Applications of information geometry to spiking neural network activity

Jacob T. Crosser and Braden A. W. Brinkman*

Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794, USA and Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, New York 11794, USA

(Received 18 May 2023; accepted 10 January 2024; published 7 February 2024)

The space of possible behaviors that complex biological systems may exhibit is unimaginably vast, and these systems often appear to be stochastic, whether due to variable noisy environmental inputs or intrinsically generated chaos. The brain is a prominent example of a biological system with complex behaviors. The number of possible patterns of spikes emitted by a local brain circuit is combinatorially large, although the brain may not make use of all of them. Understanding which of these possible patterns are actually used by the brain, and how those sets of patterns change as properties of neural circuitry change is a major goal in neuroscience. Recently, tools from information geometry have been used to study embeddings of probabilistic models onto a hierarchy of model manifolds that encode how model outputs change as a function of their parameters, giving a quantitative notion of "distances" between outputs. We apply this method to a network model of excitatory and inhibitory neural populations to understand how the competition between membrane and synaptic response timescales shapes the network's information geometry. The hyperbolic embedding allows us to identify the statistical parameters to which the model behavior is most sensitive, and demonstrate how the ranking of these coordinates changes with the balance of excitation and inhibition in the network.

DOI: 10.1103/PhysRevE.109.024302

I. INTRODUCTION

A major obstacle to understanding the computational underpinnings of the brain is the high dimensionality of its inputs—environmental stimuli such as light and sound—and its outputs—the activity of neurons and the organismal behaviors they enact [1–8]. The output space of a neural circuit with N neurons is unmanageably large: the possible spike train patterns of such a network over a trial of time length T divided into time bins of size Δt forms a data- or output-space of dimension $NT/\Delta t$. As $\Delta t \to 0$, this output space becomes infinite-dimensional. However, the activity of a neural population does not occupy this entire space, as activity is correlated across time and neurons, and the actual output of any given neural circuit constitutes just a subset of all possible observations. Perhaps surprisingly, analysis of experimental data has repeatedly found that under many conditions collective neural activity is low dimensional, often comprising less than $\approx 10^2$ dimensions of this infinite space [9–20].

Compounding this challenge is the fact that the neural properties and architectures that give rise to observed patterns of behavior are not unique: many distinct neural circuits give rise to essentially identical patterns of activity [21–24]. This multiplicity of possible underlying network properties is known as "degeneracy." While degeneracy makes it difficult, if not impossible, to precisely infer the circuitry underpinning observed activity patterns, it could prove advantageous for treating circuitry exhibiting pathological activity. If there are many possible circuit configurations that produce "normal" activity, then it may be possible to tune networks back to normal operation by targeted yet modest perturbations using, e.g., pharmacological manipulations rather than electrophysiological inputs, leading to more effective treatments of neurological disorders.

Understanding how to take advantage of degeneracy in neural circuitry requires a formalism for quantifying the similarity between network models with different parameters, and identifying directions in parameter space along which model output changes dramatically or minimally under coordinated changes in model parameters. Brute force simulations of network activity is computationally expensive and impractical for circuits larger than a few neurons. The tools of information geometry offer a possible formalism [25–32] that may be easier to apply to larger networks and begin to understand how to most effectively move a network through its parameter space to achieve desired outputs, at the cost of mapping models into abstract "output spaces." Information geometry represents the output space of a model as a manifold parametrized by the parameters of the model, which in our context will be properties of the neurons and their synaptic connections (among many other possibilities, such as stimulus or current inputs). Notably, information geometric analyses of several kinds of complex systems in biology yield model manifolds that can be described as a hierarchy of "hyperribbons" [33–40]. Hyperribbons are manifolds with a few long directions of the manifold, representing "stiff" directions that separate disparate activity states, and many thin directions, which represent "sloppy" directions that describe networks with very similar model outputs. These model manifolds come equipped with a natural metric that defines a sense of distance between model behaviors. This makes it possible to determine

*Corresponding author: braden.brinkman@stonybrook.edu

the stiff combinations of parameters that predict the bulk of the neural response space of the network model. This opens a path for better understanding of how to manipulate network properties to tune a circuit between different regimes of activity by identifying directions along hyperribbons that induce the greatest or smallest changes in model output as parameters are adjusted.

In neuroscience, tools from information theory and information geometry have traditionally been used to understand sensory coding by investigating how neural responses can optimally encode different distributions of sensory inputs [41–53]. Notably, the hierarchical nature of biology has been observed in low-dimensional—i.e., stiff—representations of biological data. For example, a highly hierarchical geometric structure underlies the natural correlations of odors from a variety of natural sources such as fruit [54], the expression of genes across cell types and stages of differentiation in humans [55], and spatial representations in the mouse hippocampus [56].

In contrast, in this work we use information geometry as a means of understanding the structure of complex models themselves, and how these models reflect changes in network activity produced by changing intrinsic or synaptic properties of neural populations, in line with foundational work applying ideas from information geometry to study abstracted representations of spiking networks [25,29], networks of rate models [28,30–32], or neural field models [26,27]. In this work we leverage recently developed information geometric tools [39,40] and path integral representations of spiking neural population models [4,57] to study a biologically plausible leaky integrate-and-fire model whose parameters are directly interpretable in terms of measurably biophysical properties, such as membrane and synaptic timescales.

While it is experimentally impractical to grow neural networks with different properties but the same output patterns, the cell and synaptic properties of neurons can be tuned pharmacologically using neuromodulators [58–60]. To this end, we investigate how the balance of single-neuron properties and the properties of the synaptic connections between neurons shape the hierarchy of possible outputs of the networks. Our focus on the contrast between single neuron versus synaptic properties is loosely motivated by experimental work studying the action of dopamine depletion and synaptic blocking on neural firing patterns [59]. We abstract the action of dopamine depletion in these experiments into changes in the synaptic and membrane timescales and study how changing the membrane and synaptic time constants of the networks shape the manifold hierarchy. We also investigate how these changes are modified if the balance of excitation and inhibition in the network is disrupted, changing the hierarchical ordering of the model output space.

We organize the paper as follows: in Sec. II we introduce the class of stochastic spiking models we will be working with and the reduction to a population-based formalism. Then, in Sec. III, we give a self-contained explanation of the "isKL" embedding method introduced by Ref. [40], and how it applies to our population model. We detail the results of the application of the isKL method in Sec. IV, and finally discuss the interpretation and significance of our results and methodology in Sec. V.

II. MODELS

A. Nonlinear Hawkes process

To model the spiking dynamics of individual neurons, we consider a nonlinear Hawkes process [4,57]

$$\frac{dV_i}{dt} = -\tau_m^{-1}(V_i - \varepsilon_i) + I_i + \tau_s^{-1}\left(\mu_{\text{ext}} - J_{\text{self}}\dot{n}_i(t) + \sum_{j=1}^{n} w_{ij}\dot{n}_j(t)\right), \tag{1a}$$

$$\dot{n}_i(t)dt \sim \text{Poiss}[\phi(V_i(t))dt], \tag{1b}$$

where V_i is the membrane potential of neuron i, ε_i is the leak reversal potential, w_{ij} is the strength of a synaptic connection from neuron j to neuron i, and $-J_{\text{self}}$ is an inhibitory self-coupling to implement post-spike refractory dynamics. The two currents μ_{ext} and I_i represent an average current received from an external network and an experimentally injected current that differs by neuron, respectively. The process $\dot{n}_i(t)$ is the spike train of neuron i, and $\phi(V_i(t))dt$ is the instantaneous firing rate nonlinearity that determines a Poisson event rate conditioned on the membrane potential of a given neuron. For the specific models studied here, $\phi(x) = \frac{1}{2}[x + (x^2 + 1/2)^{1/2}]$, which represents a soft rectified linear activation function. Finally, τ_m and τ_s are modulated parameters corresponding to membrane and synaptic timescales, respectively. Equation (1a) of this model assigns leaky integration dynamics to the membrane potential of each individual neuron, while Eq. (1b) assigns conditionally Poisson spiking dynamics to each neuron. Taken together, this model can be thought of as a soft-threshold leaky integrate-and-fire system.

Foreshadowing the coming analysis, we note that analytically calculating the statistical properties of the models in Eq. (1) is generally intractable, and to make headway we implement a Gaussian-process approximation of the network dynamics around the mean-field activity.

We can obtain a mean-field approximation of the steady-state solution for the membrane potential dynamics in Eq. (1a) by marginalizing out the spiking dynamics and assuming the distribution is sharply peaked around the most probable path of $V_i(t)$. Assuming the network achieves a steady state at long times, this procedure gives us a set of transcendental equations that can be solved numerically:

$$V_i^{\text{mf}} = \varepsilon_I + \tau_m I_i + \frac{\tau_m}{\tau_s}\left(\mu_{\text{ext}} - J_{\text{self}}\phi\left(V_i^{\text{mf}}\right) + \sum_j w_{ij}\phi\left(V_j^{\text{mf}}\right)\right), \tag{2}$$

where the V_i^{mf}, the solutions of these equations, are the mean-field predictions of the steady-state values of membrane potentials, with $\phi(V_i^{mf})$ being the corresponding mean-field prediction of the firing rates. We find the solutions to these transcendental equations using a forward-Euler integration scheme.

Following the prescription of Refs. [61–63], the time-dependent distribution of model behaviors described in Eq. (1)

can be written in the form of a path integral,

$$P[\mathbf{V}(t), \dot{\mathbf{n}}(t)] = \int \mathfrak{D}[\tilde{\mathbf{V}}, \tilde{\mathbf{n}}] e^{-S[\tilde{\mathbf{V}}, \mathbf{V}, \tilde{\mathbf{n}}, \dot{\mathbf{n}}]}, \tag{3}$$

with an action S given by

$$\begin{aligned} S[\tilde{\mathbf{V}}, \mathbf{V}, \tilde{\mathbf{n}}, \dot{\mathbf{n}}] = \int dt \sum_{i=1}^{n} \Bigg\{ \tilde{V}_i \Bigg[\dot{V}_i + \frac{V_i - \varepsilon_i}{\tau_m} - I_i \\ - \tau_s^{-1} \Bigg(\mu_{\text{ext}} - J_{\text{self}} \dot{n}_i(t) + \sum_j w_{ij} \dot{n}_j(t) \Bigg) \Bigg] \\ + \tilde{n}_i(t) \dot{n}_i(t) - (e^{\tilde{n}_i(t)} - 1) \phi(V_i) \Bigg\}. \end{aligned} \tag{4}$$

The Gaussian process approximation of the membrane dynamics in Eq. (1a) is obtained by marginalizing out the spiking dynamics from the action in Eq. (4) and taking a saddle-point approximation of the action around the mean-field solution in Eq. (2) (see Appendixes A and B for details). The resulting action corresponds to the Gaussian stochastic process given by [4,62]

$$d\mathbf{V} = \mathbf{A}(\mathbf{V}^{mf} - \mathbf{V})dt + \mathbf{\Sigma} d\mathbf{W}_t, \tag{5a}$$

$$A_{ij} = \delta_{ij} \left(\tau_m^{-1} + \tau_s^{-1} J_{\text{self}} \phi' \left(V_j^{\text{mf}} \right) \right) - \tau_s^{-1} w_{ij} \phi' \left(V_j^{\text{mf}} \right), \tag{5b}$$

$$\begin{aligned} (\mathbf{\Sigma}\mathbf{\Sigma}^T)_{ij} = \tau_s^{-2} \sum_k \big[(-\delta_{ik} J_{\text{self}} + w_{ik}) \\ \times (-\delta_{jk} J_{\text{self}} + w_{jk}) \phi \left(V_k^{\text{mf}} \right) \big], \end{aligned} \tag{5c}$$

where $d\mathbf{W}_t$ is a standard Wiener process and we use the Itô convention. We note that Eq. (5a) is an Ornstein-Uhlenbeck (OU) process, albeit one in which the drift and diffusion matrices are dependent on the mean-field values of the membrane potential.

In principle, the network modeled in Eq. (1) and approximated in Eq. (5) could be of arbitrary size. To make our information geometric analysis tractable, however, we reduce the model to a three-population model, comprising excitatory and inhibitory populations, and a single neuron targeted with an injected current; this is depicted diagrammatically in Fig. 1. We start by considering the connectivity matrix to be random with each entry being a Bernoulli variable with probability p being scaled by a connection type-dependent value w_{IJ}. To produce the more tractable reduced model, we take a population-averaging approach to the approximated process in Eq. (5). We now use an uppercase subscript to denote a population-averaged variable. For example,

$$V_I \equiv \frac{1}{N_I} \sum_{i \in I} V_i(t),$$

where we use uppercase indices $I, J, K \in \{0, 1, 2\}$ to denote the different populations, with $I = 0$ the single test neuron, $I = 1$ the excitatory population, and $I = 2$ the inhibitory population. The dynamics of the population-averaged membrane potentials under the Gaussian approximation now follow a

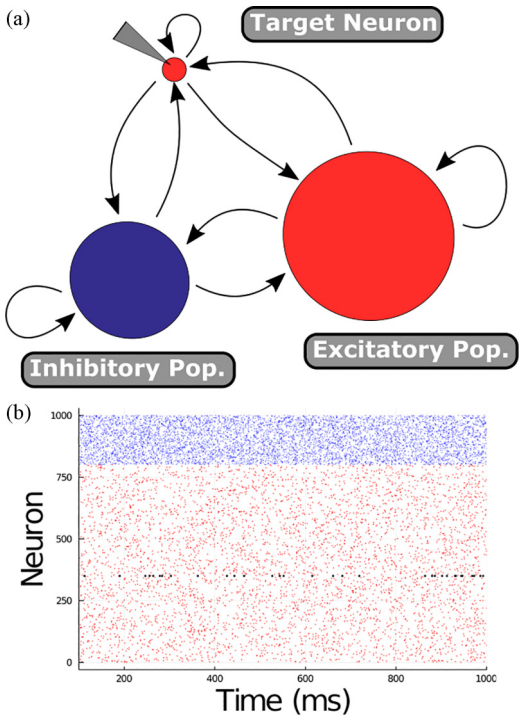

FIG. 1. Network model. (a) A graphical representation of the network architecture being studied. (b) An example raster plot generated from an extended network of spiking neurons modeled by Eq. (1).

lower-dimensional version of Eq. (5a) with drift and diffusion matrices given by

$$\begin{aligned} A_{IJ} = \delta_{IJ} \left(\tau_m^{-1} + \tau_s^{-1} J_{\text{self}} \phi' \left(V_I^{\text{mf}} \right) \right) \\ - \tau_s^{-1} p w_{IJ} N_J \phi' \left(V_J^{\text{mf}} \right), \end{aligned} \tag{6a}$$

$$\begin{aligned} (\mathbf{\Sigma}\mathbf{\Sigma}^T)_{IJ} = \tau_s^{-2} \sum_{K=0,1,2} \Bigg[\left(-\delta_{IK} \frac{J_{\text{self}}}{N_K} + p w_{IK} \right) \\ \times \left(-\delta_{JK} \frac{J_{\text{self}}}{N_K} + p w_{JK} \right) N_K \phi \left(V_K^{\text{mf}} \right) \Bigg] \\ \approx \tau_s^{-2} \sum_{K=1,2} p^2 w_{IK} w_{JK} N_K \phi \left(V_K^{\text{mf}} \right). \end{aligned} \tag{6b}$$

The approximation in the last line above comes from the fact that $N_1, N_2 \gg 0$. The population-averaged mean-field equations are now

$$\begin{aligned} V_I^{\text{mf}} = \varepsilon_I + \tau_m I_{\text{I}} + \frac{\tau_m}{\tau_s} \mu_{\text{ext}} \\ + \frac{\tau_m}{\tau_s} \left(-J_{\text{self}} \phi \left(V_I^{\text{mf}} \right) + \sum_{J=0,1,2} p w_{IJ} N_J \phi \left(V_J^{\text{mf}} \right) \right). \end{aligned} \tag{6c}$$

We formally derive the Gaussian-process approximation of the full-network [Eq. (5)] and the population-averaged approximation [Eq. (6)] in Appendix A. We also note that the statistics of the model in Eq. (6) are equivalent to those derived by first taking a population average of the membrane potential dynamics and then applying the Gaussian

approximation framework. This second derivation is provided in Appendix B.

The spiking model we study centers around a balanced network, specifically a network that is not finely tuned. This notion of fine-tuning arises from a standard derivation of balance equations for the model system (see Appendix D). In short, we can look at the average external input κ_I into population I. For our model, we can approximate κ_I to leading order as

$$\tau_s^{-1}\kappa_I \approx \sqrt{N}\Bigg(\frac{1}{\sqrt{N}}\big(I_\mathrm{I} + \tau_s^{-1}\mu_\mathrm{ext}\big) + \tau_s^{-1}\left\{pw_{I1}\frac{N_1}{\sqrt{N}}\phi(V_1) + pw_{I2}\frac{N_2}{\sqrt{N}}\phi(V_2)\right\}\Bigg),$$

where V_I is the population-averaged membrane potential for population I. For the model to be in a balanced state, the variance of the synaptic input should be $O(N^0)$ which in turn implies the synaptic weights should scale as $w_{IJ} \sim 1/\sqrt{N}$. Additionally, we assume that $N_I \propto N$ and $\langle I_I \rangle$, $\mu_\mathrm{ext} \propto \sqrt{N}$. The balanced state of the model also requires that all κ_I be $O(1)$. For this to be true as $N \to \infty$, the terms in the parentheses must vanish. This gives us a linear system that uniquely defines $(\phi(V_1),\ \phi(V_2))$:

$$-\begin{bmatrix} I_1 + \tau_s^{-1}\mu_\mathrm{ext} \\ I_2 + \tau_s^{-1}\mu_\mathrm{ext} \end{bmatrix} = \frac{1}{\tau_s}\begin{bmatrix} pw_{11}N_1 & pw_{12}N_2 \\ pw_{21}N_1 & pw_{22}N_2 \end{bmatrix}\begin{bmatrix} \phi(V_1) \\ \phi(V_2) \end{bmatrix}. \quad (7)$$

From this set of equations, we derive two cases. First, if the matrix on the right-hand side of Eq. (7) is singular and neither of the columns of the matrix are trivially the zero vector, the columns must be scalar multiples of each other. We refer to this as a "fine-tuned" spiking model. If the left-hand side of Eq. (7) is also a multiple of the columns, the system admits an infinite set of solutions $(\phi(V_1),\ \phi(V_2))$. Otherwise, it admits no solution. Such a network is thus finely tuned to specific inputs. In contrast with this, we have "untuned" spiking models. In this case, the matrix on the right-hand side of Eq. (7) is invertible and the system admits a unique solution $(\phi(V_1),\ \phi(V_2))$. This in effect applies constraints on the values of $\{w_{IJ}\}$, which we refer to as the balance equations (D1) and (D2) (see Appendix D for a derivation). Moving forward, we consider only spiking models derived from a balanced, untuned network. We also introduce a linear nonspiking model that will serve as a baseline comparison.

B. Linear nonspiking model

Although the Gaussian process approximation of the spiking network will have a Gaussian steady-state distribution of the membrane potentials, the parameters of this distribution vary nonlinearly with the self-consistent mean-field solutions. To demonstrate that the results we observe are consequences of the mean-field treatment of the spiking network, and not just the behavior of Gaussian processes more generally, we also construct a simpler model of networked, linear nonspiking (or "graded potential") neurons. We assume the neurons are injected with large numbers of synaptic input that sum together to be approximately Gaussian, with nonzero mean μ_ext, creating a stochastic system with dynamics described by

$$\frac{dV_i}{dt} = -\tau_m^{-1}(V_i - \varepsilon_I) + I_i + \tau_s^{-1}\mu_\mathrm{ext} - \tau_s^{-1}J_\mathrm{self}\phi(V_i) + \tau_s^{-1}\sum_j w_{ij}\phi(V_j) + \xi_i(t). \quad (8)$$

We choose the transfer function $\phi(\cdot)$ to be the identity function [i.e., $\phi(x) = x$], which could be viewed as an approximation of a nonlinear activation function around the setpoint $x_i = 0$, but allows for exact results. The processes $\xi_i(t)$ are zero-mean Gaussian noise synaptic input from neurons external to the network being examined, and thus they scale with τ_s^{-1}. We define the covariance of the noise processes $\{\xi_i(t)\}$ as follows:

$$\langle \xi_i(t)\xi_j(t')\rangle = \tau_s^{-2}\delta_{ij}\mu_\mathrm{ext}\delta(t - t').$$

After population-averaging, the nonspiking model becomes another OU process:

$$d\mathbf{V} = \mathbf{A}\big[\mathbf{A}^{-1}\big(\tau_s^{-1}\mu_\mathrm{ext} + \tau_m^{-1}\varepsilon_I + \mathbf{I}\big) - \mathbf{V}\big]dt + \mathbf{\Sigma}d\mathbf{W}_t = \mathbf{A}(\mu - \mathbf{V})dt + \mathbf{\Sigma}d\mathbf{W}_t. \quad (9)$$

The drift and diffusion matrices are defined as follows:

$$A_{IJ} = \delta_{IJ}\tau_m^{-1} + \tau_s^{-1}w^*_{IJ},$$
$$w^*_{IJ} = -\delta_{IJ}J_\mathrm{self} + pw^\mathrm{mod}_{IJ}N_J,$$
$$(\Sigma\Sigma^T)_{IJ} = \tau_s^{-2}\delta_{IJ}\frac{\mu_\mathrm{ext}}{N_I}.$$

Here and in the following sections, $\mathbf{w}^*$ denotes the effective connectivity matrix for the linear nonspiking models. The values of $\mathbf{w}^\mathrm{mod}$ are modulated depending on the desired excitation-inhibition conditions, which will be discussed in Sec. II D.

