805–815. <https://doi.org/10.1038/s41592-018-0109-9>.
- [14] Sussillo D, Jozefowicz R, Abbott LF, Pandarinath C. LFADS - Latent Factor Analysis via Dynamical Systems. arXiv: 160806315 2016;.
- [15] Zhao Y, Memming Park I. Variational Joint Filtering. arXiv: 170709049 2017 Jul;.
- [16] Wu A, Roy N, Keeley S, Pillow J. Gaussian process based nonlinear latent structure discovery in multivariate spike train data. NIPS 2017 2017;.
- [17] Wu A, Pashkovski S, Datta RS, Pillow JW. Learning a latent manifold of odor representations from neural responses in piriform cortex. NIPS 2018 2018;.
- [18] Zhao Y, Nassar J, Jordan I, Bugallo M, Park IM, Streaming Variational Monte Carlo; 2019.
- [19] Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Berlin, Heidelberg: Springer-Verlag; 2006.
- [20] Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An Introduction to Variational Methods for Graphical Models. Machine Learning 1999 Nov;37(2):183–233.
- [21] Kingma DP, Welling M. Auto-Encoding Variational Bayes. ArXiv: 13126114 2013 Dec;.
- [22] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR 2014;abs/1412.6980.
- [23] Jimenez Rezende D, Mohamed S, Wierstra D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ICML2014 2014 Jan;.
- [24] Trefethen LN, Bau III D. Numerical Linear Algebra. pub-SIAM:adr: pub-SIAM; 1997.
- [25] Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M. Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. Journal of Neurophysiology 2009 07;102(1):614–635.
- [26] Wu A, Roy NA, Keeley S, Pillow JW. Gaussian process based nonlinear latent structure discovery in multivariate spike train data. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30 Curran Associates, Inc.; 2017.p. 3496–3505.
- [27] Le TA, Igl M, Rainforth T, Jin T, Wood F. Auto-Encoding Sequential Monte Carlo. In: International Conference on Learning Representations; 2018. <https://openreview.net/forum?id=BJ8c3f-0b>.
- [28] Maddison CJ, Lawson J, Tucker G, Heess N, Norouzi M, Mnih A, et al. Filtering Variational Objectives. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30 Curran Associates, Inc.; 2017.p. 6573–6583.

- [29] Naesseth C, Linderman S, Ranganath R, Blei D. Variational Sequential Monte Carlo. In: Storkey A, Perez-Cruz F, editors. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, vol. 84 of *Proceedings of Machine Learning Research Playa Blanca, Lanzarote, Canary Islands: PMLR*; 2018. p. 968–977. <http://proceedings.mlr.press/v84/naesseth18a.html>.
- [30] Moretti A, Wang Z, Wu L, Pe'er I. Smoothing Nonlinear Variational Objectives with Sequential Monte Carlo. *Deep Generative Models for Highly Structured Data Workshop, ICLR 2019*; <https://openreview.net/forum?id=HJg24U8tuE>.
- [31] Moretti AK, Wang Z, Wu L, Drori I, Pe'er I. Particle Smoothing Variational Objectives. *CoRR* 2019;abs/1909.09734.
- [32] Moretti AK, Wang Z, Wu L, Drori I, Pe'er I. Variational Objectives for Markovian Dynamics with Backward Simulation. *European Conference on Artificial Intelligence 2020*;
- [33] Kalantari R, Ghosh J, Zhou M. Nonparametric Bayesian Sparse Graph Linear Dynamical Systems. *ArXiv: 180207434* 2018 Feb;.
- [34] Guo ZV, Li N, Huber D, Ophir E, Gutnisky D, Ting JT, et al. Flow of Cortical Activity Underlying a Tactile Decision in Mice. *Neuron* 2014 2018/05/16;81(1):179–194.
- [35] Li N, Chen TW, Guo ZV, Gerfen CR, Svoboda K. A motor cortex circuit for motor planning and movement. *Nature* 2015 02;519:51 EP –.
- [36] Wei Z, Inagaki H, Li N, Svoboda K, Druckmann S. An orderly single-trial organization of population dynamics in premotor cortex predicts behavioral variability. *Nature Communications* 2019;10(1):216. <https://doi.org/10.1038/s41467-018-08141-6>.
- [37] Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology* 1952;117(4):500–544.
- [38] Ma Y, Shaik MA, Kim SH, Kozberg MG, Thibodeaux DN, Zhao HT, et al. Wide-field optical mapping of neural activity and brain haemodynamics: considerations and novel approaches. *Phil Trans R Soc B* 2016;371(1705):20150360.
- [39] Ma Y, Shaik MA, Kozberg MG, Kim SH, Portes JP, Timerman D, et al. Resting-state hemodynamics are spatiotemporally coupled to synchronized and symmetric neural activity in excitatory neurons. *Proceedings of the National Academy of Sciences* 2016;113(52):E8463–E8471.
- [40] Buchanan KE, Friedrich J, Kinsella I, Stinson P, Zhou P, Gerhard F, et al. Constrained matrix factorization methods for denoising and demixing voltage imaging data. In: *Cosyne*; 2018..