The linear form of the population-averaged nonspiking model permits more analytic study than the corresponding spiking models. Ornstein-Uhlenbeck processes like those in Eqs. (6) and (9) admit a Gaussian steady-state distribution if all eigenvalues of the drift matrix are positive [64]. From the form of the drift matrix for the linear model [Eq. (9)] there is a correspondence between eigenvalues of the drift matrix $\mathbf{A}$ and the connectivity matrix $\mathbf{w}^*$:

$$\lambda_{i,\mathbf{A}} = \tau_m^{-1} - \tau_s^{-1}\lambda_{i,\mathbf{w}^*}.$$

From the stationarity condition on the eigenvalues of $\mathbf{A}$ and this correspondence between eigenvalues of $\mathbf{A}$ and $\mathbf{w}^*$, we can derive a stability boundary for the $(\tau_m^{-1}, \tau_s^{-1})$ inverse timescale space:

$$\tau_m^{-1} > \tau_s^{-1}\lambda_{\mathbf{w}^*}\ \forall\ \lambda_{\mathbf{w}^*}. \quad (10)$$

Because we choose $\phi(\cdot)$ to be linear in this nonspiking model [Eq. (9)], the stability condition in Eq. (10) is exact and predicts when the model admits a stationary statistical solution. The loss of stability observed in OU processes often corresponds to a regime in which the random variables may grow without bound. As the firing rate nonlinearity $\phi(x)$ used in the spiking model [Eq. (1)] is asymptotically linear in the $x > 0$ regime, we expect the stability condition (10) to approximately predict when a similar transition to an explosive nonstationary state occurs in spiking models that have the

TABLE I. Model parameters. Descriptions and numerical values for the parameters for the nonspiking and spiking model types.

Parameter	Description	Value
N	Total number of neurons	1000
N_e	Number of excitatory neurons	$0.8N$
N_i	Number of inhibitory neurons	$0.2N$
p	Probability of a directional synaptic connection w_{ij} between any two neurons	0.1
$-J_{\text{self}}$	Self connection for a neuron of type E-I designed to capture post-spike refractory dynamics	-5
ε_I	Leak reversal potential for neuron i	0
I_I	Injected current impinging on neuron population I	$\begin{cases}0.02 \text{ if target}\\ 0 \text{ otherwise}\end{cases}$
μ_{ext}	Mean input from network-external neurons	0.1
$w_{ee,\text{base}}$	Total expected synaptic input weight from exc. neurons to exc. neurons	285
$w_{ie,\text{base}}$	Total expected synaptic input weight from exc. neurons onto inh. neurons	300
$w_{ei,\text{base}}$	Total expected synaptic input weight from inh. neurons to exc. neurons	-902.5
$w_{ii,\text{base}}$	Total expected synaptic input weight from inh. neurons to inh. neurons	-950
$\phi(x)$	Firing rate transfer function	$\begin{cases}x \text{ if nonspiking}\\ \frac{x+\sqrt{x^2+\frac{1}{2}}}{2} \text{ if spiking}\end{cases}$
τ_m	Membrane timescale	Variable
τ_s	Synaptic timescale	Variable

same connectivity w_{IJ} as the nonspiking models. Nonstationarity in the spiking model approximations must be detected numerically.

C. Stationary distributions

As mentioned above, the stationary distributions admitted by Ornstein-Uhlenbeck processes are Gaussian when they exist [64]. Consider a general N-dimensional OU process:

$$d\mathbf{X} = \mathbf{A}(\mu - \mathbf{X})dt + \mathbf{\Sigma} d\mathbf{W}_t.$$

The stationary distribution, when it exists, is described by the multivariate normal probability density [64]

$$p(\mathbf{X}) = \frac{1}{(2\pi)^{N/2}\sqrt{\det(\mathbf{C})}} e^{-\frac{1}{2}(\mathbf{X}-\mu)^T \mathbf{C}^{-1}(\mathbf{X}-\mu)},$$

where the stationary covariance $\mathbf{C}$ is given by the solution to the matrix equation [64]

$$\mathbf{\Sigma}\mathbf{\Sigma}^T = \mathbf{A}\mathbf{C} + \mathbf{C}\mathbf{A}^T.$$

In practice, the stationary covariance matrix can by found by linearizing the matrix equation and solving the resulting linear system numerically.

D. Network architectures

Now, we turn back to our network models. We consider a population of excitatory and inhibitory neurons in which a single excitatory target neuron is injected with an external driving current. The full network contains $N = 1000$ sparsely connected neurons. We reduced the full network model into representative a three-neuron network by population-averaging, as depicted in Fig. 1(a) and described in Eq. (6), representing the excitatory target neuron, the excitatory population, and the inhibitory population. Table I contains descriptions and numerical values for the parameters used in the present study.

In addition, we would like to adjust the relative recurrent excitation and inhibition in the networks. To accomplish this, the base connection weights given in Table I are scaled by a ratio $r > 0$ depending on the desired activity regime:

$$\mathbf{w}_{Xe}^{\text{mod}} = r_e(r) w_{Xe,\text{base}} = \begin{cases} r w_{Xe,\text{base}} & \text{if } r \geqslant 1 \\ w_{Xe,\text{base}} & \text{otherwise,} \end{cases} \tag{11a}$$

$$\mathbf{w}_{Xi}^{\text{mod}} = r_i(r) w_{Xi,\text{base}} = \begin{cases} w_{Xi,\text{base}} & \text{if } r \geqslant 1 \\ \frac{1}{r} w_{Xi,\text{base}} & \text{otherwise.} \end{cases} \tag{11b}$$

The ratio function $r_e(r)$ boosts the recurrent excitatory weights when $r > 1$, but holds the excitatory weights constant when $r < 1$. Conversely, $r_i(r)$ boosts the recurrent inhibitory weights when $r < 1$ and holds them constant when $r > 1$. Taken together, these functions lead to a regime of strong recurrent inhibition (an "inhibitory regime") when $r < 1$ and a regime of strong recurrent excitation (an "excitatory regime") when $r > 1$.

The connection matrices $\mathbf{w}$ and $\mathbf{w}^*$ of the population-averaged spiking and nonspiking models, respectively, are now constructed from the full-network parameters and scaling of excitation and inhibition. All matrices $\mathbf{w}$ and $\mathbf{w}^*$ use the same indexing with $I = 0$ denoting the target neuron "population," $I = 1$ denoting the remaining excitatory neurons, and $I = 2$ denoting all inhibitory neurons. The connection strengths used in the linear nonspiking model in Eq. (9) is

then given by

$$\mathbf{w}^* = -\begin{bmatrix} J_{\text{self}} & 0 & 0 \\ 0 & J_{\text{self}} & 0 \\ 0 & 0 & J_{\text{self}} \end{bmatrix} + \frac{1}{\sqrt{pN}} \begin{bmatrix} pw_{ee} & p(N_e-1)w_{ee} & pN_i w_{ei} \\ pw_{ee} & p(N_e-1)w_{ee} & pN_i w_{ei} \\ pw_{ie} & p(N_e-1)w_{ie} & pN_i w_{ii} \end{bmatrix}$$
$$= -\begin{bmatrix} J_{\text{self}} & 0 & 0 \\ 0 & J_{\text{self}} & 0 \\ 0 & 0 & J_{\text{self}} \end{bmatrix} + \frac{1}{\sqrt{pN}} \begin{bmatrix} pr_e(r)w_{ee,\text{base}} & pr_e(r)(N_e-1)w_{ee,\text{base}} & pr_i(r)N_i w_{ei,\text{base}} \\ pr_e(r)w_{ee,\text{base}} & pr_e(r)(N_e-1)w_{ee,\text{base}} & pr_i(r)N_i w_{ei,\text{base}} \\ pr_e(r)w_{ie,\text{base}} & pr_e(r)(N_e-1)w_{ie,\text{base}} & pr_i(r)N_i w_{ii,\text{base}} \end{bmatrix}. \tag{12}$$

The $1/\sqrt{pN}$ scaling of the connection weights arises from the balance conditions mentioned at the end of Sec. II A and derived in Appendix D. The connection matrices used by the spiking models described generally in Eq. (6) is given by

$$\mathbf{w} = \frac{1}{\sqrt{pN}}\mathbf{w}^{\text{mod}}$$
$$= \frac{1}{\sqrt{pN}} \begin{bmatrix} r_e(r)w_{ee,\text{base}} & r_e(r)w_{ee,\text{base}} & r_i(r)w_{ei,\text{base}} \\ r_e(r)w_{ee,\text{base}} & r_e(r)w_{ee,\text{base}} & r_i(r)w_{ei,\text{base}} \\ r_e(r)w_{ie,\text{base}} & r_e(r)w_{ie,\text{base}} & r_i(r)w_{ii,\text{base}} \end{bmatrix}. \tag{13}$$

Finally, we would like a measure of the balance of excitation and inhibition (E-I) within a class of models. As each model type corresponds to many particular models with different values of the inverse timescales $(\tau_m^{-1}, \tau_s^{-1})$, we require a proxy measure for the E-I ratio to describe the whole class. In line with the method for adjusting the relative strength of recurrent excitation and inhibition introduced above, we assign a ratio of connection weights into the bulk excitatory population for a given model and a given modulation r. For the nonspiking models, we give the log-ratio R of these weights:

$$R = \log_{10}\left|\frac{\mathbf{w}^*_{2,1} + \mathbf{w}^*_{2,2}}{\mathbf{w}^*_{2,3}}\right|.$$

To make an accurate comparison with the nonspiking models, the E-I values for the spiking models are reported using this same measure (for a given value of modulation parameter r).

We note here that the same balanced-network calculations (see Appendix D) that gave rise to the definitions of fine-tuned and untuned formally define a notion of balance. A balanced spiking network based on the model architecture used here must satisfy constraints on the weights of $\mathbf{w}$ [Eq. (13)], either Eq. (D1) or Eq. (D2). The base connection weights for the unadjusted network—i.e., $r = 1$ in Eq. (11)—were chosen to meet these balance criteria, and the functions $r_e(r)$ and $r_i(r)$ serve to tilt the excitation-inhibition balance with respect to this measure.

E. Timescale sampling

To embed and visualize the model manifolds of interest, we must sample points on the manifold characterized by different values of the two modulated parameters. We do this by sampling a portion of the inverse-timescale parameter space that satisfies the stability condition given by Eq. (10) and where both inverse-timescales are positive. We apply a curvilinear grid to this region, uniformly sampling the radial and angular components. The radial distance components d of the grid are taken over a fixed range $d \in [0.0025, 0.03]$ ms^{-1}. To apply both the stability boundary and positivity constraints, the lower bound of the angular component α of the sample grid is set to a fixed value while the upper bound is set either by the stability boundary described by Eq. (10) or to a fixed value, whichever is more stringent:

$$\tan(\alpha) \in \left[0.1,\ \min\left(\frac{1}{\max\{\lambda_{\mathbf{w}^*}\}}, 500\right)\right].$$

The conditions for this maximal sampling are summarized in Table II. It is important to note that the stability boundary is determined by the eigenvalues of the connectivity matrix $\mathbf{w}^*$, and thus the stability boundary and the sampling region are affected by the induced E-I balance is adjusted through its affect on $\mathbf{w}^*$. The spiking models use the connection matrix from the equivalent nonspiking model to set the sampling range. The maximal sampling scheme is depicted diagrammatically in Fig. 2.

After the maximal sampling of parameter space for each model type for each E-I condition, sample points from the inverse-timescale space are subject to further exclusionary criteria. For both the spiking- and nonspiking-type models, sample points are excluded if they cause either the drift matrix $\mathbf{A}$ or the covariance matrix $\mathbf{C}$ to become singular. The singularities in these matrices have been observed to occur numerically near the theoretical stability boundary [Eq. (10)]. In addition, sample points for the spiking models are excluded if the Euler integration used to find the mean-field solutions to Eq. (6c) does not converge. The integration is determined to be numerically nonconvergent if the rate of change of the system either exceeds a predetermined value during integration

TABLE II. Maximal sampling parameters. Descriptions and numerical values for the parameters that are constant across the nonspiking, spiking with fine-tuning, and spiking without fine-tuning model types.

	Radial distance d	Angle α
Minimum value	0.0025 ms^{-1}	0.1
Maximum value	0.03 ms^{-1}	$\min(\frac{1}{\max\{\lambda_{\mathbf{w}^*}\}}, 500)$
Number of samples	301	701

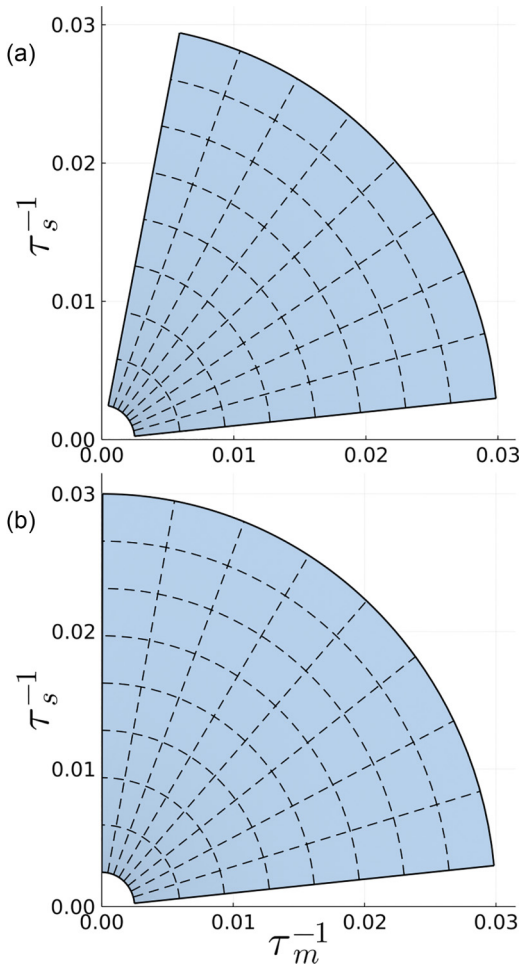

FIG. 2. Maximal sampling of the inverse timescale plane. We sample pairs of inverse-timescale values from the depicted region of the $\tau_m^{-1} - \tau_s^{-1}$ plane. (a) Sampled points are distributed evenly on a curvilinear grid between a predefined lower bound and the stability boundary for the specific connection matrix being used. (b) If sampling to the stability boundary would produce samples with negative timescales, the space is instead sampled up to a predefined value in the angular direction

or does not hit a convergence threshold before reaching the maximum number of steps.

All model manifolds studied here were generated from between 210 000 and 211 000 sampled parameter pairs.

III. isKL EMBEDDING

In this section, we recapitulate the methods developed by Teoh and colleagues [40]. This framework revolves around using the symmetric Kullback-Liebler divergence D_{sKL} as a measure of separation for different probabilistic models of the same form but with different parameters:

$$D_{sKL}(\theta,\theta') = D_{KL}(\theta:\theta') + D_{KL}(\theta':\theta) = \mathbb{E}_\theta\left[\ln\frac{p(x|\theta)}{p(x|\theta')}\right] - \mathbb{E}_{\theta'}\left[\ln\frac{p(x|\theta)}{p(x|\theta')}\right], \quad (14)$$

where x represents the (multidimensional) random variables. Teoh *et al.* apply this measure to exponential family models, which have the general form

$$p(x|\theta) = \exp\left[\sum_{i=1}^{n} t_i(x)\eta_i(\theta) + k(x) - A(\eta(\theta))\right],$$

where $\{\eta_i(\theta)\}$ are the n natural parameters of the model and $\{t_i(x)\}$ are the corresponding sufficient statistics. The D_{sKL} for exponential family models can be analytically decomposed into a *finite* number of component functions

$$D_{sKL}[\theta,\theta'] = \sum_{i=1}^{n}\{[\mathcal{T}_i^+(\theta) - \mathcal{T}_i^+(\theta')]^2 - [\mathcal{T}_i^-(\theta) - \mathcal{T}_i^-(\theta')]^2\}.$$

These component functions form a set of n spacelike ($\mathcal{T}_i^+$) and n timelike ($\mathcal{T}_i^-$) coordinates by which the model manifold may be embedded in a Minkowski-like model output space that is distinct from a data space used for embedding in methods like principal component analysis (PCA) [40]. The coordinate functions of this embedding are given in terms of the natural parameters and sufficient statistics by

$$\mathcal{T}_i^\pm = \tfrac{1}{2}[\eta_i(\theta) \pm \langle t_i(x)\rangle_\theta],$$

where $\langle\cdot\rangle_\theta$ is an average over the random variables x conditioned on the parameters θ. Alternatively, we may use an isometric embedding given by shifting and rotating the manifold [40]:

$$T_i^\pm(\theta) = \frac{1}{2}\left\{\lambda_i[\eta_i(\theta) - \overline{\eta_i}] \pm \frac{1}{\lambda_i}[\langle t_i\rangle_\theta - \overline{\langle t_i\rangle}]\right\}. \quad (15)$$

We use $T^\pm$ to distinguish the isometric embedding coordinates from the unscaled coordinates $\mathcal{T}^\pm$. Here, an over-bar $\overline{\cdots}$ denotes a mean over sampled parameters θ and $\lambda_i = [\mathrm{var}(\langle t_i\rangle)/\mathrm{var}(\eta_i)]^{1/4}$. These coordinates can be understood as an alternative definition of the exponential family. We can straightforwardly express the log-likelihood function for an exponential family in terms of the isKL coordinates:

$$\begin{aligned}\ln p(x|\theta) &= \ln k(x) + \sum_i \mathcal{T}_i^+(\theta)t_i(x) \\ &\quad + \sum_i \mathcal{T}_i^-(\theta)t_i(x) - A(\theta) \\ &= g(x) + \sum_i T_i^+(\theta)t_i(x) \\ &\quad + \sum_i T_i^-(\theta)t_i(x) - A(\theta),\end{aligned}$$

where $g(x) = \ln k(x) + \sum_i \overline{\eta}_i t_i(x)$. The authors of Ref. [40] show that the coordinates $\{T_i\}$ can also be understood in relation to the data visualization procedure multidimensional scaling (MDS). In standard MDS the data points are recorded data, whereas here each "data point" corresponds to the full distribution of an exponential family evaluated at a specific set of parameters. The doubly mean-centered matrix of MDS can be constructed in this context from the pairwise separation matrix measured by the symmetric KL-divergence $\mathbf{D}_c = -\mathbf{P}\mathbf{D}_{sKL}\mathbf{P}$ with $\mathbf{P}_{ij} = 1/n - \delta_{ij}$. In the continuous-sampling limit, the eigenvalue problem for MDS becomes an integral

equation:

$$\int D_c(\tilde{\theta}, \theta) v(\theta) d\rho(\theta) = \Lambda v(\tilde{\theta}), \tag{16}$$

where $d\rho(\theta) = \rho(\theta)d\theta$ is the measure of the distribution of parameters θ. Teoh and colleagues show [40] that the coordinates $T_i^{\pm}$ are solutions of this eigenvalue problem with corresponding eigenvalues

$$\Lambda_i^{\pm} = \frac{1}{2}[\mathrm{Cov}(\eta_i, \langle t_i \rangle) \pm \sqrt{\mathrm{var}(\eta_i)\mathrm{var}(\langle t_i \rangle)}]. \tag{17}$$

This procedure produces an embedding with only a finite and relatively small number of nonzero modes, contrasting sharply with the infinite or data-proportional embedding produced by other methods for continuous or discrete parameter sampling, respectively [40].

We complement this perspective by viewing this embedding procedure as an eigenmode expansion of the conditional probability $p(x|\theta)$ around the marginalized distribution $p(x)$ for a given prior on the parameters θ:

$$p(x|\theta) = p(x) + \sum_i c_i^+(x) T_i^+(\theta) + \sum_i c_i^-(x) T_i^-(\theta). \tag{18}$$

By defining the inner product of functions on Θ as

$$\langle f(\theta), g(\theta) \rangle = \int d\rho(\theta) f(\theta) g(\theta),$$

the modes $\sqrt{\mu(\theta)}v(\theta)$ of Eq. (16) can be shown to be orthogonal as long as the corresponding eigenvalues are distinct. Thus, the coordinate functions $T_i^{\pm}(\theta)$ are orthogonal with respect to the weight $\rho(\theta)$. Taking advantage of this orthogonality of the coordinate functions, it follows that

$$\begin{aligned}
&\int d\rho(\theta) p(x|\theta) T_j^{\pm}(\theta) \\
&\quad = p(x) \int d\rho(\theta) T_j^{\pm}(\theta) + \sum_{i,\pm} \int d\rho(\theta) c_j^{\pm}(x) T_j^{\pm}(\theta) T_i^{\pm}(\theta) \\
&\quad = c_j^{\pm}(x) \int d\rho(\theta) [T_j^{\pm}(\theta)]^2.
\end{aligned}$$

The first term on the right-hand side vanishes because the mean of each coordinate function is zero by construction, while only the $i = j$ term from the sum survives due to the orthogonality. Thus, we may calculate the coefficient functions $c_i^{\pm}(x)$ as

$$c_i^{\pm}(x) = \langle T_i^{\pm}, T_i^{\pm} \rangle^{-1} \int d\rho(\theta) p(x|\theta) T_i^{\pm}(\theta).$$

In this work we focus on applying these embedding methods to the stationary distributions of the various network models, both multivariate normal within our mean-field approximation. For a M-dimensional multivariate normal distribution with a set of means $\{\mu_i\}$ and covariance values $\{C_{ij}\}$, the $M(M+3)/2$ distinct natural parameters and sufficient statistics are given by

$$\eta = \begin{bmatrix} \sum_i C_{1i}^{-1}\mu_i \\ \vdots \\ \sum_i C_{Mi}^{-1}\mu_i \\ -\frac{1}{2}C_{11}^{-1} \\ \vdots \\ -\frac{1}{2}C_{M1}^{-1} \\ -\frac{1}{2}C_{22}^{-1} \\ \vdots \\ -\frac{1}{2}C_{MM}^{-1} \end{bmatrix}, \quad \langle t_i \rangle_\theta = \begin{bmatrix} \langle x_1 \rangle \\ \vdots \\ \langle x_M \rangle \\ \langle x_1^2 \rangle \\ \vdots \\ \langle x_M x_1 \rangle \\ \langle x_2^2 \rangle \\ \vdots \\ \langle x_M^2 \rangle \end{bmatrix}. \tag{19}$$

Before we present the embedding and analysis of the models from Sec. II, we provide two simpler models as illustrative examples that are related to the Poissonian and Gaussian characteristics of our model.