## A | VIND'S INTRACTABILITY

Consider a simple toy model comprising just two time steps. According to Eq. (7),  $Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$  would be given by:

$$Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X}) = \kappa_{\phi,\varphi}(\mathbf{X}) \tilde{Q}_{\phi,\varphi}(\mathbf{Z}|\mathbf{X}), \quad (27)$$

where  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$  and

$$\tilde{Q}(\mathbf{Z}|\mathbf{X}) = g(\mathbf{z}_0|\mathbf{x}_0)g(\mathbf{z}_1|\mathbf{x}_1) \cdot h_0(\mathbf{z}_0)h(\mathbf{z}_1|\mathbf{z}_0), \quad (28)$$

stands for the unnormalized distribution:

$$\kappa_{\phi,\varphi}^{-1}(\mathbf{X}) = \int \tilde{Q}(\mathbf{Z}|\mathbf{X}) d\mathbf{Z}. \quad (29)$$

Parameter subindices are suppressed in what follows for convenience.

Even in this simple setup, direct integration of  $\tilde{Q}$ , as in Eq. (29), is unsuccessful. To illustrate this, consider the simplified case in which the variance parameters are all set to the identity:

$$\Gamma_0 = \Gamma = \sigma_\varphi = \mathbb{I}_{d_Z}. \quad (30)$$

Then, marginalizing first with respect to  $\mathbf{z}_1$ :

$$\int \tilde{Q} d\mathbf{z}_1 = h(\mathbf{z}_0)g(\mathbf{z}_0|\mathbf{x}_0) \cdot I(\mathbf{z}_0|\mathbf{x}_1) \quad (31)$$

where  $I(\mathbf{z}_0|\mathbf{x}_1)$  is given by

$$I(\mathbf{z}_0|\mathbf{x}_1) = \int \exp \left\{ -\frac{1}{2} \Delta(\mathbf{z}_1|\mathbf{z}_0)^T \Delta(\mathbf{z}_1|\mathbf{z}_0) - \frac{1}{2} \Delta(\mathbf{z}_1|\mathbf{x}_1)^T \Delta(\mathbf{z}_1|\mathbf{x}_1) \right\} d\mathbf{z}_1, \quad (32)$$

with

$$\Delta(\mathbf{z}_1|\mathbf{z}_0) = \mathbf{z}_1 - a_\phi(\mathbf{z}_0), \Delta(\mathbf{z}_1|\mathbf{x}_1) = \mathbf{z}_1 - \mu_\varphi(\mathbf{x}_1). \quad (33)$$

Carrying out the integral,

$$I(\mathbf{z}_0|\mathbf{x}_1) = \frac{1}{(2\pi)^{d_Z}} \exp \left\{ -\frac{1}{4} (a_\phi(\mathbf{z}_0) - \mu_\varphi(\mathbf{x}_1))^2 \right\}. \quad (34)$$

The desired normalizing constant would then be given by

$$\kappa^{-1} = \int h(\mathbf{z}_0)g(\mathbf{z}_0|\mathbf{x}_0)I(\mathbf{z}_0|\mathbf{x}_1) d\mathbf{z}_0. \quad (35)$$

However, the argument of the exponential in the integrand includes terms in  $a_\phi(\mathbf{z}_0)$  and  $a_\phi(\mathbf{z}_0)^2$  which are non-quadratic in  $\mathbf{z}_0$ . They are the source of the intractability. In turn, these are mandated by VIND's factorization of the approximate

posterior, inherited from the Generative Model.

## B | REVIEW OF THE FIXED-POINT ITERATION METHOD

The FPI method (also known as Picard Fixed-Point Iteration) yields a numerical approximation to the solution of a system of  $k$  nonlinear equations in  $k$  independent variables:

$$F_i(x) = 0, \quad i = 1, \dots, k \quad (36)$$

where  $x \in \mathbb{R}^k$ . To apply the FPI the system is transformed into the form

$$x = T(x) \quad (37)$$

where  $T : \mathbb{R}^k \rightarrow \mathbb{R}^k$ . An initial estimate  $x_0$  is subsequently picked. The FPI algorithm then generates the sequence  $x_n$  by applying  $T$  repeatedly:

$$x_n = T(x_{n-1}). \quad (38)$$

If this sequence converges, then it is Cauchy and its limit is the solution of Eq. (37).

The fundamental convergence result for Picard iterations is the Picard-Banach-Cacciopoli (PBC) theorem, formulated for operators  $T, T : X \rightarrow X$  where  $(X, d_X)$  is a complete metric space:

**Theorem 1. (PBC)** *Let  $T$  be Lipschitz-continuous in  $U \subset X$ . That is*

$$d_X(T(x), T(y)) \leq K \cdot d_X(x, y), \quad \text{for } x, y \in U \quad (39)$$

*for some real number  $K$ . If  $K \in [0, 1)$  then  $T$  has a unique fixed point  $x^* \in U$  and the Picard sequence  $\{x_n\}$  for  $n = 0, \dots, \infty$  where*

$$x_n = T(x_{n-1}) = T^n(x_0) \quad (40)$$

*converges to  $x^*$  for any initial guess  $x_0 \in U$ .*

It can be further shown that the rate of convergence is exponential in the iteration number

$$d_X(x_n, x^*) \leq K^n \cdot d_X(x_0, x^*). \quad (41)$$

When the PBC theorem holds, we say the map  $T$  is a  $K$ -contraction.

Let  $J_{ij}(x)$  be the Jacobian of the map  $T$ ,  $i, j = 1, \dots, k$ . Let  $\{\lambda_i(x_0)\}$  be the eigenvalues of  $J_{ij}$  evaluated at  $x_0$ . A common way to show that a mapping  $T : \mathbb{R}^k \rightarrow \mathbb{R}^k$  is a contraction under the Euclidean distance in a neighborhood of  $x_0 \in \mathbb{R}^k$ , is to show that  $\max \lambda_i < 1$ . In turn this can be proven using the Gershgorin Circle Theorem that gives a bound to the spectrum of a square matrix  $A$ :

**Theorem 2. (Gershgorin)** *Let  $a_{ij}$  be the entries of the square matrix  $A$  and  $r_i = \sum_{j \neq i} |a_{ij}|$ . Then every eigenvalue of  $A$  lies within a disc centered at  $a_{ii}$  with radius  $r_i$ .*

As a corollary, an upper bound on the maximum absolute value for the eigenvalues of  $A$  is obtained:

$$\max_i \lambda_i \leq \max_i \sum_j |a_{ij}|. \quad (42)$$

Applied to the Picard iteration, a sufficient condition for its convergence is obtained:

$$\max_i \sum_j |J_{ij}| = \max_i \sum_j \left| \frac{\partial T_i}{\partial x_j} \right| < 1. \quad (43)$$

In what follows, we use this result to obtain an order-of-magnitude estimate for the VIND hyperparameters such that convergence of the VIND FPI is plausible.

## C | IMPLEMENTATION DETAILS OF LLDS/VIND

In this appendix, we provide extra details of the VIND framework for the LLDS parameterization of the hidden dynamics.