A. Example: One-dimensional exponential model

Let X be exponentially distributed with rate υ, i.e., $X \sim \mathrm{Exp}(\upsilon)$. In the exponential family formalism, we have

$$\eta = -\upsilon, \quad \langle t_i \rangle_\theta = \langle x \rangle_\upsilon = \upsilon^{-1}, \quad k(x) = 1,$$
$$A(\upsilon) = -\ln \upsilon.$$

The isKL embedding coordinates for this model are one-dimensional functions given by

$$T^{\pm}(\upsilon) = \frac{1}{2}\left\{\lambda[\overline{\upsilon} - \upsilon] \pm \frac{1}{\lambda}[\upsilon^{-1} - \overline{\upsilon^{-1}}]\right\}.$$

These embedding functions are shown in Fig. 3(a) using a parameter distribution $\rho(\upsilon) = (8\upsilon \ln 10)^{-1}$ with support $\upsilon \in [10^{-5}, 10^5]$ for illustration.

We may also explicitly calculate the coefficients $c^{\pm}(x)$ for this example,

$$c^{\pm}(x) = \langle T^{\pm}, T^{\pm} \rangle^{-1} \left[\left(\frac{\lambda}{2}\overline{\upsilon} \mp \frac{1}{2\lambda}\overline{\upsilon^{-1}} \right) \left(-\frac{dZ}{dx} \right) - \frac{\lambda}{2}\frac{d^2 Z}{dx^2} \pm \frac{1}{2\lambda} Z \right],$$

where $Z(x) \equiv \int e^{-\upsilon x} d\rho(\upsilon)$ is the moment-generating function of the distribution $\rho(\upsilon)$ with source $-x$. The full model manifold is depicted in Fig. 3(f), where points are colored by the logarithm of the rate parameter υ. Here, we see the manifold is neatly broken into two branches corresponding to a low event-rate ($\log_{10}\upsilon < 0$) and a high event-rate ($\log_{10}\upsilon > 0$).

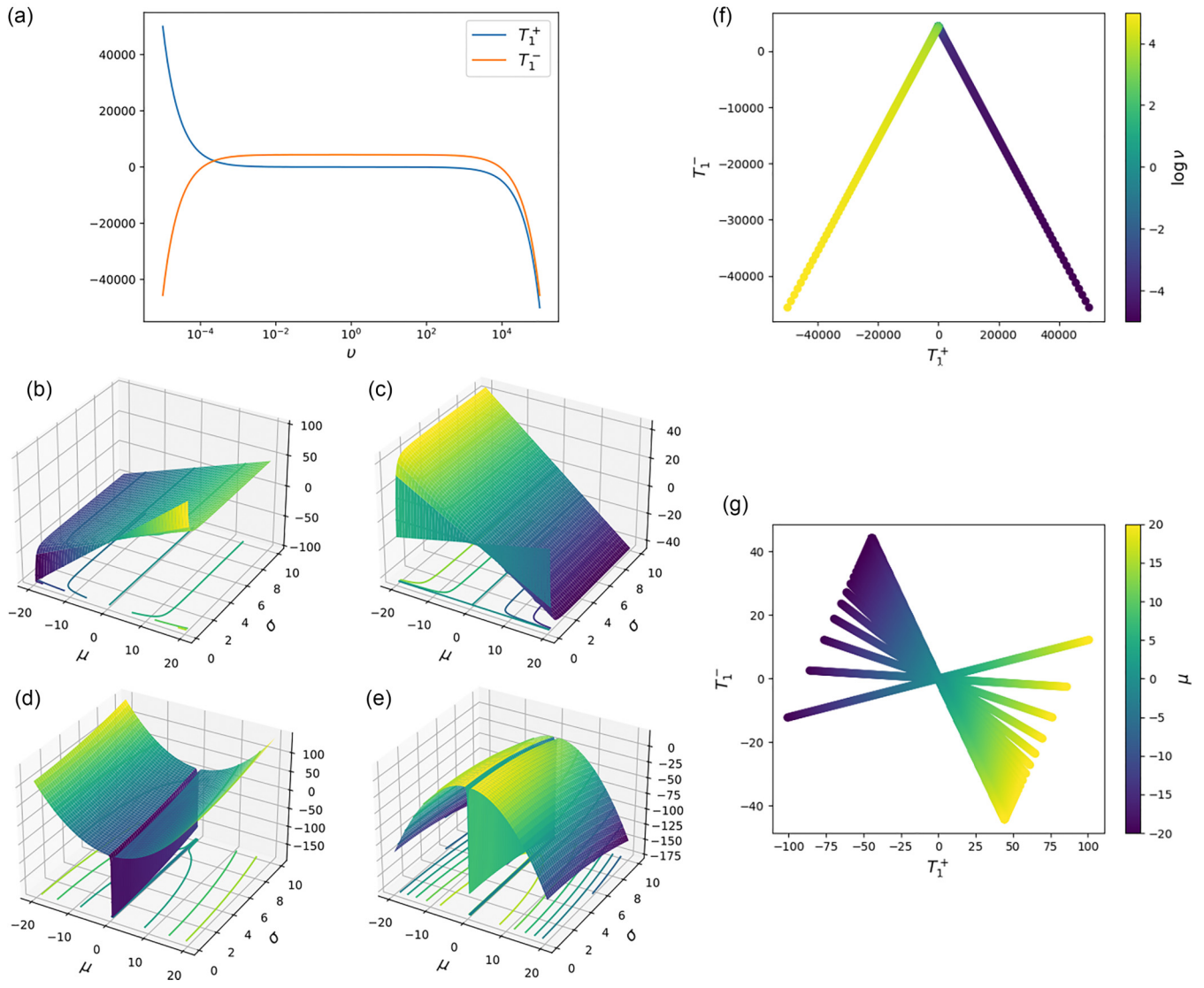

FIG. 3. Visualizations of the isKL embedding coordinate functions. (a) The two coordinate functions for the one-dimensional exponential model. (b)–(e) Coordinate functions for the one-dimensional Gaussian model: (b) T_1^+, (c) T_1^-, (d) T_2^+, (e) T_2^-. (f) The model manifold for the one-dimensional exponential model colored by $\log_{10} \upsilon$. (g) The model manifold for the one-dimensional Gaussian model projected onto just $T_1^\pm$, colored by μ

B. Example: One-dimensional Gaussian model

Let X be normally distributed as $X \sim \mathcal{N}(\mu, \sigma)$. In the exponential family formalism, we have

$$\eta = \begin{bmatrix} \mu/\sigma^2 \\ -\sigma^{-2} \end{bmatrix}, \quad \langle t_i \rangle_\theta = \begin{bmatrix} \langle x \rangle \\ \langle x^2 \rangle \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma^2 + \mu^2 \end{bmatrix},$$

$$k(x) = \frac{1}{\sqrt{2\pi}}, \quad A(\mu, \sigma) = \frac{\mu^2}{2\sigma^2} + \ln \sigma.$$

The isKL embedding coordinates are then two-dimensional functions:

$$T_1^\pm(\mu, \sigma) = \frac{1}{2}\left\{\lambda_1\left[\frac{\mu}{\sigma^2} - \overline{\left(\frac{\mu}{\sigma^2}\right)}\right] \pm \frac{1}{\lambda_1}[\mu - \overline{\mu}]\right\},$$

$$T_2^\pm(\mu, \sigma) = \frac{1}{2}\left\{\lambda_2\left[-\frac{1}{\sigma^2} + \overline{\left(\frac{1}{\sigma^2}\right)}\right] \pm \frac{1}{\lambda_2}[\sigma^2 + \mu^2 - \overline{(\sigma^2 + \mu^2)}]\right\}.$$

The Gaussian model coordinate functions are depicted in Figs. 3(b)–3(e) using a parameter distribution

$$\rho(\mu, \sigma) = \begin{cases} 1/800 & \text{if } -20 \leqslant \mu \leqslant 20, \quad 0 < \sigma \leqslant 20 \\ 0 & \text{otherwise.} \end{cases}$$

A projection of the model manifold onto the spacelike and timelike coordinates corresponding to first moment of the model is depicted in Fig. 3(g). The points on this projection

are colored by the mean parameter μ. Here, we see a degree of rotational symmetry in the manifold projection, separated into negative mean values on the left and positive mean values on the right of the T_1^+ centerline. Also note that there are apparent breaks in this manifold projection. These breaks do not reflect a true discontinuity in the structure of the manifold, but instead reflect the density with which the (μ, σ) space is sampled. We see manifold breaks related to the sampling density in our results for the network models.

IV. RESULTS

Before proceeding with results, it is helpful to briefly summarize the goal of this paper and the workflow constructed in prior sections. We wish to study the population-averaged behavior of stochastic spiking models as we vary synaptic and membrane timescales, repeating this across a range of relative excitation and inhibition. To do this, we approximate the full spiking network dynamics as a population-averaged multivariate Gaussian process [Eq. (6)]. We choose a sample of inverse timescales as discussed in Sec. II E, and in particular constrain the sampling based on the stability condition for the corresponding nonspiking model [Eq. (10)]. Within this sampled regime of timescales, the Gaussian process approximations should be mean-reverting and thus reach a stationary Gaussian distribution. We numerically solve for the vector-mean and the covariance matrix of the stationary Gaussian distribution at each sampled timescale-point. Finally, we embed this manifold of stationary Gaussian distributions into a behavioral space using the isKL methods introduced in Sec. III. With the workflow summarized, we may proceed.

A. Gaussian process approximations are stable

A key step in the analysis workflow is to find the stationary distribution for the approximated processes at each sampled timescale point. For the stationary distribution of an Ornstein-Uhlenbeck process to exist, all of the eigenvalues $\lambda_{\mathbf{A}}$ of the drift matrix $\mathbf{A}$ must have a positive real component. Basing the upper sampling boundary on the theoretical stability boundary of the related nonspiking linear model, as well as the check for singularities in the drift matrices $\mathbf{A}$ and covariance matrices $\mathbf{C}$, should ensure this requirement is met. We confirm this by explicitly examining the eigenvalues of the sampled models.

Each individual model—as specified by the model type, E-I $\log_{10}$-ratio R, and a pair of inverse timescales $(\tau_m^{-1}, \tau_s^{-1})$—has three drift-matrix eigenvalues. We pool together the eigenvalues of all particular models on a given model manifold as specified just by the model type and E-I log-ratio R. The resulting eigenvalue distributions for a subset of E-I conditions R are given in Fig. 4. Points in the eigenvalue distribution are colored by the log-distance of the particular model [specified by $(\tau_m^{-1}, \tau_s^{-1})$] from the origin in the inverse-timescale parameter space, shown in the inset. Each individual eigenvalue distribution is accompanied by marginal histograms where appropriate.

The subset of manifolds shown in Fig. 4 highlights a portion of the inhibition-dominated regime for which the models were observed to have complex eigenvalues. These complex distributions seen in the spiking-type model manifolds have a relatively small range in the imaginary direction and the imaginary components tend to pool near the origin in the along the real axis. Additionally, most of the density for these complex distributions are along the real axis itself, indicating that models with complex eigenvalues are relatively rare within their corresponding manifolds. The manifolds for the remaining E-I conditions have eigenvalue distributions qualitatively very similar to those at the extremes: purely real and positive eigenvalues spanning roughly the same range and skewed toward the origin. A key takeaway from Fig. 4 is that all of the eigenvalues have strictly positive real components, confirming that the sampled models for each manifold are stable and therefore appropriate for embedding analysis.

B. Behavior of full spiking network models

As discussed in Sec. III, the isKL embedding methods take the model manifold from the parameter space and position it in a hyperbolic space using the symmetric Kullback-Liebler divergence. The KL-divergence functions similarly to a distance between models based on their (sufficient) statistics which determine the statistical behavior of a particular model from the manifold. The isKL method thus embeds the model manifold in a behavioral space. This connection to the underlying output of the sampled models can be obscured when looking only at the results of the embedding analysis. As such, we take some time here to discuss the behavior of the full-network model described by Eq. (1).

Full-network spiking models were generated from the appropriate parameters in Table I with a sparse, random connection matrix as described in Sec. II A. A subset of model manifolds were chosen from across the range of E-I conditions R, and individual models from these manifolds were taken from along an arc of radius ≈ 0.01 (see, e.g., Fig. 5, left column). The membrane and spiking dynamics described by Eq. (1) with a specific choice of timescales were simulated using a basic forward-Euler integration scheme with a time step $dt = 0.1$ ms. Models were simulated for 20 000 ms, except the models in rightmost column, which were simulated for an increased duration of 150 000 ms to ensure convergence to a stationary regime.

The long-term dynamics of the population-averaged membrane potentials and the membrane potentials of six excitatory and six inhibitory neurons are shown in Fig. 5. The mean-field values of the membrane potentials predicted by Eq. (6c) are represented by the red and blue dotted lines. From this figure, we see that the population-averaged membrane potentials in the full spiking network do indeed reach stationary values as predicted by the drift eigenvalues λ_A for the approximated spiking models shown in Sec. IV A. Additionally, we see that the mean-field membrane potential values correspond fairly well to the stationary population-averaged potentials (Fig. 5, columns 1 and 2), but this breaks down near the upper bound of the arc (Fig. 5, column 3). This upper boundary of the arc corresponds to the stability boundary in first two manifolds (Fig. 5, rows 1 and 2) and a bifurcation boundary in the remaining manifolds. The breakdown of the mean-field approximations at these limits thus lines up with the colloquial understanding of their accuracy. Knowing that the population-averaged potentials of the full spiking networks converge to

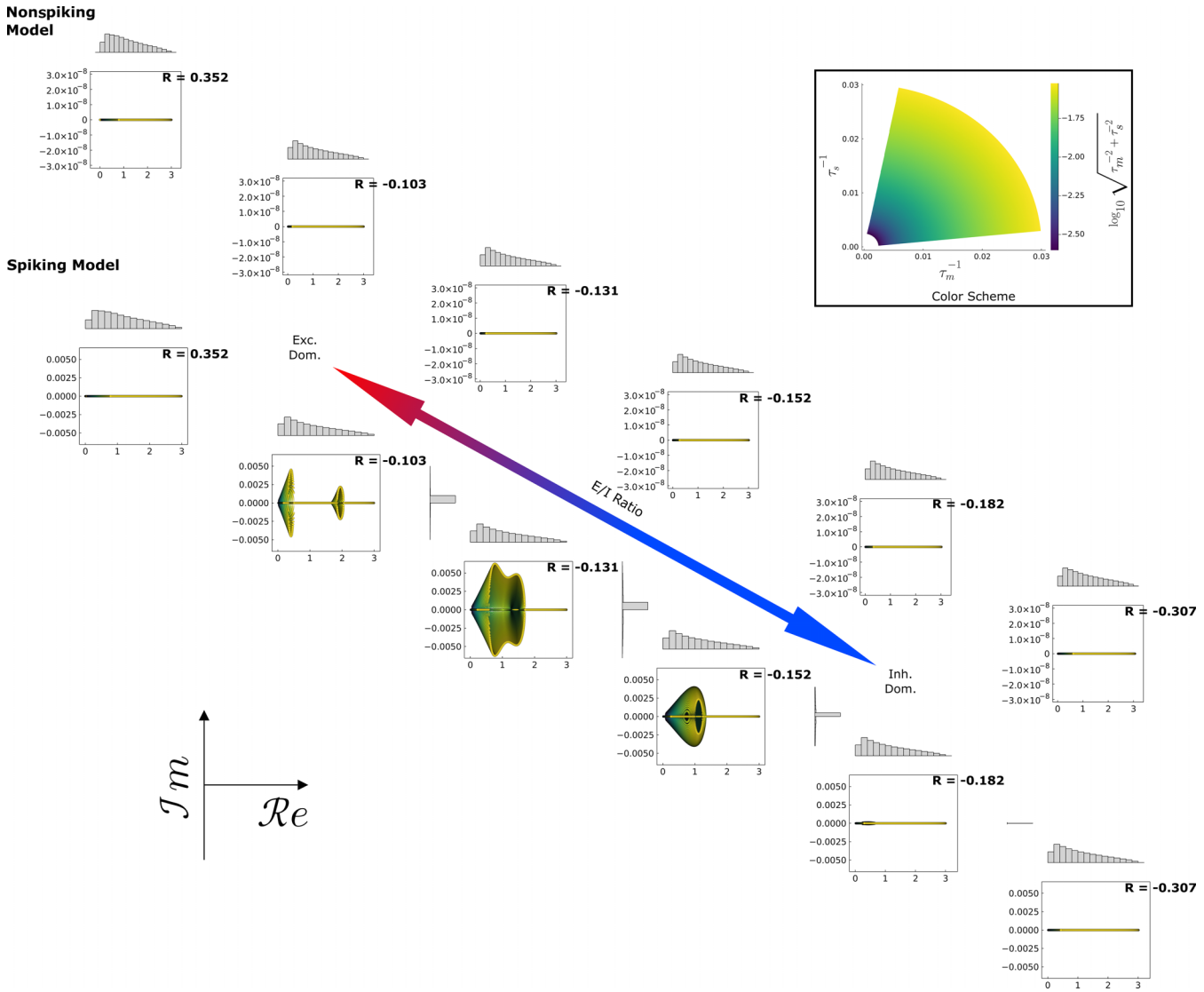

FIG. 4. Eigenvalues of drift matrices. Eigenvalue distributions for the drift matrices **A** of all sampled network models for the nonspiking model manifolds (top) and the spiking model (bottom) as the excitation-inhibition ratio R is adjusted. Points are colored by the radial distance d of model from the origin in inverse-timescale space, as illustrated by the inset. The marginal histograms for the real ($\mathcal{RE}$, top) and imaginary ($\mathcal{IM}$, right) are given for each distribution. Plots without a histogram in the imaginary dimension indicates a marginal δ distribution. Emphasis is placed on a portion of the inhibitory regime ($-0.19 < R < -0.09$, middle columns) in which some eigenvalue distributions display imaginary components.

a stationary condition, we next look at how these stationary solutions of these models differ.

The membrane potential dynamics in the stationary regime of these same models are shown in Fig. 6. Here, we show data from the last 1000 ms of simulation. We see that the population-averaged membrane potentials visibly fluctuate around an average value for most of the simulated models. These fluctuations in the population averages become less noticeable as the overall magnitude of the averages and standard deviations increase, e.g., along column 3 of Fig. 6. A similar trend is seen in the membrane dynamics for individual neurons in the network. Fluctuations in individual potentials are very large relative to their mean values for the first two models along the arc (Fig. 6, columns 1 and 2). The population-variance in membrane potentials for these models is thus highly dependent on the fluctuations of individual membrane potentials. Towards the upper end of the arc (Fig. 6, column 3), the magnitude of the membrane potentials increases and the fluctuations of individual potentials are less pronounced. For these models, the population-variance of the membrane potentials is much more dependent on the spread of individuals around the population mean as opposed to the fluctuations of those individuals.

We finish this section by examining the actual spiking dynamics for the example models in Fig. 7. Here we show raster plots for each model during the last 1000 ms of simulation. First, we see that the target neuron (black spikes) fires more frequently than other excitatory neurons in the network, which is to be expected because it receives extra current input. For each model manifold (different rows in Fig. 7), we observe an overall increase in the rate of spiking in the network as we move along the arc from point 1 to point 3. This aligns with

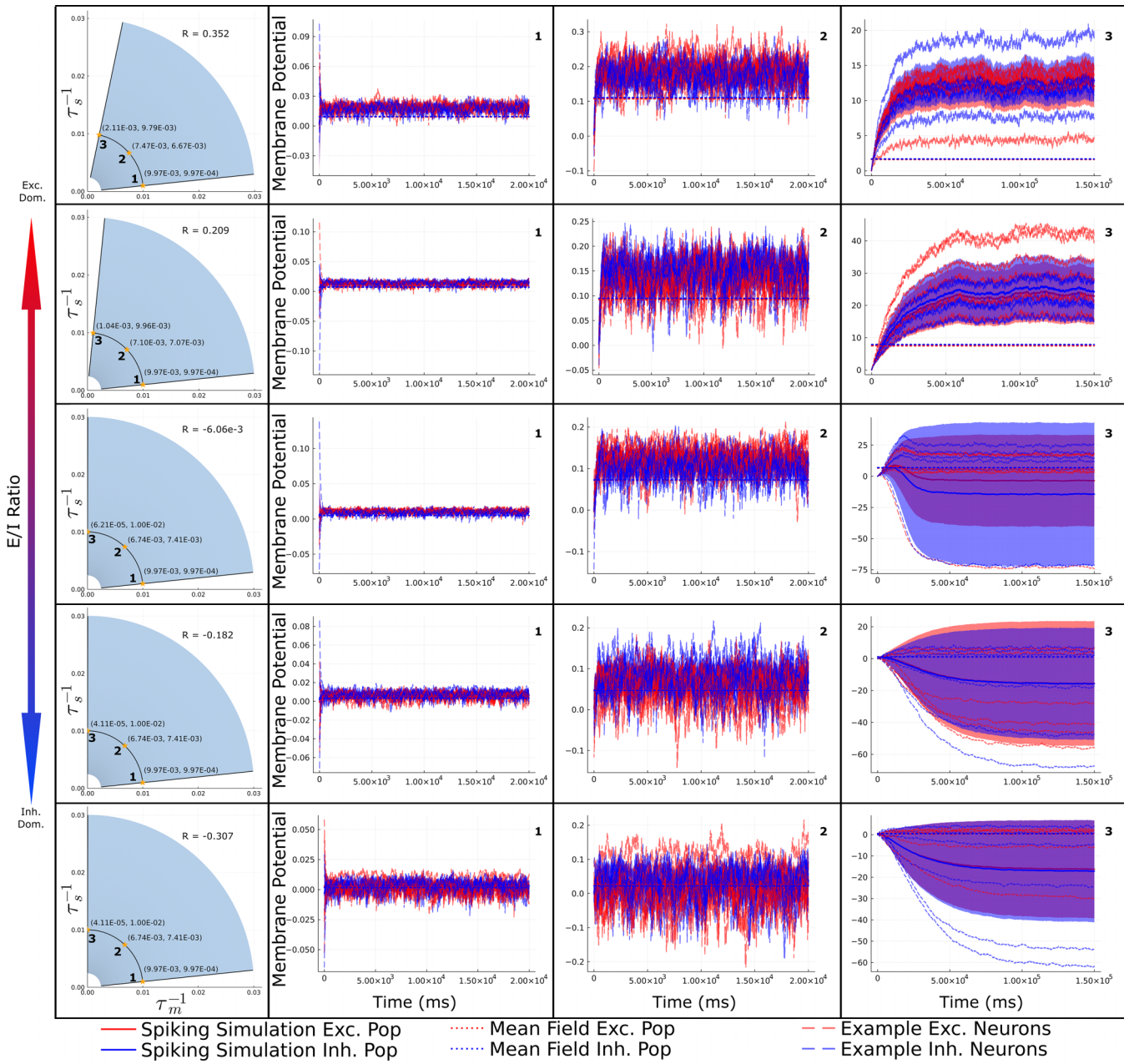

FIG. 5. Membrane potential dynamics of spiking network models. The dynamic population-averaged membrane potentials (solid lines) are plotted against the predicted mean-field values (dotted lines) for three different parameter pairs across the examined range of E-I ratios R. The membrane dynamics of six excitatory and six inhibitory neurons (dashed lines) are also shown for each condition. The shaded bands around the population-averaged potentials are the standard deviation of the membrane potentials within the corresponding population at each time point. The sampled parameter distribution from the embedding calculations, along with the chosen points for spiking simulation, are given in the rightmost column. Full-network spiking simulations were run until a stationary state was reached.

the change in overall magnitude of the membrane potentials seen in Figs. 5 and 6. This observation also accords with an intuitive understanding of the timescales: along this arc, the relative rate of synaptic input becomes much faster than the relaxation dynamics. This trend is taken to the extreme in the models at the top of the arc for each manifold (Fig. 7, column 3) where we see unrealistically high spiking rates in the last three rows. Finally, the overall spike rate of neurons decreases as the strength of the recurrent inhibition increases from the top row to the bottom row of Fig. 7, being most pronounced in the models with the most extreme behavior in the rightmost column. This aligns with intuition about the role of recurrent inhibition and the shifting of the stationary membrane potential values across the rows of Fig. 6.

Through the examination of these exemplary models from across the model manifolds of interest, we have built an intuitive understanding of the variation in model behavior. Having a sense of the interplay between the (inverse) timescales, the E-I modulation parameter R, and the stationary statistical behavior of the models in this way helps to ground the remaining results in a tangible set of observations. We now move on to the embedding analysis for these models.

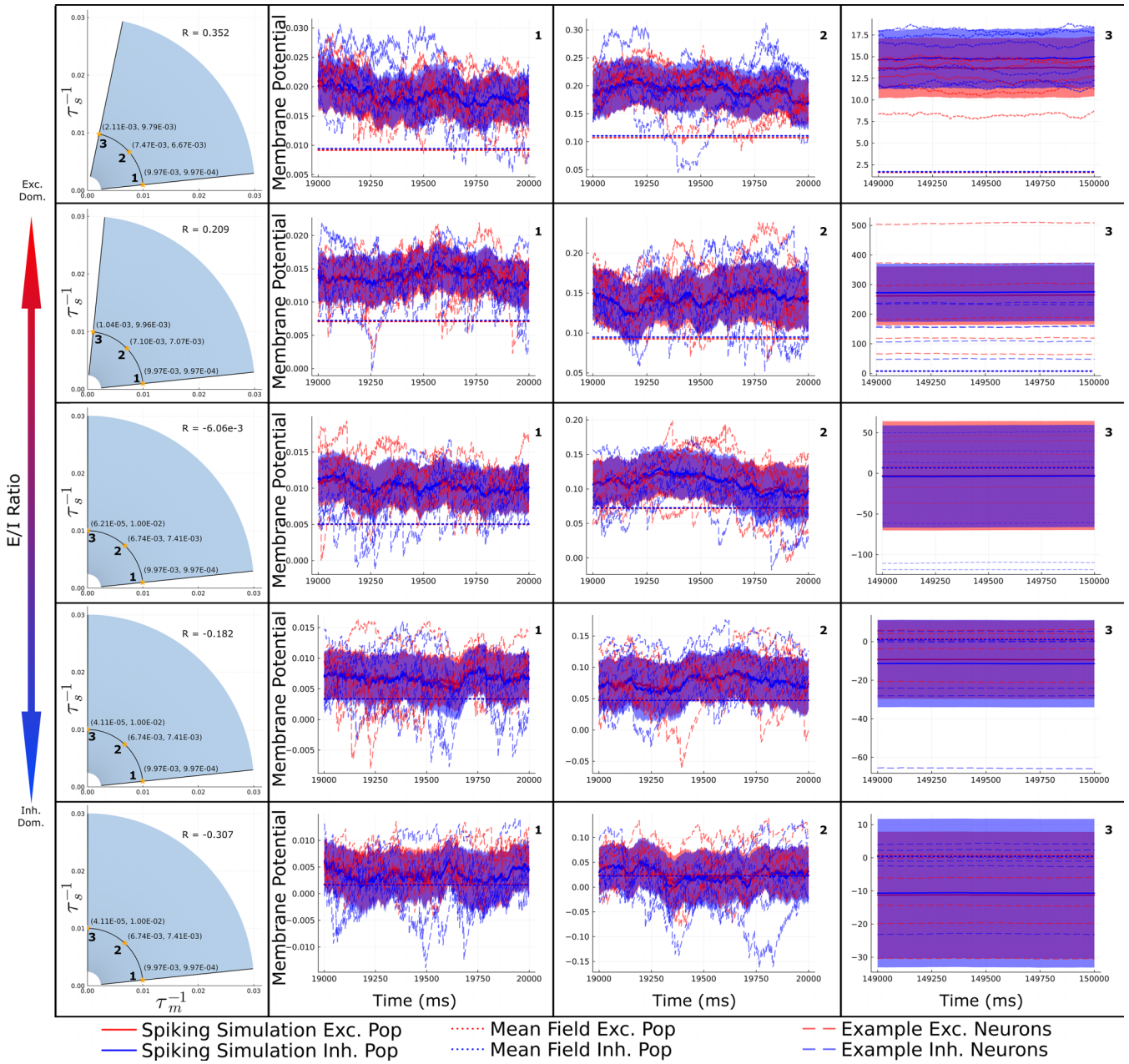

FIG. 6. Stationary membrane potential dynamics of spiking network models. The dynamic population-averaged membrane potentials (solid lines) are plotted against the predicted mean-field values (dotted lines) for three different parameter pairs across the examined range of E-I ratios R. The membrane dynamics of six excitatory and six inhibitory neurons (dashed lines) are also shown for each condition. The shaded bands around the population-averaged potentials are the standard deviation of the membrane potentials within the corresponding population at each time point. The sampled parameter distribution from the embedding calculations, along with the chosen points for spiking simulation, are given in the rightmost column. Full-network spiking simulations were run until a stationary state was reached, and the membrane dynamics for the last 1000 ms of simulation time are plotted.

C. Network embedding is hierarchical

It has been previously reported that biological models exhibit a hierarchy of sensitivities to different parameter combinations relative to some cost function on model behavior [38]. A similar hierarchical structure has been observed in the widths of model manifolds, which correlate strongly with the corresponding eigenvalues induced by a particular embedding [34,39]. The current modeling and embedding differs from these prior cases in that we are embedding probabilistic models in behavioral space. Considering also the limited dimensionality of the current embedding, it is unclear if this hierarchical property should manifest in the current system. We show below that the manifolds for models of the types in Eqs. (6) and (9) are indeed hierarchical under the isKL embedding framework, with coordinate eigenvalues spanning several orders of magnitude for each E-I condition.

We used the isKL methods (Sec. III) to embed the stationary distributions for both the spiking and nonspiking model

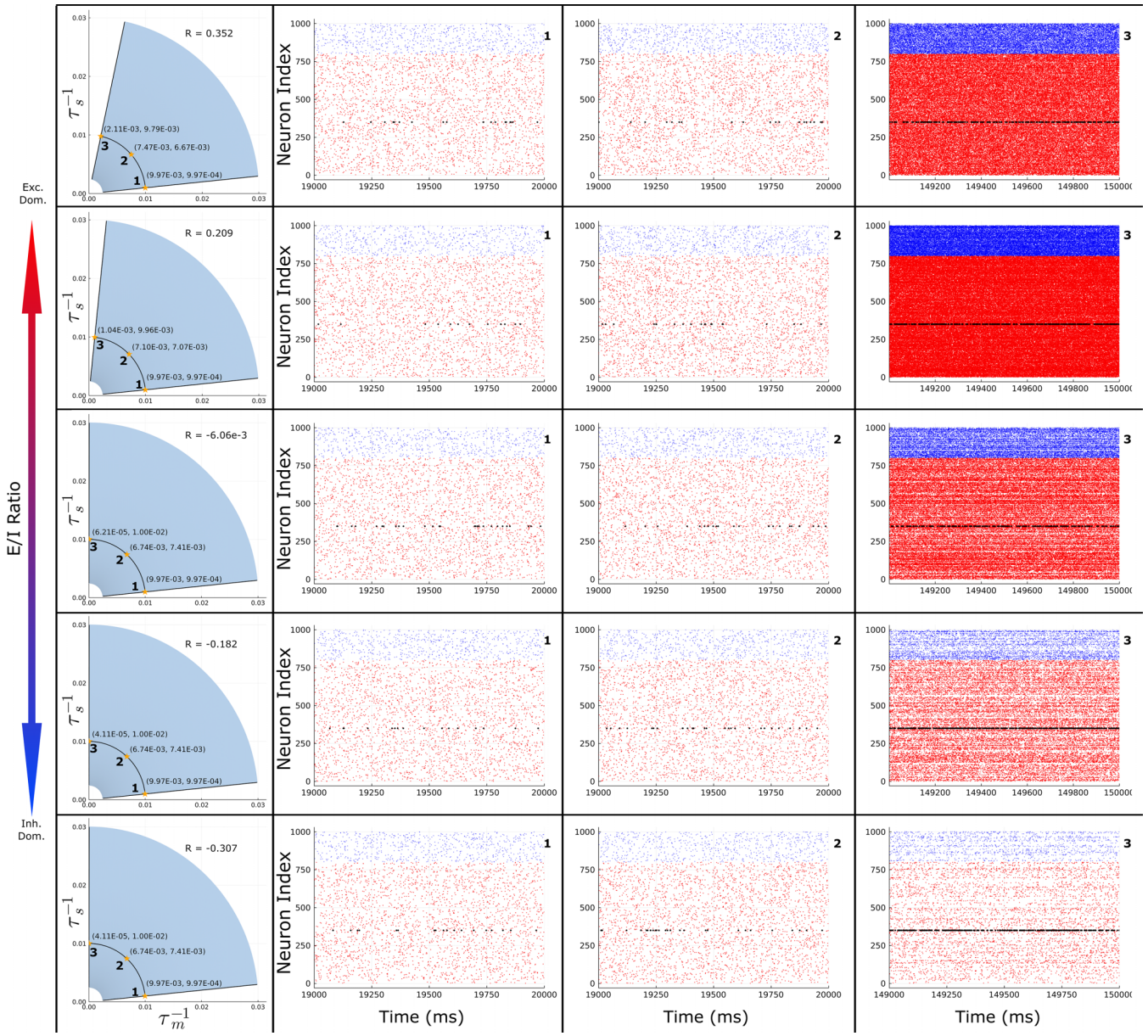

FIG. 7. Stationary spiking behavior of full network models. Example raster plots for individual timescale pairs for spiking models across the sampled range of E-I ratios R. The sampled parameter distribution from the embedding calculations, along with the chosen points for spiking simulation, are given in the rightmost column. Full-network spiking simulations were run until an apparent stationary state was reached, and the spikes from the last 1000 ms are plotted. In each model, neuron index 350 was designated as the target neuron and its spikes are shown in black.

types across 25 E-I conditions ranging from the excitation-dominated to the inhibition-dominated, and approximately centered at $R = 0$. The root absolute eigenvalues for the embedding coordinates $\{\Lambda_i^{\pm}\}$ are plotted against the observed manifold width along the same coordinate for the nonspiking models [Fig. 8(a)] and the spiking models [Fig. 8(c)]. Here, the manifold width is taken to be the simple range across a given coordinate. We see that the widths and eigenvalues are indeed correlated across E-I conditions for both model types, following with previous observations [34,39].This suggests these two measures may be used interchangeably in further analysis. The coordinate eigenvalues of the nonspiking model [Fig. 8(b)] span at least three orders of magnitude for any given E-I condition tested, and up to nearly fifteen orders of magnitude at the most extreme. The coordinate eigenvalues for the spiking models [Fig. 8(d)] span roughly two to three orders of magnitude on the extreme ends of the E-I spectrum and upwards of four in at some points in the center, with eigenvalues peaking towards the center as you approach from either extreme. Taken together, both model types studied here exhibit a hierarchical structure in line with prior observations of other systems, albeit with a more limited degree of separation in the case of the spiking-type models.

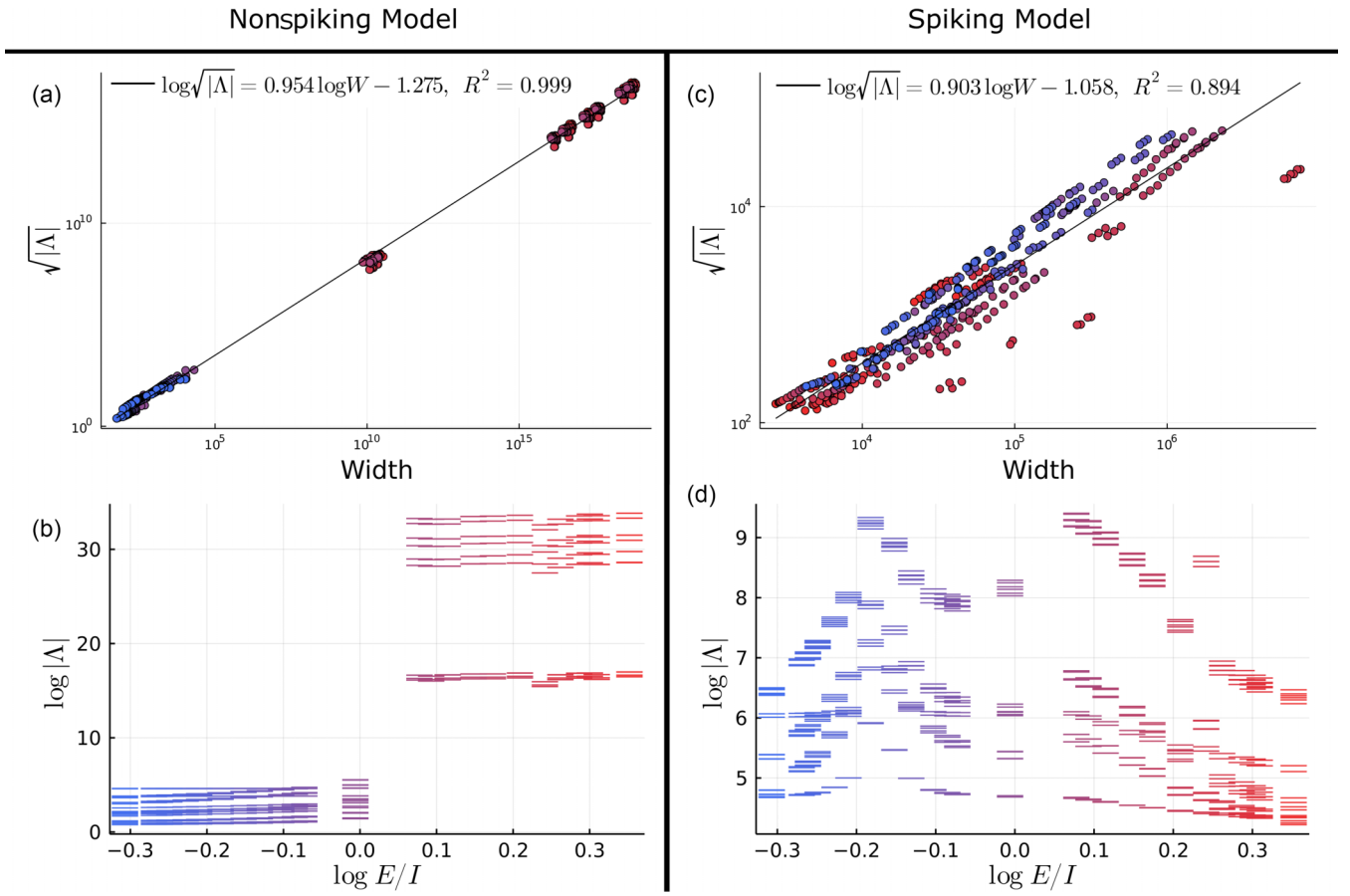

FIG. 8. Hierarchies of manifold widths. (top row) The correlation between the coordinate eigenvalues $\Lambda_i^{\pm}$ and the width W across the manifold in that coordinate direction as the E-I ratio is varied for the (a) nonspiking models and (c) spiking models. A simple linear regression is applied to the $\log_{10}$-widths and $\log_{10}$-absolute eigenvalues for visualization. The distribution of the scale of coordinate eigenvalues as the E-I ratio is varied for the (b) nonspiking models and (d) spiking models.

Before proceeding, we make some comparative observations between the two model categories. The scale and range of eigenvalues for the nonspiking model significantly larger than those for the spiking model type when in the excitation-dominated regime. Additionally, the nonspiking models show a sharp jump in eigenvalues when moving from the inhibition-dominated regime to the excitation-dominated one. This jump in eigenvalues may indicate a sort of bifurcation in the overall manifold. The eigenvalues $\{\Lambda_i^{\pm}\}$ directly reflect the covariance and—anecdotally more importantly—the variance in the corresponding sufficient statistics and natural parameters. A jump in the magnitude of the eigenvalues thus indicates a sudden increase in the variability of model activity, and this could correspond to sampling near the stability boundary [Eq. (10)] in the case of the transition seen the nonspiking models. A similar transition may be happening at the peaks in the eigenvalue distributions of the spiking-type models, however, these are much less pronounced than the one seen in the nonspiking models and the increase does not persist through the excitation-dominated regime as in the nonspiking models. We note that the firing rate nonlinearity for the spiking model-type (Table I) is asymptotically linear when $x \gg 1$. A naïve prediction would be a qualitatively similar eigenvalue distribution between model types when the membrane potentials become more positive as in the excitation-dominated regime. However, this is not reflected in the observed distributions of $\{\Lambda_i^{\pm}\}$.

D. Projection hierarchies

Having established that the isKL embedding of the spiking models and the nonspiking models exhibit a hierarchical structure, we next want to interrogate this structure across our model manifolds. We do this by examining projections of the manifolds onto lower-dimensional spaces along the largest widths and smallest widths. We focus on two-dimensional projections.

Figure 9 shows the largest manifold projections in behavioral space for the nonspiking models and spiking models across a subset of E-I conditions. Points on these manifolds are colored by the mean value of the membrane potential for the test neuron $\langle V_0 \rangle$. It is visually clear that the manifold projections are shrinking from top left to bottom right for each condition, reflecting the hierarchical structure of the manifold. A large fraction of projections—for example, Fig. 9 column 3, row 2—have apparent gaps in their structure. These are similar to the gaps seen in the projection of the example Gaussian distribution shown in Fig. 3(g) and are tied to the sampling density used across the inverse-timescale space near key boundaries (data not shown).

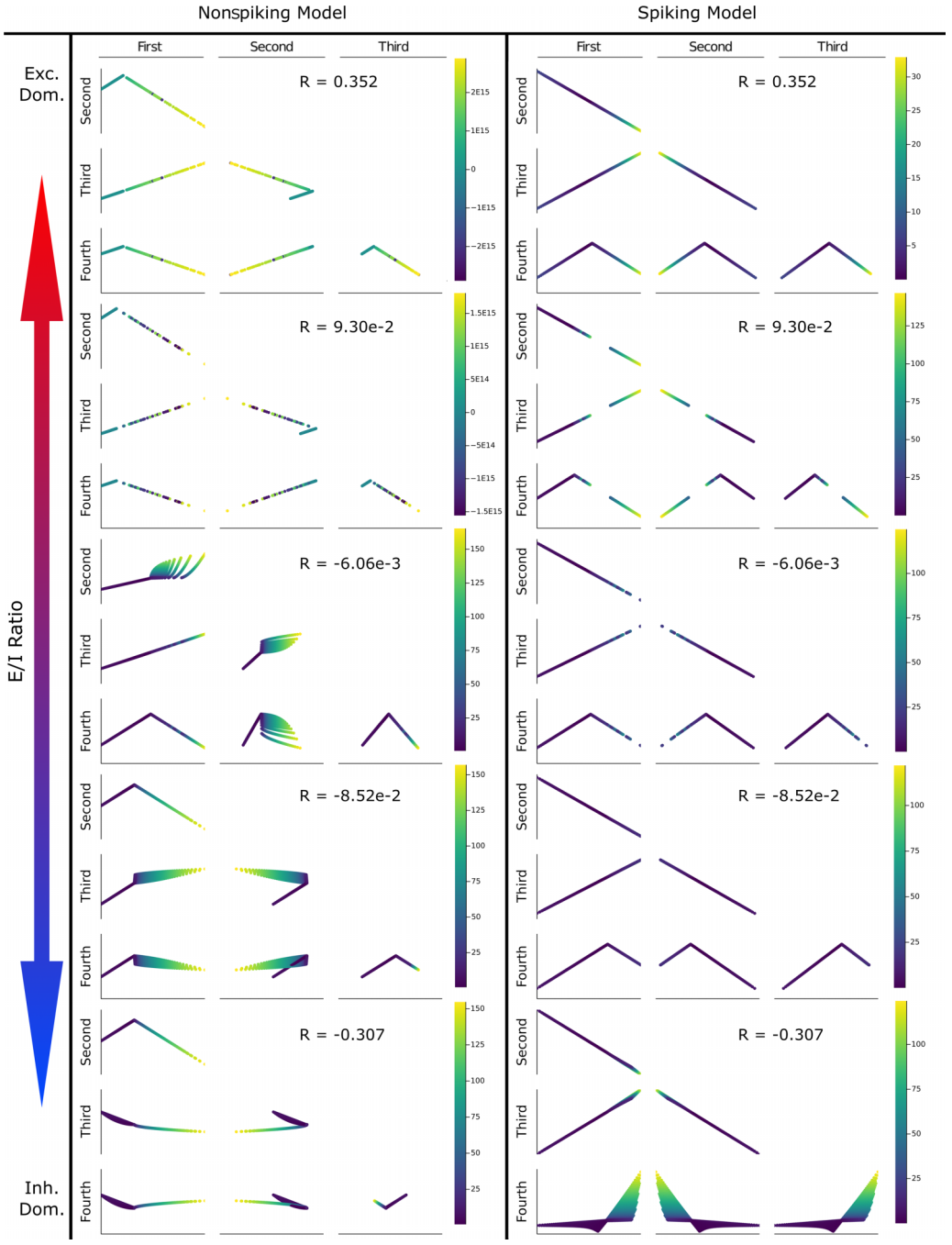

FIG. 9. Largest manifold projections. Projections of the manifolds for the two model types onto the largest four coordinates as determined by the observed manifold width. The hierarchy of projections is shown as the excitation-inhibition ratio is changed from excitation dominant regime (top) to an inhibition-dominant regime (bottom). Manifolds are colored by $\langle V_0 \rangle$—the mean membrane potential of the test neuron—and each projection is scaled by the largest observed width. These projections are the stiffest, contributing the most to the behavior of the distribution of activities.