- VIND is initialization-sensitive. The initial estimates for the latent path  $\mathbf{P}_i^{(0)}$ , the starting point for the FPI, are taken to be  $\mathbf{M}_\phi(\mathbf{X}_i)$ . Moreover, empirically, we found that it is important that the initial path estimates fall within a region where the nonlinearity is not severe ( $\max_{\mathbf{P}_i} \|A_\phi(\mathbf{z}_t) - \mathbb{I}\| \lesssim 0.1$  for every trial  $i$ ). This is guaranteed by proper initialization of the parameters of the recognition network.
- The local transformation  $A_\phi(\mathbf{z}_t)$  is redundant (it is akin to a gauge transformation in physics parlance). To see this, note that for every  $\mathbf{z}_t$ , the image of the transformation  $A_\phi(\mathbf{z}_t)\mathbf{z}_t$  is a subset of  $\mathbb{R}^n$ . On the other hand  $A_\phi(\mathbf{z}_t)$  has dimensionality  $\mathbb{R}^{n^2}$ . In other words, given  $\mathbf{z}_t$  and  $\mathbf{z}_{t+1}$ , there is a continuum of matrices  $A_\phi(\mathbf{z}_t)$  that satisfy  $\mathbf{z}_{t+1} = A_\phi(\mathbf{z}_t)\mathbf{z}_t$ . As a consequence,  $A_\phi(\mathbf{z}_t)$  can be substantially restricted without loss of generality (“fixing the gauge”). In our code,  $A_\phi(\mathbf{z}_t)$  was constrained to be symmetric with good results.
- The number of FPIs to produce good convergence results in Algorithm 1 (see main text) depends on the dataset. We found that  $n = 2$  was a good compromise that yielded convergence across datasets.
- In all experiments, no noticeable decrease in performance was found if the gradient terms in  $r_{\phi,\varphi}$  - see for instance Eq. (47) - and the corresponding ones for  $s_{\phi,\varphi}$  are neglected. These terms are subleading compared to  $\Lambda_\varphi \mathbf{M}_\varphi$  both because they are proportional  $\alpha$ , and because the gradient is applied on a deep neural network.

**FPI convergence.** As detailed in the main text, Algorithm 1, a VIND training epoch consists of two steps that are carried in alternate fashion: the FPI that updates the best estimate of the latent path, and the gradient descent step that updates the model parameters. For LLDS/VIND to converge, the FPI, defined by the map  $r_{\phi,\varphi}$ :

$$\mathbf{P} = r_{\phi,\varphi}(\mathbf{P}, \mathbf{X}) \quad (44)$$

$$r_{\phi,\varphi}(\mathbf{P}, \mathbf{X}) = \tilde{\Lambda}^{-1} \cdot \mathbf{Y}(\mathbf{P}) \quad (45)$$

$$\tilde{\Lambda} = \Lambda + \mathbf{S}(\mathbf{Z}) \quad (46)$$

$$\mathbf{Y}(\mathbf{P}) = \Lambda_\varphi \mathbf{M}_\varphi - \frac{1}{2} \mathbf{P}^T \frac{\partial \mathbf{S}_\phi(\mathbf{P})}{\partial \mathbf{P}} \mathbf{P}. \quad (47)$$

must be in the contractive regime within a domain  $D$ ,  $D \subset \mathbb{R}^{T \times d_Z}$ . As remarked in App. B, a necessary condition for this

to occur is that the Jacobian  $J$  of the map  $r_{\phi,\varphi}$ :

$$J_{ij}(\mathbf{Z}) = \frac{\partial r_i}{\partial Z_j}, \quad \text{for } i, j \in 1, \dots, T \times d_Z. \quad (48)$$

satisfies Eq. (43).

For  $\alpha = 0$ ,  $\log Q_{\phi,\varphi}(\mathbf{Z}|\mathbf{X})$  is a quadratic form in  $\mathbf{Z}$ . In this case the VIND FPI is a convex optimization problem, Eq.(11) is linear with a closed form solution. Therefore, deviations from convergence and convexity are always  $O(\alpha)$ .

In what follows, we provide an order-of-magnitude estimation to argue that  $\alpha$  can be chosen to be small enough to preserve convergence and convexity in LLDS/VIND. For the sake of clarity, we remove the parameter subindices.

For the specific case of LLDS/VIND, the entries  $J_{ij}$  are suppressed both by the small hyperparameter  $\alpha$  and by the gradients of the deep neural network  $B_\phi(\mathbf{z}_t)$ , Eq. (17). Neglecting the subleading terms in Eq. (47) proportional to the gradient of  $\mathbf{S}(\mathbf{Z})$ :

$$\frac{\partial r_i}{\partial Z_j} \simeq \tilde{\Lambda}^{-1} \frac{\partial \tilde{\Lambda}}{\partial Z_j} \tilde{\Lambda}^{-1} \cdot \Lambda \mathbf{M}_\varphi \simeq \tilde{\Lambda}_{ik}^{-1} \frac{\partial \tilde{\Lambda}_{kl}}{\partial Z_j} \cdot r_l. \quad (49)$$