Many of the projections across model types and E-I conditions appear very linear or piece-wise linear, for example, Fig. 9 column 2 rows 1–4. This thinness at the largest scales would suggest a relatively simple relationship between the largest coordinates and that the model manifold is relatively flat. The difficulty of overcoming under-sampling of the parameter space complicates this interpretation slightly. The gaps in the projections seen in the excitation-dominated regime are clear evidence of some under-sampling, but interpolating the data across gaps suggests that the projections in these conditions may still be piecewise- or quasilinear. These stick-like projections both model-types in the excitation-dominated regime. The projections are also seen to change shape qualitatively as the E-I conditions change. In the nonspiking models, we see the appearance of spoon-shaped projections as we move into the inhibition-dominated regime. In contrast we see knife-like projections in the spiking models, albeit only under the most inhibitory of E-I conditions. This

qualitative change in the manifold projections seen across the two model types could be caused either by warping of the manifold along each coordinate as the conditions change or by changes in ranking of each coordinate. This point will be revisited in Secs. IV E and IV G.

Additionally, we note that many of the projections separate points on the manifold according to the value of $\langle V_0 \rangle$, as was seen in the example embeddings shown in Figs. 3(f) and 3(g). This is particularly clear, for example, in the inhibition-dominated regime of the two model types (Fig. 9, row 5). The separation of the manifold into sections based on behavioral regimes depends on more than just $\langle V_0 \rangle$, however. For example, we see no such trend in Fig. 9, column 2, row 4. In this case the individual models within the manifold are more significantly separated by (a combination of) statistical parameters that, in a sense, have a dependence on the timescale parameters that is orthogonal to the way $\langle V_0 \rangle$ depends on the timescales. Exceptions aside, this noted $\langle V_0 \rangle$-aligned separation serves as a clear demonstration of the behavioral clustering induced by the isKL method.

We can also examine the smallest projections of the model manifolds for the two model types, which correspond to the least important modes of the expansion in Eq. (18). Figure 10 shows the smallest projections of the model manifolds for all three model types across E-I conditions. As was the case for the largest coordinate projections, there is evidence of undersampling in the smallest projections also. This particularly evident in the nonspiking model manifold in the excitation-dominated regime (Fig. 10, column 1, rows 1 and 2). In contrast with the largest projections, the stick-like projections comprise the minority of the small-projection shapes. The model manifolds appear instead to be highly curved at the fine-grained level. Following the observation from the large-scale projections, we see the smallest manifold projections separate points in line with $\langle V_0 \rangle$. In particular, the counterexample mentioned before (Fig. 9, column 2, row 4) now also shows a degree of alignment with changes in $\langle V_0 \rangle$, now in Fig. 10, column 2, row 4. This highlights the fact that model separation along different directions on the manifold can be more or less tied to a particular behavioral or statistical parameter.

Before moving on, we should highlight the relationship between the projection coordinates and model behavior. Coordinates with larger eigenvalues contribute more to separation between $p(x|\theta)$ and $p(x|\theta')$ as measured by the D_{sKL}. More specifically, a relatively large absolute-eigenvalue $|\Lambda_i^{\pm}|$ indicates that the corresponding natural parameter is a relatively better way to separate individual models on the manifold by their behavioral predictions, or alternatively that a larger part of the variance of behavioral predictions across the manifold are explained by the associated natural parameter. Given a tractable mapping between the underlying model parameters and the natural parameters, the magnitude of coordinate eigenvalues can also give a sense of the relative importance of different parameter combinations aligning with a given coordinate direction. Lastly, the high degree of correlation between coordinate eigenvalues and the manifold widths along those coordinates means that the relative size of a particular projection gives a visual representation of the importance of the corresponding parameter combination.

E. Coordinate rankings

We saw in Sec. IV D that the projections hierarchies of the model manifolds changed across the examined range of E-I conditions R. One possible explanation for this is that the rankings of coordinates by manifold width change with R. This potential aspect of the changing projections can be interrogated by tracking the rank of each coordinate across the range of R. This will additionally provide insight into what aspects of the statistical model have the greatest (or least) impact on the overall model behavior for both model types. Figure 11 depicts the ranking for each coordinate by both the length of the manifold along that coordinate (top row) and eigenvalue-magnitude (middle row) for the spiking and nonspiking model types for a subset of E-I conditions. As each coordinate corresponds directly to a single sufficient statistic, we color code and shape code the rank of each coordinate according to this correspondence. The bottom row of Fig. 11 reproduces the eigenvalue distributions shown in Fig. 8, except each point is now color and shape coded according to the sufficient statistic instead of the E-I measure R.

We see in Fig. 11 that the ranking of coordinates changes across E-I conditions for both the nonspiking and the spiking models. We also note that the rankings by observed width (top row) and by eigenvalue-magnitude (middle row) agree fairly well across the range of R. This agreement between the two sets of rankings makes sense when considering the high degree of correlation between the coordinate eigenvalues and manifold widths shown in Fig. 8. It is interesting that the eigenvalue distributions—particularly those for the spiking models—seem to separate into clusters of coordinates that do not intersect for much of the range of R. For example, the tan-orange-steel blue (eight point star-five point star-hexagon) cluster at the top of the eigenvalue spectrum corresponds to the second moments involving the inhibitory and excitatory populations. This cluster remains consistently above the gray-magenta (downward triangle-pentagon) and pink-red (diamond-square) clusters—which correspond to the second moments involving the target neuron and first moments for the bulk populations, respectively, across R for the spiking models.

Knowing that the eigenvalues and thus the eigenvalue-magnitude rankings form these clusters across the E-I spectrum, it is natural to examine the sufficient statistics that correspond to coordinates in these clusters. We focus on the spiking-type models that exhibit these clusters. The tan-orange-steel blue (eight point star-five point star-hexagon) cluster noted before dominates over other clusters in the spiking models, and these coordinates correspond to the second-order moments of the membrane potentials of the excitatory and inhibitory populations (Fig. 11, columns 2 and 3). This indicates that they are the most important statistics for distinguishing models across the manifolds. These second moments are also important for the nonspiking model manifolds, but they only sit at the top of the hierarchy in the excitation dominated regime $R > 0$. The gray-magenta (downward triangle-pentagon) cluster in the spiking models corresponds to the mixed second moments involving the target neuron V_0 and either V_1 or V_0. This cluster is above the green (triangles) cluster in the mid-range of R and just below

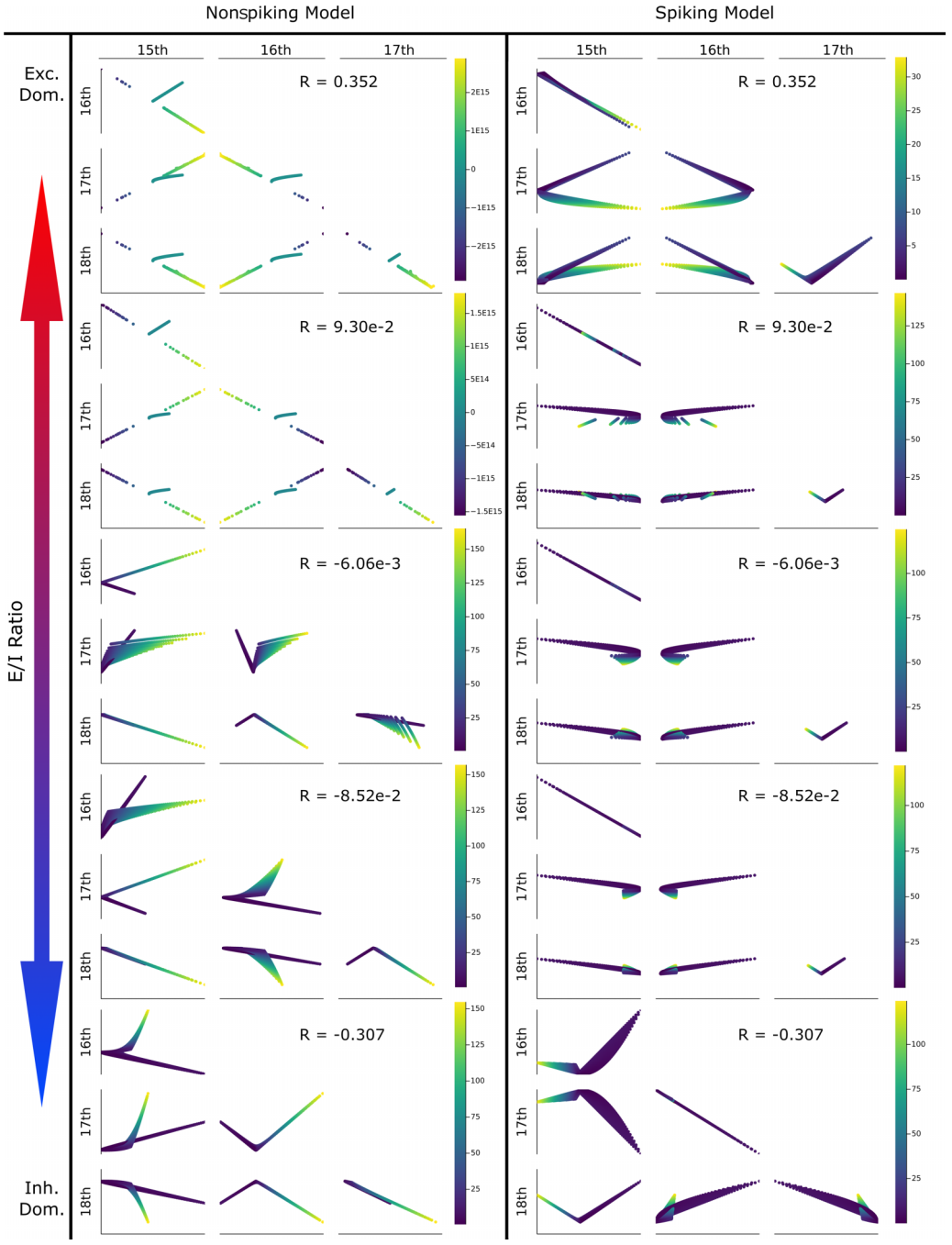

FIG. 10. Smallest manifold projections. Projections of the model manifold for the two model types onto the smallest four coordinates as determined by the observed manifold width. The hierarchy of projections is shown as the excitation-inhibition ratio is changed from excitation dominant regime (top) to an inhibition-dominant regime (bottom). Manifolds are colored by $\langle V_0 \rangle$, and each projection is scaled by the observed width along the coordinate ranked $(M-3) = 15$. These projections are the sloppiest, contributing the least to the behavior of the distribution of activities.

it in the extreme E-I conditions, and this green (triangles) cluster is the pure second moment $\langle V_0 \rangle$. The green cluster is generally above the pink-red (diamonds-squares) and the blue (circles) clusters, except for a brief crossing of the pink-red and the green clusters around $R \approx -0.18$. These last two clusters correspond to the mean values of all three membrane potentials. Taken together, these observations say that for the spiking-type models the fluctuations are more important for distinguishing between individual models on a given manifold and—for both the first and second moments—the statistics that involve the test neuron are generally less important than those that do not. These observations hold for the nonspiking model manifolds in the excitation-dominated regime, but not in the inhibition-dominated regime (Fig. 11, column 1).

Let us discuss the coordinate rankings at a more granular level of detail. While both model types have the second moments at the top of their respective hierarchies in the excitation-dominated regime, it is interesting to note how

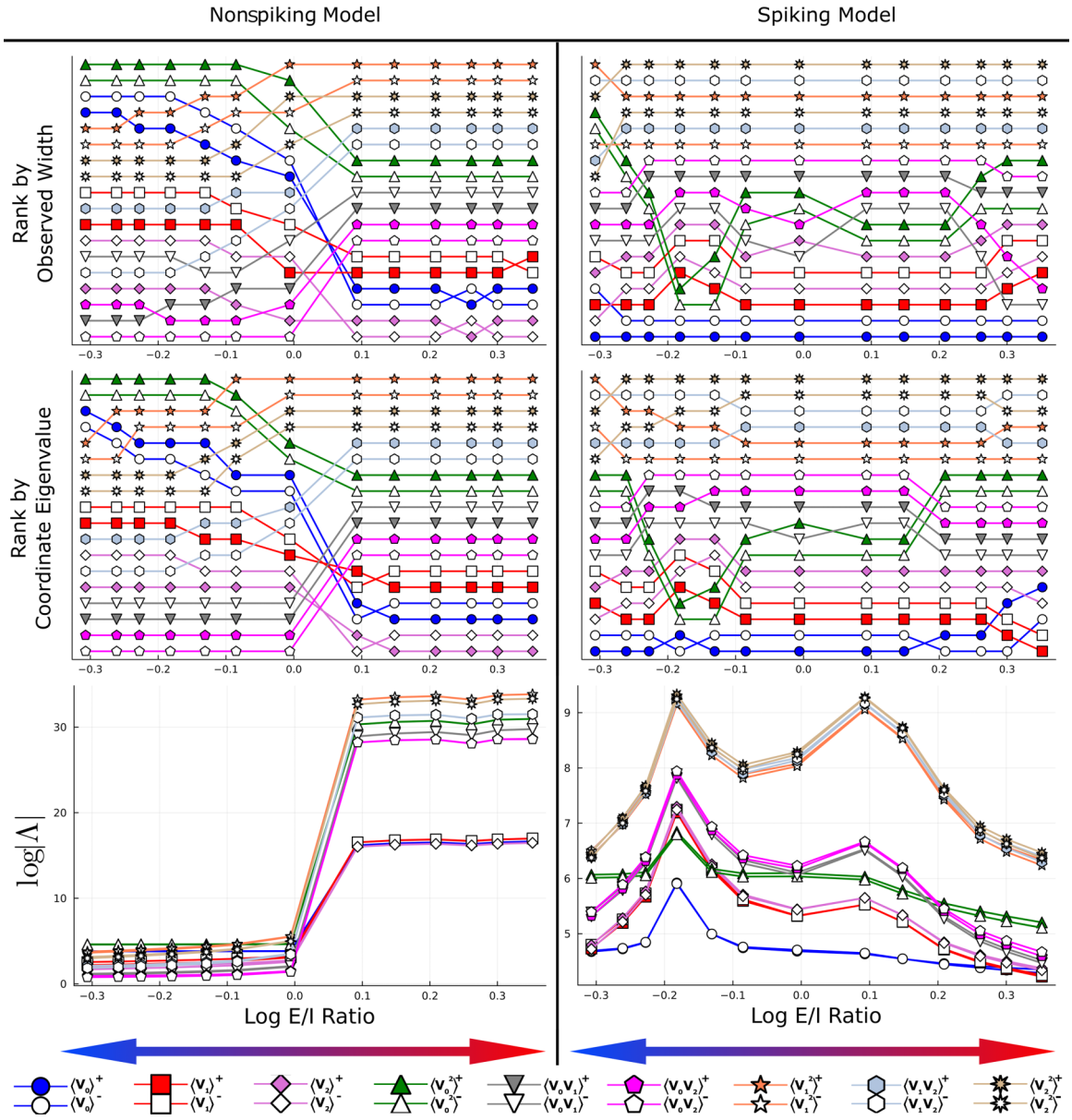

FIG. 11. Coordinate rankings. The ranking of each manifold coordinate $T_i^{\pm}$ as the E-I balance is changed in both model types. Coordinates are ranked from most important (top of each plot) to least important (bottom of each plot) based on the observed width of the manifold along said coordinate (top row) or the magnitude of the corresponding eigenvalue $|\Lambda_i^{\pm}|$ (middle row). The $\log_{10}$-magnitude of the eigenvalue for each coordinate is given in the bottom plot as the E-I balance R is changed as in Fig. 8. The legend renames each coordinate $T_i^{\pm}$ to the corresponding sufficient statistic $\langle t_i \rangle^{\pm}$ for ready interpretation.

they differ here. The spiking-type models have the $\langle V_2^2 \rangle^+$-related coordinates at the top while the nonspiking model is topped by the $\langle V_1^2 \rangle^+$-related coordinates. This suggests that the degree of fluctuations in the inhibitory population are the most varied for the spiking models in this regime, but the excitatory population fluctuations take that title in the excitation-dominated nonspiking models. The last fine-grained detail we highlight here is the increased importance of the V_0 moments in distinguishing the behavior of the inhibition-dominated nonspiking models relative to their importance in the inhibition-dominated spiking-type models.

In addition to visualizing the relative importance of certain sufficient statistics for distinguishing between particular models across a given model manifold, we get another piece of information visualized for free through the eigenvalue-magnitude ranking plots (Fig. 11, middle row). Recall from Eq. (17) that the eigenvalues $\Lambda_i^{\pm}$ are given by the covariance of the sufficient statistic and natural parameter across the manifold $[\mathrm{Cov}(\eta_i, \langle t_i \rangle_\theta)]$ and the geometric means of their individual variances $[\sqrt{\mathrm{var}(\eta_i)\mathrm{var}(\langle t_i \rangle_\theta)}]$. As we know, the Λ_i^- eigenvalues are negative and of the same order of magnitude as the corresponding Λ_i^+ (Fig. 11, row 3), we know that geometric mean of those variances greatly outweighs their covariance. Furthermore, the relative ranking of Λ_i^+ and Λ_i^- imply the sign of the covariance $\mathrm{Cov}(\eta_i, \langle t_i \rangle)$: if $\Lambda_i^+ > \Lambda_i^-$ then $\mathrm{Cov}(\eta_i, \langle t_i \rangle_\theta) > 0$ and *vice versa*. For example, by

looking at the eigenvalue-magnitude ranking of the steel blue (hexagon) coordinates in the spiking-type models (Fig. 11, row 2, columns 2 and 3) we see that $\mathrm{Cov}(-\frac{1}{2}C_{12}^{-1}, \langle V_1 V_2 \rangle) < 0$ across all E-I conditions examined here. While this could very easily be determined by looking at these covariances themselves—and they must be calculated in order to determine $\Lambda_i^{\pm}$—it is convenient to be able to glean this from a plot already produced for another purpose.

We will briefly summarize. Figure 11 shows that the coordinate rankings do change across the sampled E-I range, thus explaining the changing projection hierarchies in Figs. 9 and 10 at least in part. We found that the coordinates form clusters in the eigenvalue distribution that behave in a correlated manner across the E-I spectrum and with which they share relations to similar types sufficient statistics. In particular, the cluster of coordinates corresponding to the fluctuations $\langle V_I V_J \rangle^{\pm}$ for $I, J \in \{1, 2\}$ have the most impact on the activity of the spiking-type models, as well as in the excitation-dominated nonspiking models. We made observations of which types of fluctuations were most important to model distinction across the manifold for different model-types and different E-I conditions. Finally, we highlighted a secondary visual interpretation of the eigenvalue-magnitude ranking plots relating to the base statistical model.

F. Transforming base parameters

We highlighted in Sec. IV E that the statistical parameters from the stationary Gaussian distribution of membrane potentials—discussed in terms of the sufficient statistics—have a hierarchical impact on the possible behaviors of the spiking model that changes across the E-I spectrum R. Furthermore, we identified clusters of parameters that tend to change in similar ways with R. In the case of the spiking models, the fluctuations $\langle V_I V_J \rangle^{\pm}$ for $I, J \in \{1, 2\}$ were the most impactful while the mean coordinates $\langle V_0 \rangle^{\pm}$ had a relatively small impact. While important, these observations do not directly address the role of the inverse timescales $(\tau_m^{-1}, \tau_s^{-1})$ on model behavior. Unfortunately, the mapping from the timescale parameters to statistical parameters is intractable, owing primarily to the transcendental system of mean-field equations [Eq. (6c)]. Closed forms for the stationary distribution parameters of the linear nonspiking models can be found, but these expressions are ratios of high-degree polynomial functions of the timescales and do not directly reflect the mapping in the spiking model context. To begin untangling the impact of the timescale parameters on the range of model behaviors, we must thus rely on a qualitative understanding of the relationship between the timescales and, e.g., the sufficient statistics.

In Fig. 12, we plot the logarithm of several sufficient statistics as a function of position in inverse-timescale space for the spiking network with $R = -6.06 \times 10^{-3}$. We include $\langle V_1^2 \rangle$ [Fig. 12(b)] and $\langle V_2^2 \rangle$ [Fig. 12(c)] from the upper cluster as well $\langle V_0 \rangle$ [Fig. 12(a)] to serve as a representative set from across the hierarchies in Fig. 11. Note that the second moments have different units than those of $\langle V_0 \rangle$, which should be kept in mind when comparing the color scales. That said, the D_{sKL} between two members of the same exponential family

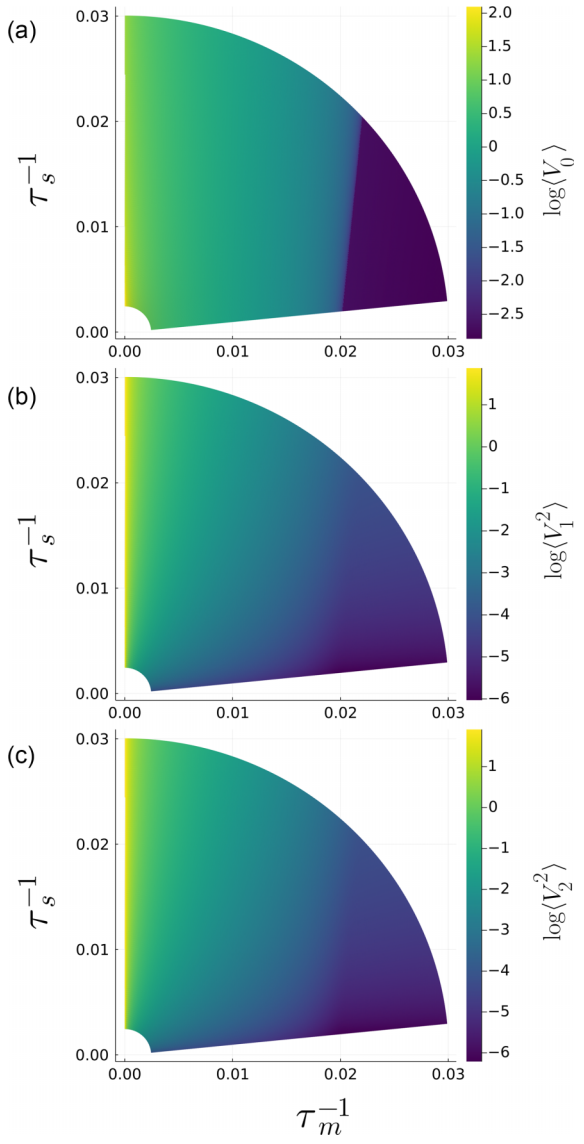

FIG. 12. Mapping from inverse timescales to statistical parameters. The relationship between the inverse timescales $(\tau_m^{-1}, \tau_s^{-1})$ and a select set of statistical parameters from the corresponding stationary Gaussian distribution for the spiking network with $R = -6.06 \times 10^{-3}$. Inverse timescale spaces are colored by the $\log_{10}$-value of one of the sufficient statistics: (a) $\log_{10}\langle V_0 \rangle$, (b) $\log_{10}\langle V_1^2 \rangle$, (c) $\log_{10}\langle V_2^2 \rangle$.

can be rewritten as [40]

$$D_{sKL}(\theta, \theta') = \sum_i [\eta_i(\theta) - \eta_i(\theta')][\langle t_i(\theta) \rangle - \langle t_i(\theta') \rangle].$$

Paired with the sufficient statistics of a multivariate normal distribution [Eq. (19)], we see that the D_{sKL} is in some sense weighing the first and second moments directly against each other. This in mind, the variability in the second moments is $\approx$2 orders of magnitude larger than that for the mean of the test neuron, in line with their relative ranking in Fig. 11. Furthermore, we note the similar dependence of all three statistical parameters on the inverse timescales, increasing in magnitude radially from the τ_m^{-1} axis to the τ_s^{-1} axis as well as exhibiting a "cold spot" triangle on the rightmost corner

of the sampled wedge. The trends between the means and covariances differ most significantly along the τ_s^{-1} boundary. Here, the magnitude of the mean increases towards the inverse-timescale origin (i.e., very long timescales) while the second moments increase moving away from the origin (i.e., very short timescales).

The presence of the cold spot in the mappings to the statistical parameters—particularly the sharpness of the transition seen for $\langle V_0 \rangle^{\pm}$—reinforce the intuition that translating changes in the statistical parameters back onto the timescale parameters is nontrivial. That said, the shared general trend in the mappings suggest a possible avenue for model reduction if some loss of expressivity is permitted. Reducing the two-dimensional sampled space to an arc around the origin and through the cold spot could be used to capture the concomitant increases in the magnitude of the first and second moments, capturing the majority of their respective variability. Alternatively, radial sampling along the τ_s^{-1} boundary could be used to study the apparent trade-off in magnitude of the means and covariances. This idea of model reduction is intimately tied to notions of model dimensionality, which we return to in Secs. IV H and V.

G. Manifold projections change smoothly with E-I balance

We now return to the question raised at the beginning of Sec. IV E: What causes the projection hierarchies to change across E-I conditions? While the changing coordinate ranking observed across E-I conditions for both models can explain the changing manifold projections, it does not rule out the possibility that the projections along a given coordinate are themselves changing. To address this possibility, we project the model manifolds for both of the model-types onto the same pair of coordinates across the E-I spectrum in Fig. 13. We chose to project the manifold onto the spacelike $\langle V_0^2 \rangle^+$ and $\langle V_0 \rangle^+$ coordinates because the statistical behavior of the test neuron may be of particular interest in some scenarios. For the sake of visualization, each projection along the E-I spectrum is scaled by the larger of the two manifold projections at each condition. The overall scale of the projection is given by the axis scale.

We can see in the projections of the nonspiking model manifolds (Fig. 13, upper diagonal) that there is a squashing and stretching of the manifold relative to the overall change in scaling as the E-I conditions are changed. Additionally, these transformations appear to act smoothly on the manifold projections until the manifold flattens going from the inhibition-dominated regime to the excitation-dominated regime in the range $-6.06 \times 10^{-3} \leqslant R \leqslant 0.115$. This flattening reflects a radical increase in the manifold scale along the $\langle V_0^2 \rangle^+$ coordinate relative to the $\langle V_0 \rangle^+$ coordinate because all of the eigenvalues are seen to jump (Fig. 11, column 1, row 3). This interpretation is corroborated by the change in overall scale of the axes—from $\approx O(10^3)$ for $R < 0.115$ to $\approx O(10^{17})$ for $R \geqslant 0.115$ (Fig. 13, upper diagonal)—and the correlation between eigenvalue magnitude and manifold width discussed in Sec. IV C (Fig. 8). The projections of the spiking manifolds (Fig. 13, lower diagonal) are also seen to transform smoothly with R with the fork-shaped projections (e.g., lower diagonal, $R = -0.307$) collapsing into the spoon projections (e.g., lower diagonal, $R = -0.103$) seen in the nonspiking model around $R \approx -0.1$. The projections for the spiking-type manifolds do change along R in line with the changes in their respective eigenvalue distributions (Fig. 11, columns 2 and 3, row 3), but these changes are more subtle than in the nonspiking model. The eigenvalue distributions for the spiking-type models drift downwards as we move from $R < -0.18$ to $R > 0.10$, and this is mirrored in the manifold projects by a slight decrease in projection scale moving in the same direction.

We have shown here that the manifold projections for both model-types do indeed change across the sampled E-I range, which plays a subsequent role in the changing of projection hierarchies across E-I conditions noted in Sec. IV D. The individual projections were shown to undergo potentially significant rescaling across values of R that alter it visually, as noted in the nonspiking manifold. Additionally, the manifold can exhibit a warping, as in the fork-spoon-fork transition noted in the spiking-type models.

H. Manifold dimensionality

As mentioned at the outset (Sec. I), a key issue when analyzing the spike responses of a collection of large spiking networks is the dimensionality of the behavioral output space. The true behavioral space of the model—as expressed through the spiking activity—grows with both an increasing network size N and a decreasing time bin size Δt. A goal of the current work is to understand the behavioral output of these models in a lower-dimensional framing. Thus, we briefly interrogate the dimensionality of our spiking and nonspiking models—as they are expressed through the various steps of our analysis—before proceeding with the final discussion.

The dimensionality of model output following Eq. (1) is $2NT/\Delta t$ for a discrete-time trial of length T, the combined size of the two matrices containing the membrane and spiking processes for each neuron at each discrete point in time. The dimensionality of the output space is reduced to $NT/\Delta t$ when moving to the Gaussian process approximation of the model in Eq. (5) because the explicit spiking dynamics are effectively marginalized away. Population-averaging these approximated dynamics to Eq. (6) further decreases the dimensionality to $N_{\text{pop}}T/\Delta t$, where $N_{\text{pop}} < N$ is the number of populations being considered. By simplifying our analysis to studying the *stationary distribution* of population behaviors, the dimensionality drops to $N_{\text{pop}}(N_{\text{pop}}+3)/2$, corresponding to the maximal number of independent sufficient statistics (see Sec. III). Finally, the isKL methods embed this manifold of population activity into an $N_{\text{pop}}(N_{\text{pop}}+3)$-dimensional space, which determines the upper limit of dimensionality that may be measured from embedded data (i.e., sampled models). To summarize, we have traced the maximal dimensionality of the model output through the various simplification steps before talking about the dimensionality of the model manifold, demonstrating how the dimensionality of the manifold is tied to the dimensionality of the output space of the model.

Having considered how the dimensionality changes across the steps for our analysis, two key questions remain: (i) If we can only see the results of the embedding, how do we gauge the dimensionality of the manifold being analyzed? (ii) If we

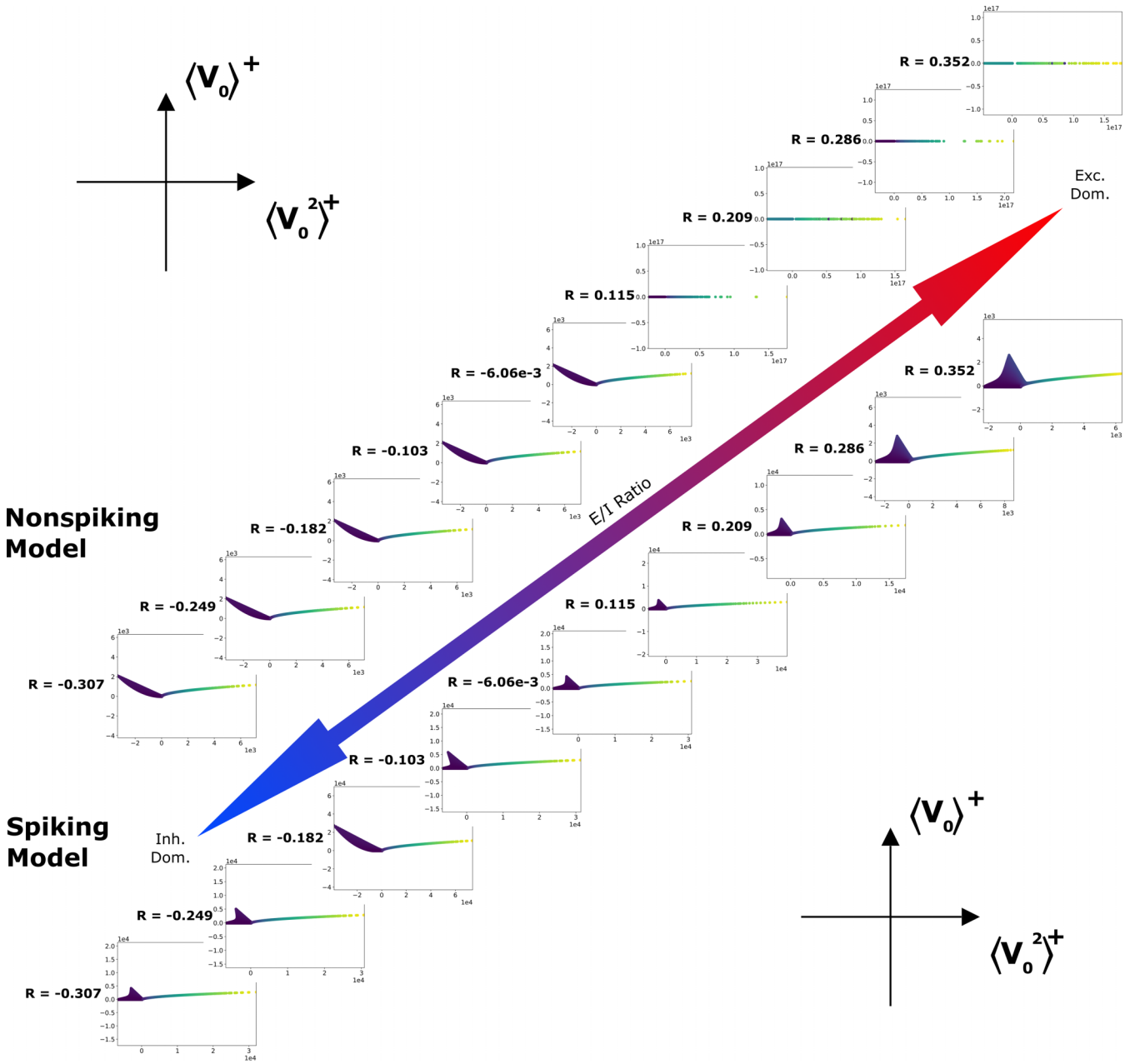

FIG. 13. Coordinate evolution as E-I balance is tuned. Projections onto the $\langle V_0^2\rangle^+ - \langle V_0\rangle^+$ plane of the nonspiking model manifolds (top) and the spiking model manifolds (bottom) as the excitation-inhibition ratio is adjusted. The axes of each two-dimensional projection are scaled to the larger width for the given model and E-I condition for the convenience of visualization. Note that the projections change in size by an order of magnitude as the E-I balance is adjusted. The projections are colored by $\langle V_0\rangle$ as in prior figures.

instead have an understanding of the maximal dimensionality of the system, is there any effective reduction in dimensionality that we can measure? To address these questions, we start with a measure of effective manifold dimensionality commonly used in principal component analysis (PCA) known as the participation ratio (PR):

$$\mathrm{PR} = \frac{\left(\sum_i \Lambda_i\right)^2}{\sum_i \Lambda_i^2}.$$

As the isKL embedding methods are intimately tied to multidimensional scaling (MDS)—an extension of PCA—the PR should serve as a useful base for measuring the effective dimensionality of our embedded model manifolds. This is complicated slightly by the presence of negative eigenvalues $\{\Lambda_i^-\}$ that arise in MDS, so we use an altered PR as our measure of effective dimensionality:

$$\mathrm{PR} = \frac{\left(\sum_{i,\pm} |\Lambda_i^\pm|\right)^2}{\sum_{i,\pm} (\Lambda_i^\pm)^2}. \tag{20}$$

The effective dimensionality of our two model types across the examined E-I spectrum is shown in Fig. 14. We see that the spiking-type models begin with a relatively high PR≈8 in the inhibition-dominated regime before dropping to PR≈6 in the middle regime and then rising slightly again in the excitation-

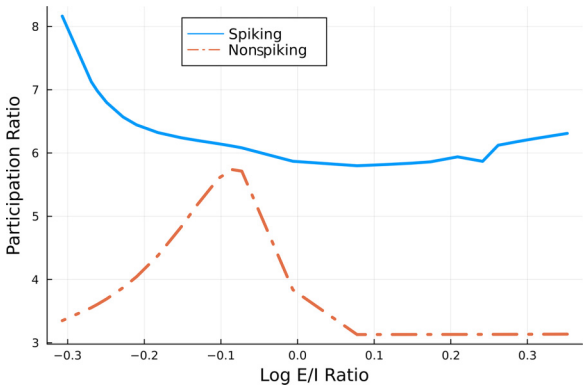

FIG. 14. Dimensionality of model manifolds. The effective dimensionality of nonspiking and spiking models across the spectrum of E-I conditions is given. Effective dimensionality is given by the altered participation ratio given by Eq. (20).

dominated regime. By contrast, the nonspiking model has PR ≈ 3 in the excitation-dominated regime. The PR for the nonspiking model then peaks at PR ≈ 6 around $R \approx -0.1$ before decaying back down further into the inhibition-dominated regime. The nonspiking models thus have a lower effective dimensionality than the spiking-type models.

How do we contextualize the measured PR for these models? First, we note that the maximal possible measured dimensionality for both the spiking and nonspiking model types is $N_{\text{pop}}(N_{\text{pop}}+3) = 18$, and the *statistical model* dimensionality is $N_{\text{pop}}(N_{\text{pop}}+3)/2 = 9$. This indicates that the approximate models show a dimensionality reduction compared with both the model dimensionality and the maximal embedding dimensionality. This seems trivial until one examines the PR measure for the example embeddings given in Sec. III. The simple Poisson example has just one parameter Λ and subsequently a maximum embedding dimension of two. Despite the intrinsic parameter density of one, its isKL embedding [Fig. 3(f)] gives an effective dimensionality is much closer to its maximal embedding dimension and gives PR ≈ 1.982. In a similar vein, the example Gaussian model has two parameters (μ, σ), yet it has an embedding dimension of PR ≈ 3.929 which is nearly its maximum embedding dimensionality of four [see Fig. 3(g) for one of the manifolds projections]. The measured PR thus does not seem to reflect the dimensionality of the intrinsic manifold structure, but instead the number of embedding dimensions within the isKL framework required to capture most of the variability in model behaviors. This will be discussed further in Sec. V.

V. DISCUSSION

The central motivation of this paper is to tease apart the impact of cellular and synaptic model parameters—internal timescales and relative synaptic strengths, respectively on the complex and high-dimensional behavioral space of spiking network models. Taking inspiration from prior applications of information geometry to neural systems [25–32], we approached this Herculean task by leveraging recently developed methods for studying the information geometry of complex biology models [39,40] and applying them to spiking network models with more biological features than those considered previously. We began by defining our spiking model [4,57] and then simplifying it through population-averaging, using a path-integral formalism to approximate the membrane dynamics as a Gaussian process [4,62], and then calculating the stationary distribution for that approximation [64]. The stationary distributions for these were then analyzed using the information geometric framework introduced by Teoh and colleagues [40]. This workflow is the core of the work presented.

Before diving into the results of the geometric embedding analysis, we briefly examined the behaviors of full spiking networks across various E-I conditions and for a few different timescale points. We showed that the behavior of the actual networks change distinctly across the variables at both the level of spiking and of population-averaged membrane dynamics. Importantly, the spiking models reach a stationary state in the long-time limit. This agreed qualitatively with the mean-field predictions and supported the analysis of the stationary distributions from the reduced model.

The information-geometric analysis demonstrated that the approximated models are hierarchical. Manifold widths and coordinate eigenvalues spanned several orders of magnitudes, pointing to a hyperribbon structures with stiff and sloppy coordinate directions. The distribution of these coordinate eigenvalues changed across E-I conditions and with it the hierarchy of two-dimensional manifold projections. These changes in the manifold projections arose from a smooth warping of projections onto specific coordinate pairs as well as a reordering of the coordinate rankings. Identifying each coordinates with their corresponding sufficient statistic highlighted a clustered structure in the eigenvalue distribution of the spiking models across E-I conditions. From this clustered structure, it is possible to pick out the most and least important sufficient statistics for distinguishing between models on a given manifold—these are the parameter combinations that underlie the stiff and sloppy coordinate directions, respectively. In particular, the stiffest directions on the spiking-model manifolds corresponded to the second moments of the excitatory and inhibitory population membrane potentials while the sloppiest directions were those corresponding to the first moments. This suggests that bulk fluctuations are key for determining the behavior of a specific network. It is unfortunately difficult to tie this understanding of stiff and sloppy statistical parameters to the timescale parameters in a manner that is satisfactorily analytical, owing primarily to the transcendental mean-field equations [Eq. (6c)] that must be solved numerically. That said, the implication of the sloppy and stiff coordinate observations is that an adjustment of the membrane and synaptic timescales tends to have a larger effect on the large-population fluctuations than it does on the means.

At the end of our isKL analysis, we discussed the dimensionality of the models, their behavior, and their manifolds. The largest reductions in the size of the model occur when moving to population-averaged models and when focusing on the stationary distribution. The combined effect decreases the dimensionality of the model output space being studied from $NT/\Delta t$ to $N_{\text{pop}}(N_{\text{pop}}+3)/2$, in which we essentially shift from a study of particular spike patterns to a study of probability distributions. From here, the isKL methods em-

bed the distribution in $N_{\text{pop}}(N_{\text{pop}}+3)$ dimensions. This sets the upper limit of dimensionality at the end of our analysis, being 18 for the particular population-averaged networks studied here. Using a modified participation ratio to measure the effective dimensionality of our embedded spiking models gave us a range of $6 \lesssim \text{PR} \lesssim 8$—depending on the E-I measure R—less than the maximum possible dimension. It was illustrated elegantly through the example embedding of the one-dimensional Poisson model that the participation ratio measures the number of dimensions needed to hold a sufficiently representative version of the model manifold rather than the intrinsic dimensionality of the manifold. In fact, the participation ratio for both toy models indicated that they basically "filled" their respective embedding spaces. Taken together, these show that approximated spiking models are definitely undergoing a degree of dimensionality reduction because they are not filling the embedding space like the toy models did. The participation ratio can then be interpreted as giving a sense of how "pointed" a change in parameters is. If a six-dimensional space is needed to represent most features of the manifold, this likely implies that the modulated parameters are mostly affecting three natural parameters. However, this is a measure of the effect of base parameter (i.e., τ_m and τ_s) on model output and does not necessarily reflect a minimal structure in the base parameter space needed for nearly full expressivity of the model.