For an order of magnitude estimate of the necessary scales involved, let  $L$  be the typical linear dimension of a bounding box in latent space inside which the latent paths are contained,

$$r \sim L. \quad (50)$$

Let  $\sigma^2$  represent the typical scale of the entries of the diagonal recognition covariance matrix  $\Lambda$ , and let  $\sigma_{\text{ev}}^2 = \Gamma^{-1}$  represent the typical scale of the evolution covariance. Moreover, for simplicity consider the case in which  $\Lambda \gtrsim \mathbf{S}(\mathbf{Z})$ , so that in magnitude,

$$\tilde{\Lambda}^{-1} \sim \sigma^2 \cdot \mathbb{I} \quad (51)$$

Let  $\Delta$  represent the typical rate of variation of the entries of the matrix  $B(\mathbf{z}_t)$ . Then we have

$$\frac{\partial \tilde{\Lambda}_{kl}}{\partial Z_j} \sim \frac{\alpha \Delta}{\sigma_{\text{ev}}^2} V_{klj} \quad (52)$$

where  $V_{klj}$  is a sparse tensor (only the  $(j, j)$ ,  $(j, j + 1)$  and  $(j + 1, j)$  blocks in  $\tilde{\Lambda}_{kl}$  can depend on  $Z_j$ ). Replacing all these into Eq. (43) we obtain a simple rule that, when satisfied, suggests the FPI is in the contractive regime

$$\max_i \sum_j \left| \frac{\partial r_i}{\partial Z_j} \right| \sim c \frac{\sigma^2}{\sigma_{\text{ev}}^2} \alpha \Delta L. \quad (53)$$

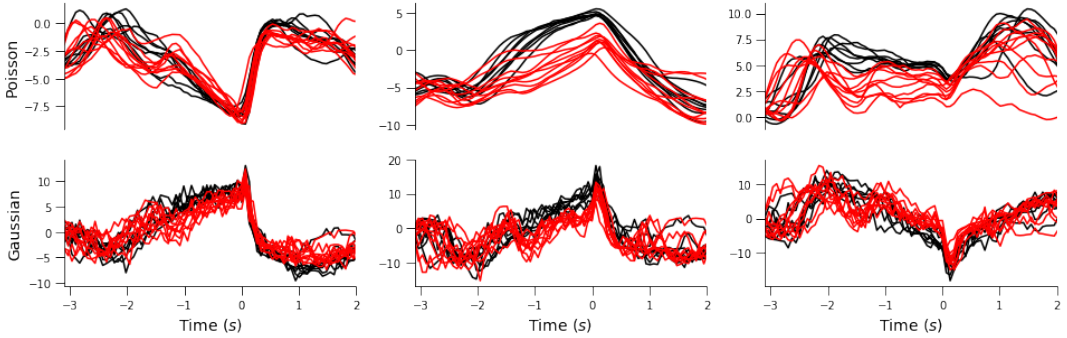
where  $c$  is an  $O(1)$  constant.

In practice, and guided by this analysis, we tune the hyperparameters and architecture of the evolution network so that

$$\alpha \Delta \ll \frac{\sigma_{\text{ev}}^2}{L \sigma^2} \quad (54)$$

at initialization with good results.

## D | DETAILS OF THE ELECTROPHYSIOLOGY TASK AND RESULTS



**FIGURE 8** Examples of latent dimension dynamics for Gaussian and Poisson VIND in validation data. Black lines, posterior pole location; red lines, anterior pole location. Notice how the inferred paths differ for posterior and anterior pole locations. Also note visible changes in dynamics at  $t = -1.3$  (stimulus), and  $t = 0$  (go cue).

A recording session contains 18 simultaneous recorded units, with 74 lick-left trials, (posterior pole location), and 100 lick-right trials (anterior pole location). Spike counts were binned in a 67 ms non-overlapped time window, which resulted in a number of spike counts per bin between 0 and 10. The fit covers the time interval  $[-0.5, 2.0]$ , going from the onset epoch to the end of the response epoch. At time = 0s, the mouse receives the go cue. Each trial contains 77 time bins.

Fig. 8 shows the average neuronal activity of 3 representative cells in the recordings. Cell #1 is a typical neuron with small separation of trials, but strong peaking activity at transition from delay to response epochs. Cells #2, 3 exhibit the stereotypical ramping activity and separations of different trial types, which are assumed for preparation of the movements. Both VIND setups (Poisson and Gaussian observations, nonlinear evolutions,  $d_z = 5$ ) can reproduce the complex and variable neural dynamics in the held-out trials (9 lick-left trials; 9 lick-right trials).