From the copious stick-like projections seen in the hierarchies (Figs. 9 and 10), we may intuit that the embedded manifolds are of an even smaller dimension than is represented by the participation ratio. We can take this a step further by understanding the entire embedding process as a transformation of a manifold originally in the parameter space, implying that it should intrinsically be, at most, two-dimensional. We also noted in Sec. IV F that there are *ad hoc* ways of reducing the parameter space to a one-dimensional curve while seemingly preserving much of the variability in statistical parameters. If the goal is to find a reduced number of base parameters to approximately cover the manifold in a more principled way, this would likely require estimating the intrinsic dimensionality of the model manifolds with more sophisticated tools than those discussed here. This would provide a number of parameters—or parameter combinations—needed to understand and express the model. Thus, a combination of both an intrinsic measure and the participation ratio provides a complimentary understanding of model manifold dimensionality through the lenses of necessary base parameters and range of impact, respectively.

It is important to note that the properties of the spiking model manifolds are likely to differ depending on inputs and other contexts, such as networks that perform specific tasks like encoding sensory input. The models studied here are functionally in a spontaneous regime with external tonic drives that maintain the network in a balanced state when the excitatory and inhibitory synaptic inputs are balanced. The structure of a given task is known to collapse high-dimensional spontaneous activity into a lower-dimensional space [11,17], which might be seen directly in information-geometric interrogations such as the one performed in this paper. Furthermore, this may well affect which statistical parameters are important, in turn changing the coordinate rankings, projection hierarchies, and potentially even the degree to which the resulting manifolds are hierarchical. It is certainly possible to study the model manifolds as a function of external inputs to the network, rather than intrinsic network properties, using the methods outlined in this work. These possibilities require their own analysis in follow-up work.

The relationship between the hierarchical structures seen in these models and the hyperbolic nature of the isKL embedding methods has a deeper connection to prior work than initially mentioned in Sec. I. It has been shown that an underlying hyperbolic geometry can embed the exponentially expanding branches of hierarchical tree-like processes nearly isometrically, and conversely that tree-like processes can be thought of as a discretization of the hyperbolic space [65]; this has been taken to mean that an underlying hyperbolic geometry implies a fundamentally hierarchical structure and *vice versa*. The initial study of this relationship was meant to provide a geometric description of complex hierarchical networks from earlier work [66], but the scope of application expanded from there. With a combination of topological tools to match data correlations to an optimal underlying geometry and hyperbolic embedding methods, the observed hyperbolicity was interpreted as evidence of the existence of hierarchical tree structures in the correlations of natural odors and their perceptual organization [54], used to visualize and understand gene expression patterns across cell types and through the process of differentiation [55], and provided a mechanistic understanding of place representation and the influence of experience on those representations in mouse hippocampal CA1 spiking patterns [56]. The natural pairing of the hyperbolic isKL embedding methods to exponential family models [40] thus fundamentally implies an underlying hierarchical structure per these findings [65]. Conversely, this connection implies the hierarchical structure of many biological models from the sloppy modeling literature [33–40] may be understood further in terms of an underlying hyperbolic geometry.

A. Limitations

It is important to highlight some limitations of the framework of modeling and analysis in this paper. The primary hurdle to using these methods is the fact that the base calculations required for each step, combined with the number of samples needed to visually resolve the embedded manifolds, make it costly to increase the dimensionality of the parameter space or the number of network elements. The manifolds embedded here require a large number of sampled parameter points to resolve adequately, restricting the number of parameters considered. Similarly, calculating $\approx N^2$ statistical parameters under the Gaussian process approximation would be computationally infeasible. The first of these restrictions led to the choice of only two key parameters—the timescales—in the current work. The second restriction motivates the reduction of the model by population-averaging. The embedding of the inverse timescale plane (see Sec. II E) revealed that much of the manifold was comprised of points near the boundaries where the behavior became pathological. This suggests that a principled or data-informed restriction of parameter space may lead to a decrease in the necessary per-parameter sampling density and ease the restrictions presented here.

B. Future directions

There are several avenues along which the next investigations could follow; in closing we highlight a few of the most promising directions.

As noted earlier, one could use external current inputs to the two populations to parametrize the manifolds instead of the membrane and synaptic timescales used in this work. This would allow us to compare and contrast, at an idealized level, differences in efficacy between manipulating network activity using pharmacological means (e.g., applying different neuromodulators to adjust circuit properties) versus direct stimulation (such as deep brain stimulation).

To build toward making more specific predictions for real biological networks it would be valuable to apply our analyses to models of specific experimental studies, tailoring the details of the model to concrete brain areas by incorporating known features of the circuitry, such as cell types beyond an excitatory-inhibitory dichotomy or spatial organization of the network. Such a model would still need to be relatively idealized in order for our current methods to remain tractable, but our general framework—marginalize, approximate, population-average, and then embed—can apply to cases beyond the general networks presented here.

Extensions of the isKL method that could enable application to more complex models include studying the evolution of manifolds in time, instead of the stationary scenarios considered here. A first-attempt way to do this would be to discretize time, apply the embedding procedure at each time-step, and trace points through the embedding space. While conceptually straightforward, this approach would involve significantly more computational investment and the interpretation of the results would be more complicated than in the system discussed in this paper.

One could instead try to extend the iSKL embedding framework to apply directly to the nonequilibrium path integral representations of the network statistical dynamics. The major obstacle to such an extension is that the representation of the probability of a trajectory involves a functional integral over auxiliary response fields [like $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{n}}$ in Eq. (3)] [61], rendering direct computation of the D_{sKL} intractable for most models. This will require further approximations using, e.g., the replica trick [67] to deal with the logarithms in the definition of the D_{sKL} [Eq. (14)]. Nevertheless, this is a promising route for extending the manifold embedding methods used in this work to time-dependent contexts.

ACKNOWLEDGMENTS

J.T.C. and B.A.W.B. thank the National Institute of Mental Health and National Institute for Neurological Disorders and Stroke under Grant No. UF-1NS115779-01 and Stony Brook University for financial support for this work.

APPENDIX A: POPULATION AVERAGING OF THE GAUSSIAN PROCESS APPROXIMATED NETWORK

Here, we derive the reduced model from Eq. (6) by first making a Gaussian process approximation on the full-network spiking model and then averaging the resulting dynamics by population. The stochastically spiking full network, modeled using a nonlinear Hawkes process, is reproduced here:

$$\frac{dV_i}{dt} = -\tau_m^{-1}(V_i - \varepsilon_I) + I_i + \tau_s^{-1}\left(\mu_{\text{ext}} - J_{\text{self}}\dot{n}_i(t) + \sum_j w_{ij}\dot{n}_j(t)\right),$$

$$\dot{n}_i(t)dt \sim \text{Poiss}[\phi(V_i(t))dt].$$

Recall that the lowercase subscripts (i, j, etc.) denote individual neurons within the network. V_i is the membrane potential of neuron i, ε_i is the leak reversal potential, w_{ij} is the strength of a synaptic connection from neuron j to neuron i, and $-J_{\text{self}}$ is an inhibitory self-coupling. The two currents μ_{ext} and I_i represent an average current from an external network and an experimentally injected current, respectively. The process $\dot{n}_i(t)$ is the spike train of neuron i, and $\phi(\cdot)dt$ is the instantaneous firing rate nonlinearity, here given by $\phi(x) = \frac{1}{2}[x + (x^2 + 1/2)^{1/2}]$. Finally, τ_m and τ_s are modulated membrane and synaptic timescales, respectively. The mean-field equations for the steady-state membrane potentials can be obtained directly from these equations by using the fact that the approximation neglects fluctuations, and hence $\langle n_i(t)\rangle = \langle\phi(V_i(t))\rangle \approx \phi(\langle V_i(t)\rangle)$, yielding

$$V_i^{\text{mf}} = \varepsilon_I + \tau_m I_i + \frac{\tau_m}{\tau_s}\left(\mu_{\text{ext}} - J_{\text{self}}\phi\left(V_i^{\text{mf}}\right) + \sum_j w_{ij}\phi\left(V_j^{\text{mf}}\right)\right). \quad \text{(A1)}$$

To obtain the dynamics of fluctuations around the mean-field predictions, and to set up for future calculations that go even beyond the Gaussian approximation, it is useful to introduce a path integral representation of this stochastic process, using techniques from statistical physics [61]. In discrete time, we can write the joint probability for the membrane potential $\mathbf{V}(t)$ and the spike trains $\dot{\mathbf{n}}(t)$ as follows:

$$P[\mathbf{V}(t), \dot{\mathbf{n}}(t)] = \prod_{t,i} P[V_i(t)|\dot{\mathbf{n}}(t - dt)(t - dt)] \times P[\dot{n}_i(t - dt)|\mathbf{V}(t - dt)],$$

where the dynamics of the membrane potential are deterministic given a particular history of the spike trains,

$$P[V_i(t)|\dot{\mathbf{n}}(t - dt)] \propto \delta\left(\frac{dV_i}{dt} + \tau_m^{-1}(V_i - \varepsilon_I) - I_i - \tau_s^{-1} \times \left(\mu_{\text{ext}} - J_{\text{self}}\dot{n}_i(t) + \sum_j w_{ij}\dot{n}_j(t)\right)\right).$$

Here, the proportionality hides a Jacobian factor that arises from a change of variables from $V_I(t)$ to $\dot{V}_I(t)$; this factor is constant for an Itô time discretization, which we assume here.

Next, we take the spike train process to be conditionally Poisson given the current value of the membrane

potentials:

$$P[\dot{n}_i(t-dt)|V_i(t-dt)] = \frac{\phi(V_i(t-dt))^{\dot{n}_i(t-dt)dt}}{[\dot{n}_i(t-dt)dt]!}e^{-\phi(V_i(t-dt))dt},$$

giving an overall representation

$$P[\mathbf{V}(t),\dot{\mathbf{n}}(t)] = \prod_{t,i}\delta\left(\frac{dV_i}{dt}+\tau_m^{-1}(V_i-\varepsilon_I)-I_i - \tau_s^{-1}\left(\mu_{\text{ext}}-J_{\text{self}}\dot{n}_i(t)+\sum_j w_{ij}\dot{n}_j(t)\right)\right) \times\left[\frac{\phi(V_i(t-dt))^{\dot{n}_i(t-dt)dt}}{[\dot{n}_i(t-dt)dt]!}e^{-\phi(V_i(t-dt))dt}\right].$$

To cast this in a path integral representation, the standard approach is to represent the probability distributions in terms of a Fourier space representation. For the δ distribution we have

$$\delta(x) = \int_{-i\infty}^{i\infty}\frac{d\tilde{x}}{2\pi}e^{-\tilde{x}x},$$

and for a Poisson distribution with rate λ we have

$$p(x) = \int_{-i\infty}^{i\infty}\frac{d\tilde{x}}{2\pi}e^{-\tilde{x}x+W(\tilde{x})} = \int_{-i\infty}^{i\infty}\frac{d\tilde{x}}{2\pi}e^{-\tilde{x}x+\lambda\left(e^{\tilde{x}}-1\right)},$$

where $W(\tilde{x}) = \lambda[\exp(\tilde{x})-1]$ is the cumulant generating function for the Poisson process. We have adopted the standard notation from physics of writing the auxiliary variables this process introduces with tildes, and absorbing the factor of the imaginary unit i into the notation (giving imaginary units of integration). The path integral representation of the spiking process above is then given by

$$P[\mathbf{V}(t),\dot{\mathbf{n}}(t)] = \int\mathfrak{D}[\tilde{\mathbf{V}},\tilde{\mathbf{n}}]e^{-S[\tilde{\mathbf{V}},\mathbf{V},\tilde{\mathbf{n}},\dot{\mathbf{n}}]},$$

where $S[\tilde{\mathbf{V}},\mathbf{V},\tilde{\mathbf{n}},\dot{\mathbf{n}}]$ is referred to as the "action" of the process. We take the continuous-time limit, converting the product over time into an integral over time in the exponent. For this particular model, the action is given by

$$S[\tilde{\mathbf{V}},\mathbf{V},\tilde{\mathbf{n}},\dot{\mathbf{n}}] = \int dt\sum_{i=1}^{n}\left\{\tilde{V}_i\left[\dot{V}_i+\frac{V_i-\varepsilon_i}{\tau_m}-I_i - \tau_s^{-1}\left(\mu_{\text{ext}}-J_{\text{self}}\dot{n}_i(t)+\sum_j w_{ij}\dot{n}_j(t)\right)\right] + \tilde{n}_i(t)\dot{n}_i(t)-(e^{\tilde{n}_i(t)}-1)\phi(V_i)\right\}.$$

For our purposes, it will be convenient to marginalize out the dynamics of the spiking process $\dot{\mathbf{n}}(t)$ and its conjugate variable $\tilde{\mathbf{n}}(t)$ to obtain a representation for the stochastic dynamics of the membrane potentials [along with their auxiliary variables $\tilde{\mathbf{V}}(t)$]. The spike-marginalized action is

$$S[\tilde{\mathbf{V}},\mathbf{V}] = \int dt\sum_{i=1}^{n}\left\{\tilde{V}_i\left[\dot{V}_i+\frac{V_i-\varepsilon_I}{\tau_m}-I_i-\tau_s^{-1}\mu_{\text{ext}}\right] - \left(e^{\tau_s^{-1}(-J_{\text{self}}\tilde{V}_i+\sum_j\tilde{V}_j w_{ji})}-1\right)\phi(V_i)\right\}.$$

The Gaussian process approximation is derived by expanding this action around the mean-field solution, retaining only terms up to quadratic order in $\mathbf{V}(t)-\mathbf{V}^{\text{mf}}$ and $\tilde{\mathbf{V}}(t)$. The mean-field solution is obtained by the saddle-points of the action with respect to $\mathbf{V}(t)$ and $\tilde{\mathbf{V}}(t)$, which reproduce Eq. (A1) for $\mathbf{V}^{\text{mf}}$ and yield $\tilde{\mathbf{V}}^{\text{mf}} = \mathbf{0}$. We thus perform a functional Taylor-series expansion of the action around $(\tilde{\mathbf{V}},\mathbf{V}) = (\mathbf{0},\mathbf{V}^{\text{mf}})$, keeping only terms to the second order in $\delta\mathbf{V} = \mathbf{V}-\mathbf{V}^{\text{mf}}$ and $\tilde{\mathbf{V}}$. The result is

$$S[\tilde{\mathbf{V}},\mathbf{V}] = \frac{1}{2}\int dtdt'\sum_{ij}\tilde{V}_i(t)\left[-\tau_s^{-2}\sum_k(-\delta_{ik}J_{\text{self}}+w_{ik}) \times(-\delta_{jk}J_{\text{self}}+w_{jk})\phi\left(V_k^{\text{mf}}\right)\right]\tilde{V}_j(t') + \int dtdt'\sum_{ij}\tilde{V}_i(t)\left[\delta_{ij}\delta(t-t')\frac{d}{dt}+\delta_{ij}\left(\tau_m^{-1}+\tau_s^{-1}J_{\text{self}}\phi'\left(V_j^{\text{mf}}\right)\right)-\tau_s^{-1}w_{ij}\phi'\left(V_j^{\text{mf}}\right)\right]\delta V_j(t').$$

The form of the truncated action is the same as the path integral representation of an Ornstein-Uhlenbeck process derived explicitly by Chow and Buice [61]. We may therefore match terms to identify the effective stochastic process described by this action:

$$\frac{d\delta V_i}{dt} = -\sum_{j=1}^{n}\left[\delta_{ij}\left(\tau_m^{-1}+\tau_s^{-1}J_{\text{self}}\phi'\left(V_j^{\text{mf}}\right)\right)-\tau_s^{-1}w_{ij}\phi'\left(V_j^{\text{mf}}\right)\right] \times\delta V_j+\xi_i(t)\quad\text{for } i=1,2,\ldots,n,$$

where $\xi_i(t)$ is a zero-mean Gaussian noise with covariance

$$\langle\xi_i(t)\xi_j(t')\rangle = \tau_s^{-2}\sum_k(-\delta_{ik}J_{\text{self}}+w_{ik})(-\delta_{jk}J_{\text{self}}+w_{jk}) \times\phi\left(V_k^{\text{mf}}\right)\delta(t-t').$$

Casting this as a proper Itô stochastic differential equation, we get

$$d\delta\mathbf{V} = -\mathbf{A}\delta\mathbf{V}dt+\mathbf{\Sigma}d\mathbf{W}_t,$$

or equivalently

$$d\mathbf{V} = \mathbf{A}(\mathbf{V}^{\text{mf}}-\mathbf{V})dt+\mathbf{\Sigma}d\mathbf{W}_t,$$

where

$$A_{ij} = \delta_{ij}\left(\tau_m^{-1}+\tau_s^{-1}J_{\text{self}}\phi'\left(V_j^{\text{mf}}\right)\right)-\tau_s^{-1}w_{ij}\phi'\left(V_j^{\text{mf}}\right),$$

$$(\mathbf{\Sigma}\mathbf{\Sigma}^T)_{ij} = \tau_s^{-2}\sum_k(-\delta_{ik}J_{\text{self}}+w_{ik})(-\delta_{jk}J_{\text{self}}+w_{jk})\phi\left(V_k^{\text{mf}}\right).$$

In deriving the reduced dynamics for the population averages, we begin with the Langevin dynamics derived for the full network. We consider the network to have weakly heterogeneous populations in which the connections $w_{ij} = w_{IJ}x_{ij}$ are given by Bernoulli variables, i.e., $w_{ij} = w_{IJ}x_{ij}$, where w_{IJ} is a constant depending on the pre- and postsynaptic population identities (J and I, respectively). We take each connection variable x_{ij} to be independent:

$$x_{ij} \sim \text{Bernoulli}(p_{IJ}).$$

We formally define the average of variable A_i across population I as

$$\langle\langle A_i\rangle\rangle_I = A_I \equiv \frac{1}{N_I}\sum_{i\in I} A_i(t).$$

At this point, we write the population-averaged connection weights as follows:

$$\langle\langle w_{ij}\rangle\rangle_I \approx p_{IJ}w_{IJ}.$$

We derive the effective equations for $V_{I=0} \equiv V_{i=0}$ (the test neuron) and the population averages:

$$V_{I=1} \equiv \frac{1}{N_1}\sum_{i\in 1} V_i, \quad V_{I=2} \equiv \frac{1}{N_2}\sum_{i\in 2} V_i.$$

We make mean-field-like approximations on the population average of terms like $\langle\langle f(A_i)\rangle\rangle_I \approx f(\langle\langle A_i\rangle\rangle_i = f(A_I)$, and we additionally assume approximate independence between the distributions of the synaptic connections, the stationary mean-field potentials V_i^{mf}, and the potentials V_i. We thus have

$$\begin{aligned}
\frac{d}{dt}\left(\frac{1}{N_I}\sum_{i\in I} V_i\right) &= \left\langle\left\langle \sum_j \left[\delta_{ij}\left(\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(V_j^{\text{mf}}\right)\right) - \tau_s^{-1}w_{ij}\phi'\left(V_j^{\text{mf}}\right)\right]\left(V_j^{\text{mf}} - V_j\right) + \xi_i(t)\right\rangle\right\rangle_I \\
&= \left\langle\left\langle\left[\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(V_i^{\text{mf}}\right)\right]V_i^{\text{mf}}\right\rangle\right\rangle_I - \left\langle\left\langle\left[\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(V_i^{\text{mf}}\right)\right]V_i\right\rangle\right\rangle_I \\
&\quad - \tau_s^{-1}\left\langle\left\langle \sum_j w_{ij}\phi'\left(V_j^{\text{mf}}\right)V_j^{\text{mf}}\right\rangle\right\rangle_I - \tau_s^{-1}\left\langle\left\langle \sum_j w_{ij}\phi'\left(V_j^{\text{mf}}\right)V_j\right\rangle\right\rangle_I + \langle\langle\xi_i(t)\rangle\rangle \\
&\approx \left[\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(\left\langle\left\langle V_i^{\text{mf}}\right\rangle\right\rangle_I\right)\right]\left\langle\left\langle V_i^{\text{mf}}\right\rangle\right\rangle_I - \left[\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(\left\langle\left\langle V_i^{\text{mf}}\right\rangle\right\rangle_I\right)\right]\langle\langle V_i\rangle\rangle_I \\
&\quad - \tau_s^{-1}\left\langle\left\langle \sum_J N_J\left\langle\left\langle w_{ij}\phi'\left(V_j^{\text{mf}}\right)V_j^{\text{mf}}\right\rangle\right\rangle_J\right\rangle\right\rangle_I - \tau_s^{-1}\left\langle\left\langle \sum_J N_J\left\langle\left\langle w_{ij}\phi'\left(V_j^{\text{mf}}\right)V_j\right\rangle\right\rangle_J\right\rangle\right\rangle_I + \Xi_I(t) \\
&\approx \left[\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(V_I^{\text{mf}}\right)\right]\left(V_I^{\text{mf}} - V_I\right) - \tau_s^{-1}\left\langle\left\langle \sum_J N_J\langle\langle w_{ij}\rangle\rangle_J\phi'\left(\left\langle\left\langle V_j^{\text{mf}}\right\rangle\right\rangle_J\right)\left\langle\left\langle V_j^{\text{mf}}\right\rangle\right\rangle_J\right\rangle\right\rangle_I \\
&\quad - \tau_s^{-1}\left\langle\left\langle \sum_J N_J\langle\langle w_{ij}\rangle\rangle_J\phi'\left(\left\langle\left\langle V_j^{\text{mf}}\right\rangle\right\rangle_J\right)\langle\langle V_j\rangle\rangle_J\right\rangle\right\rangle_I + \Xi_I(t) \\
&\approx \left[\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(V_I^{\text{mf}}\right)\right]\left(V_I^{\text{mf}} - V_I\right) - \tau_s^{-1}\sum_J N_J p_{IJ}w_{IJ}\phi'\left(V_J^{\text{mf}}\right)\left(V_J^{\text{mf}} - V_J\right) + \Xi_I(t) \\
\Rightarrow \frac{dV_{I=1}}{dt} &\approx \sum_J \left[\delta_{IJ}\left(\tau_m^{-1} + \tau_s^{-1}J_{\text{self}}\phi'\left(V_I^{\text{mf}}\right)\right) - \tau_s^{-1}w_{IJ}p_{IJ}N_J\phi'\left(V_J^{\text{mf}}\right)\right]\left(V_J^{\text{mf}} - V_J\right) + \Xi_I(t).
\end{aligned}$$

In the last line above, the population-averaged effective noise processes are defined by $\Xi_I(t) = \frac{1}{N_I}\sum_{i\in I}\xi_i(t)$, and the sum over J is over an arbitrary definition of subpopulations. In our particular case, we have $J \in \{0, 1, 2\}$ as defined in Sec. II D with $N_0 = 1$.