In particular, the Gaussian VIND model can capture the changes of dynamics on finer timescales. On the other hand, the latent dynamics are smoother in the Poisson VIND model, Fig. 8. Smoother trajectories are correlated with superior performance in the forward interpolation tasks. Intuitively, for noisier latent paths, the algorithm attempts to ascribe some of the variance to the dynamical system, which hurts the forward interpolation capabilities. In the Poisson VIND fit represented in Fig. 8, the latent dynamics in dimensions 2 and 3 appears to represent the preparation of the choice where the neural dynamics for different trial types gradually diverges with time. The dynamics in latent dimension 1 shows rapid peaking dynamics at the transitions of the behavioral epochs. However, those two types of dynamics were mixed in the Gaussian VIND fit. In general, ramping and peaking dynamics is not operated by distinguishable groups of neurons, yet to our surprise they are separated in the latent space.

## E | PREPROCESSING OF WIDEFIELD IMAGING DATA

Macro-scale wide-field optical mapping (WFOM) is an increasingly popular technique for surveying neural activity over very large areas of cortex with high temporal resolution. WFOM can image the fluorescence of genetically-encoded calcium (GCaMP6f) indicators using LED illumination and camera detection scheme. We use methods for correcting fluorescence recordings of neural activity for confounding contamination by changes in hemoglobin concentration and oxygenation as in [39], by measuring both neural fluorescence signals and hemodynamics. This correction provides us with an accurate change in fluorescence of neural regions ( $\Delta F/F$ ).

An example frame of the data is shown in the main text, Fig. 5, 464-by-473 pixels. The activity of the mouse is simultaneously recorded using a webcam pointed at the mouse's body, and the movement speed at time  $t$  is taken as a 1D signal consisting of the standard deviation of the difference in value of all pixels from time  $t - 1$  to time  $t$ .

We use a WFOM recording of length 2 minutes, where the signals are sampled at 10Hz, thus leading to 1200 time points. We normalize  $\Delta F/F$  to lie between 0 and 1 for every video, and then apply block singular value decomposition (SVD) to the videos for denoising and dimensionality reduction [40]. First, we fit an anisotropic Wiener filter in a  $4 \times 4$  neighborhood of each pixel to reduce uncorrelated noise while preserving spatially-local, time-correlated signals. Next, the video is partitioned into 25 ( $5 \times 5$ ) blocks, and SVD is performed on the pixels in each block. The temporal components are ranked according to a metric defined on their empirical autocorrelation function, and components that fall within a 99% confidence interval of Gaussian white noise are discarded. Moreover, those temporal components that have a signal-to-noise ratio lower than 1.6 are also discarded. The remaining temporal components from each block are concatenated, and these form the  $X$  matrix, here  $147 \times 1200$ . This is augmented using a 1D behavior signal that is extracted using the standard deviation of successive frames from a webcam recording the lateral view of the mouse's body, representing the speed of the mouse's movements in arbitrary units. We used different sessions of recording from the same mouse, preprocessed in the same way, to obtain training and validation data.

## F | DETAILS OF THE SEQUENTIAL MONTE CARLO FITS

Auto-Encoding Sequential Monte Carlo [27] is a method for model inference and learning using a variant of the ELBO constructed from the Sequential Monte Carlo marginal likelihood estimator. In our experiments the proposal distribution factorizes into separate functions for an evolution of the latent dynamics and an encoding of the data:

$$Q_{SMC}(\mathbf{Z}_{1:T} | \mathbf{X}_{1:T}) = \prod_{t=1}^T h_{SMC}(\mathbf{z}_t | \psi(\mathbf{z}_{t-1}), \Gamma) g_{SMC}(\mathbf{z}_t | \gamma(\mathbf{x}_t), \Lambda)$$

This choice is advantageous because  $h_{SMC}(\mathbf{z}_t | \mathbf{z}_{t-1})$  is designed to share parameters with the evolution term of the generative model. In this way the resulting evolution term of the approximate posterior is exact. The functions  $\psi : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{d_z}$  where  $\psi(\mathbf{z}_t) = \mathbf{z}_{t+1}$  and  $\gamma : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_z}$  where  $\gamma(\mathbf{x}_t) = \mathbf{z}_t$  are nonlinear time invariant represented with two layer neural networks. We found that training separate networks for both the evolution term of the proposal and the evolution term of the generative model resulted in numerical issues when computing importance weights that caused AESMC to fail to converge.