Next, we calculate the covariance of the population-averaged noise processes $\Xi_I(t)$. We make the mean-field-like approximations as before:

$$\begin{aligned}
\langle\Xi_I, \Xi_J\rangle &= \left\langle\frac{1}{N_I}\sum_{i\in I}\xi_i, \frac{1}{N_J}\sum_{j\in J}\xi_j\right\rangle = \frac{1}{N_I N_J}\sum_{i\in I, j\in J}\left[\langle\xi_i\xi_j\rangle - \langle\xi_i\rangle\langle\xi_j\rangle\right] \\
&\approx \frac{\tau_s^{-2}}{N_I N_J}\sum_{i\in I, j\in J}\left[\sum_k (-\delta_{ik}J_{\text{self}} + w_{ik})(-\delta_{jk}J_{\text{self}} + w_{jk})\phi\left(V_k^{\text{mf}}\right)\right]\delta(t - t') \\
&= \frac{\tau_s^{-2}}{N_I N_J}\sum_{i\in I, j\in J}\left[\delta_{ij}J_{\text{self}}^2\phi\left(V_i^{\text{mf}}\right) - J_{\text{self}}w_{ji}\phi\left(V_i^{\text{mf}}\right) - w_{ij}J_{\text{self}}\phi\left(V_j^{\text{mf}}\right) + \sum_K\sum_{k\in K} w_{ik}w_{jk}\phi\left(V_k^{\text{mf}}\right)\right]\delta(t - t')
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\tau_s^{-2}}{N_I N_J}\Bigg[\delta_{IJ} N_I \left\langle\left\langle J_{\text{self}}^2 \phi\left(V_i^{\text{mf}}\right)\right\rangle\right\rangle_I - N_I N_J \left\langle\left\langle J_{\text{self}} w_{ji} \phi\left(V_i^{\text{mf}}\right)\right\rangle\right\rangle_{I,J} - N_I N_J \left\langle\left\langle w_{ij} J_{\text{self}} \phi\left(V_j^{\text{mf}}\right)\right\rangle\right\rangle_{I,J} \\
&\quad + N_I N_J \left\langle\left\langle \sum_K \sum_{k\in K} \left\langle\left\langle w_{ik} w_{jk} \phi\left(V_k^{\text{mf}}\right)\right\rangle\right\rangle_K \right\rangle\right\rangle_{I,J}\Bigg]\delta(t-t') \\
&= \tau_s^{-2}\left[\delta_{IJ}\frac{J_{\text{self}}^2}{N_I}\phi\left(V_{\text{I}}^{\text{mf}}\right) - J_{\text{self}}\langle\langle w_{ji}\rangle\rangle_{I,J}\phi\left(V_I^{\text{mf}}\right) - \langle\langle w_{IJ}\rangle\rangle_{I,J} J_{\text{self}}\phi\left(V_J^{\text{mf}}\right) + \sum_K \langle\langle w_{jk}\rangle\rangle_{K,I,J} N_K \phi\left(V_K^{\text{mf}}\right)\right]\delta(t-t') \\
&= \tau_s^{-2}\left[\delta_{IJ}\frac{J_{\text{self}}^2}{N_I}\phi\left(V_{\text{I}}^{\text{mf}}\right) - J_{\text{self}} p_{JI} w_{JI}\phi\left(V_I^{\text{mf}}\right) - p_{JI} w_{IJ} J_{\text{self}}\phi\left(V_j^{\text{mf}}\right) + \sum_K p_{JK} w_{JK} N_K \phi\left(V_K^{\text{mf}}\right)\right]\delta(t-t') \\
&= \tau_s^{-2}\sum_{\text{K}}\left(-\delta_{IK}\frac{J_{\text{self}}}{N_K} + p_{IK} w_{IK}\right)\left(-\delta_{JK}\frac{J_{\text{self}}}{N_K} + p_{JK} w_{JK}\right) N_K \phi(V_K^*)\delta(t-t').
\end{aligned}
$$

Note that in this derivation we are assuming an equivalence between the temporal mean-field membrane potential for each individual neuron V_i (used in the previous section) with the mean-field value of the population-averaged membrane potential V_I. This amounts to saying the network is sufficiently large and thus the mean of the membrane potential V_i for $i \in I$ tends toward the mean of V_I. This yields a stochastic differential equation of the form

$$d\mathbf{V} = \mathbf{A}(\mathbf{V}^{\text{mf}} - \mathbf{V})dt + \mathbf{\Sigma} d\mathbf{W}_t,$$

where

$$
\begin{aligned}
A_{ij} &= \delta_{ij}\left(\tau_m^{-1} + \tau_s^{-1} J_{\text{self}}\phi'\left(V_j^{\text{mf}}\right)\right) - \tau_s^{-1} w_{ij}\phi'\left(V_j^{\text{mf}}\right), \\
(\mathbf{\Sigma}\mathbf{\Sigma}^T)_{ij} &= \tau_s^{-2}\sum_k (-\delta_{ik} J_{\text{self}} + w_{ik})(-\delta_{jk} J_{\text{self}} + w_{jk})\phi\left(V_k^{\text{mf}}\right).
\end{aligned}
$$

APPENDIX B: GAUSSIAN PROCESS APPROXIMATION OF A POPULATION-AVERAGED NETWORK

In this Appendix, we derive the reduced model from Eq. (6) by first averaging the Hawkes process dynamics across subpopulations and then making a Gaussian approximation, reversing the order of operations in Appendix A. We begin with the base model:

$$
\begin{aligned}
\frac{dV_i}{dt} &= -\tau_m^{-1}(V_i - \varepsilon_I) + I_i + \tau_s^{-1}\left(\mu_{\text{ext}} - J_{\text{self}}\dot{n}_i(t) + \sum_J \sum_{j\in J} w_{ij}\dot{n}_j(t)\right), \\
\dot{n}_i(t)dt &\sim \text{Poiss}[\phi(V_i(t))dt].
\end{aligned}
$$

The population-averaged membrane potential dynamics are given by

$$
\begin{aligned}
\frac{d}{dt}V_I = \frac{d}{dt}\langle\langle V_i\rangle\rangle_I &= -\frac{\langle\langle V_i\rangle\rangle_I - \varepsilon_I}{\tau_m} + \langle\langle I_i\rangle\rangle_I + \frac{\mu_{\text{ext}}}{\tau_s} - \tau_s^{-1} J_{\text{self}}\langle\langle\dot{n}_i(t)\rangle\rangle_I + \tau_s^{-1}\left\langle\left\langle \sum_J \sum_{j\in J} w_{ij}\dot{n}_j(t)\right\rangle\right\rangle \\
&= -\frac{V_I - \varepsilon_I}{\tau_m} + I_{\text{I}} + \frac{\mu_{\text{ext}}}{\tau_s} - \tau_s^{-1} J_{\text{self}}\langle\langle\dot{n}_i(t)\rangle\rangle_I + \tau_s^{-1}\sum_J \sum_{j\in J}\langle\langle w_{ij}\rangle\rangle\dot{n}_j(t).
\end{aligned}
$$

As before, we take the connections w_{ij} to be scaled Bernoulli variables, i.e., $w_{ij} = w_{IJ}x_{ij}$, where w_{IJ} is a constant depending on the pre- and postsynaptic population identities (J and I, respectively) and $x_{ij} \sim \text{Bernoulli}(p_{IJ})$. The population-averaged connections are again given by $\langle\langle w_{ij}\rangle\rangle_I \approx p_{IJ} w_{IJ}$. We next recast the spiking processes into population spiking processes using the following definition:

$$\dot{m}_I(t) \equiv \sum_{i\in I}\dot{n}_i(t) = N_I\langle\langle\dot{n}_i\rangle\rangle_I.$$

As each $\dot{m}_I(t)$ is a sum of conditionally Poisson processes, it is also a conditionally Poisson process. Using the same mean-field-esque approximation as before, we may approximate the conditional rate of each $\dot{m}_I(t)$ as follows:

$$\begin{aligned}\dot{m}_I = \sum_{i\in I}\dot{n}_i(t) \sim \text{Poiss}\left(\sum_{i\in I}\phi(V_i(t))dt\right) &= \text{Poiss}(N_I\langle\langle\phi(V_i(t))\rangle\rangle_I dt)\\ &\approx \text{Poiss}(N_I\phi(\langle\langle V_i(t)\rangle\rangle_I)dt) = \text{Poiss}(N_I\phi(V_I(t))dt).\end{aligned}$$

With this, the population-averaged Hawkes process dynamics become

$$\begin{aligned}\frac{d}{dt}V_I &= -\frac{V_I-\varepsilon_I}{\tau_m} + I_\text{I} + \frac{\mu_\text{ext}}{\tau_s} - \tau_s^{-1}J_\text{self}\frac{\dot{m}_I(t)}{N_I} + \tau_s^{-1}\sum_J p_{IJ}w_{IJ}\dot{m}_J(t)\\ &= -\frac{V_I-\varepsilon_I}{\tau_m} + I_\text{I} + \tau_s^{-1}\left[\mu_\text{ext} + \sum_J\left(-\delta_{IJ}\frac{J_\text{self}}{N_I} + p_{IJ}w_{IJ}\right)\dot{m}_J(t)\right],\\ \dot{m}_I(t)dt &\sim \text{Poiss}[N_I\phi(V_I(t))dt].\end{aligned}$$

After deriving the population-averaged dynamics for the nonlinear Hawkes process, we apply the Gaussian-process approximation scheme to the new dynamics. We begin by applying a mean-field-like approximation to the average of the population-spiking processes, namely, $\langle\dot{m}_I(t)\rangle \approx N_I\phi(\langle V_I\rangle)$. This is used to find the stationary mean-field solution for the population-averaged membrane potential dynamics, given by a set of transcendental equations:

$$V_I^\text{mf} = \varepsilon_I + \tau_m I_\text{I} + \frac{\tau_m}{\tau_s}\left[\mu_\text{ext} + \sum_J\left(-\delta_{IJ}\frac{J_\text{self}}{N_I} + p_{IJ}w_{IJ}\right)N_J\phi\big(V_J^\text{mf}\big)\right].$$

As in Appendix A, we represent the joint probability distribution $P[\mathbf{V}(t), \dot{\mathbf{m}}(t)]$ as a path integral by discretizing time, making appropriate Fourier transforms, and taking a continuous-time limit. This yields the expression

$$P[\mathbf{V}(t), \dot{\mathbf{m}}(t)] = \int \mathfrak{D}[\tilde{\mathbf{V}}, \tilde{\mathbf{m}}]e^{-S[\tilde{\mathbf{V}},\mathbf{V},\tilde{\mathbf{m}},\dot{\mathbf{m}}]},$$

with

$$\begin{aligned}S[\tilde{\mathbf{V}}, \mathbf{V}, \tilde{\mathbf{m}}, \dot{\mathbf{m}}] = \int dt \sum_{I=0,1,2}\Bigg\{&\tilde{V}_I\left[\dot{V}_I + \frac{V_I-\varepsilon_i}{\tau_m} - I_\text{I} - \tau_s^{-1}\left(\mu_\text{ext} + \sum_J\left(-\delta_{IJ}\frac{J_\text{self}}{N_I} + p_{IJ}w_{IJ}\right)\dot{m}_J(t)\right)\right]\\ &+ \tilde{m}_I(t)\dot{m}_I(t) - (e^{\tilde{m}_I(t)} - 1)N_I\phi(V_I)\Bigg\}.\end{aligned}$$

We marginalize out the explicit spiking dynamics as before by finding the zeros of the derivatives of the action with respect to $\dot{\mathbf{m}}(t)$ and its conjugate variables $\tilde{\mathbf{m}}(t)$. This yields the following marginalized action:

$$S[\tilde{\mathbf{V}}, \mathbf{V}] = \int dt \sum_{I=0,1,2}\left\{\tilde{V}_I\left[\dot{V}_I + \frac{V_I-\varepsilon_I}{\tau_m} - I_\text{I} - \tau_s^{-1}\mu_\text{ext}\right] - \left(e^{\tau_s^{-1}(-\frac{J_\text{self}}{N_I}\tilde{V}_I + \sum_J \tilde{V}_j p_{JI}w_{JI})} - 1\right)N_I\phi(V_I)\right\}.$$

We expand this action around the mean-field solution $(\tilde{\mathbf{V}}, \mathbf{V}) = (\mathbf{0}, \mathbf{V}^\text{mf})$ to quadratic order. Evaluating individual terms and derivatives at the mean-field solution, we get

$$\begin{aligned}S[\mathbf{0}, \mathbf{V}^\text{mf}] &= 0,\\ S_{V_I}[\mathbf{0}, \mathbf{V}^\text{mf}] &= 0,\\ S_{\tilde{V}_I}[\mathbf{0}, \mathbf{V}^\text{mf}] &= \int dt\left[\dot{V}_i + \frac{V_i^\text{mf}-\varepsilon_I}{\tau_m} - \frac{V_i^\text{mf}-\varepsilon_I}{\tau_m}\right] = \int dt\big[\dot{V}_i - \dot{V}_i^\text{mf}\big] = \int dt\delta\dot{V}_i,\\ S_{\tilde{V}_I\tilde{V}_J}[\mathbf{0}, \mathbf{V}^\text{mf}] &= \int dt\left[-\tau_s^{-2}\sum_K\left(-\delta_{IK}\frac{J_\text{self}}{N_I} + p_{IK}w_{IK}\right)\left(-\delta_{JK}\frac{J_\text{self}}{N_J} + p_{JK}w_{JK}\right)N_K\phi(V_K)\right],\\ S_{\tilde{V}_I V_J}[\mathbf{0}, \mathbf{V}^\text{mf}] &= \int dt\big[\delta_{IJ}\big(\tau_m^{-1} + \tau_s^{-1}J_\text{self}\phi'\big(V_I^\text{mf}\big)\big) - \tau_s^{-1}p_{IJ}w_{IJ}N_J\phi'\big(V_J^\text{mf}\big)\big],\end{aligned}$$

and

$$S_{V_I V_J}[\mathbf{0}, \mathbf{V}^{\mathrm{mf}}] = 0.$$

Again defining fluctuations in the membrane potential as $\delta V_I := V_I - V_I^{\mathrm{mf}}$, the approximated action can be written as

$$S[\tilde{\mathbf{V}}, \mathbf{V}] = \int dt \sum_I \left\{ \delta \dot{V}_I + \sum_J \left[\delta_{IJ} \left(\tau_m^{-1} + \tau_s^{-1} J_{\mathrm{self}} \phi' \left(V_I^{\mathrm{mf}} \right) \right) - \tau_s^{-1} p_{IJ} w_{IJ} N_J \phi' \left(V_J^{\mathrm{mf}} \right) \right] \delta V_J \right\} \tilde{V}_I(t)$$
$$+ \frac{1}{2} \int dt dt' \sum_{ij} \left[-\tau_s^{-2} \sum_K \left(-\delta_{IK} \frac{J_{\mathrm{self}}}{N_I} + p_{IK} w_{IK} \right) \left(-\delta_{JK} \frac{J_{\mathrm{self}}}{N_J} + p_{JK} w_{JK} \right) N_K \phi(V_K) \right] \tilde{V}_I(t) \tilde{V}_J(t').$$

As before, we can identify the GPA dynamics of the population-averaged Hawkes process as corresponding to an Ornstein-Uhlenbeck process. We may therefore match terms to identify the effective stochastic process described by this action:

$$\frac{d\delta V_I}{dt} = - \sum_{J=0,1,2} \left[\delta_{IJ} \left(\tau_m^{-1} + \tau_s^{-1} J_{\mathrm{self}} \phi' \left(V_I^{\mathrm{mf}} \right) \right) - \tau_s^{-1} p_{IJ} w_{IJ} N_J \phi' \left(V_J^{\mathrm{mf}} \right) \right] \delta V_J + \xi_I(t) \quad \text{for } I = 0, 1, 2,$$

where $\xi_I(t)$ is a zero-mean Gaussian noise with covariance

$$\langle \xi_I(t) \xi_J(t') \rangle = \tau_s^{-2} \sum_K \left(-\delta_{IK} \frac{J_{\mathrm{self}}}{N_I} + p_{IK} w_{IK} \right) \left(-\delta_{JK} \frac{J_{\mathrm{self}}}{N_J} + p_{JK} w_{JK} \right) N_K \phi(V_K) \delta(t - t').$$

Casting this as a proper Itô stochastic differential equation, we get

$$d\mathbf{V} = \mathbf{A}(\mathbf{V}^{\mathrm{mf}} - \mathbf{V}) dt + \mathbf{\Sigma} d\mathbf{W}_t,$$

where

$$A_{ij} = \delta_{IJ} \left(\tau_m^{-1} + \tau_s^{-1} J_{\mathrm{self}} \phi' \left(V_I^{\mathrm{mf}} \right) \right) - \tau_s^{-1} p_{IJ} w_{IJ} N_J \phi' \left(V_J^{\mathrm{mf}} \right),$$
$$(\mathbf{\Sigma} \mathbf{\Sigma}^T)_{ij} = \tau_s^{-2} \sum_K \left(-\delta_{IK} \frac{J_{\mathrm{self}}}{N_I} + p_{IK} w_{IK} \right) \left(-\delta_{JK} \frac{J_{\mathrm{self}}}{N_J} + p_{JK} w_{JK} \right) N_K \phi(V_K).$$

We note that this is consistent with the form derived in Appendix A.

APPENDIX C: POPULATION AVERAGING FOR THE LINEAR NONSPIKING MODEL

We also construct a simpler model of networked, linear nonspiking (or graded potential) neurons. We assume the neurons are injected with large numbers of synaptic input that sum together to be approximately Gaussian, with nonzero mean μ_{ext}, creating a stochastic system with dynamics described by

$$\frac{dV_i}{dt} = - \tau_m^{-1} (V_i - \varepsilon_I) + I_i + \tau_s^{-1} \left(\mu_{\mathrm{ext}} - J_{\mathrm{self}} \phi(V_i) + \sum_j w_{ij} \phi \left(V_j \right) \right) + \xi_i(t).$$

We begin this derivation by assuming the connections $w_{ij} = w_{IJ} x_{ij}$ are scaled Bernoulli variables as in Appendixes A and B. Here, the transfer function $\phi(\cdot)$ is a simple linear function [i.e., $\phi(x) = x$]. The processes $\xi_i(t)$ are zero-mean Gaussian noise synaptic input from neurons external to the network being examined, and thus they scale with τ_s^{-1} [i.e., $\xi_i(t) \sim \tau_s^{-1}$]. We define the covariance of the noise processes $\xi_i(t)$ as follows:

$$\langle \xi_i(t) \xi_j(t') \rangle = \tau_s^{-2} \mu_{\mathrm{ext}} \delta(t - t').$$

Here, k_J is a constant potentially depending on the identity of the receiving population J. We wish to derive a population-averaged model for the membrane potential dynamics for comparison with the Gaussian-process-approximated spiking models. Again, we define

$$\langle \langle A_i \rangle \rangle_I \equiv \frac{1}{N_I} \sum_{i \in I} A_i(t).$$

We thus derive the population-averaged dynamics for population I:

$$\frac{d}{dt}\left(\frac{1}{N_I}\sum_{i\in 1} V_i\right) = -\left\langle\left\langle \frac{V_i-\varepsilon_I}{\tau_m}\right\rangle\right\rangle_I + \langle\langle I_i\rangle\rangle_I + \tau_s^{-1}\left(\mu_{\text{ext}} - \langle\langle J_{\text{self}}\phi(V_i)\rangle\rangle_I + \left\langle\left\langle \sum_J N_J\langle\langle w_{ij}\phi(V_j)\rangle\rangle_J\right\rangle\right\rangle_I\right) + \langle\langle \xi_i(t)\rangle\rangle_I,$$

$$\Rightarrow \frac{dV_1}{dt} \approx -\frac{V_I-\varepsilon_I}{\tau_m} + I_I + \tau_s^{-1}\left(\mu_{\text{ext}} - J_{\text{self}}\phi(V_I) + \sum_J p_{IJ}w_{IJ}N_J\phi(V_J)\right) + \Xi_I(t),$$

where we have defined $\Xi_I(t) \equiv \frac{1}{N_I}\sum_{i\in I}\xi_i(t)$ for $I = 0, 1, 2$. The means and covariances of the population-averaged noise processes are as follows:

$$\langle \Xi_I(t)\rangle = \left\langle \frac{1}{N_I}\sum_{i\in I}\xi_i(t)\right\rangle = \frac{1}{N_I}\sum_{i\in I}\langle\xi_i(t)\rangle = 0,$$

and

$$\langle \Xi_I(t), \Xi_J(t)\rangle = \left\langle \frac{1}{N_I}\sum_{i\in I}\xi_i(t), \frac{1}{N_J}\sum_{j\in J}\xi_j(t)\right\rangle = \frac{1}{N_I N_J}\sum_{i\in I, j\in J}\langle\xi_i(t)\xi_j(t)\rangle - \langle\xi_i(t)\rangle\langle\xi_j(t)\rangle = \frac{1}{\tau_s^2 N_I N_J}\sum_{i\in I, j\in J}\delta_{ij}\mu_{\text{ext}}\delta(t-t')$$

$$= \frac{\delta_{IJ}}{\tau_s^2 N_I^2}\sum_{i\in I}\mu_{\text{ext}}\delta(t-t') = \frac{\delta_{IJ}}{\tau_s^2 N_I^2}N_I\mu_{\text{ext}}\delta(t-t') = \frac{\delta_{IJ}}{\tau_s^2 N_I}\mu_{\text{ext}}\delta(t-t').$$

We can then rewrite the population dynamics as

$$dV_I = \left[-\frac{V_I-\varepsilon_I}{\tau_m} + I_I + \tau_s^{-1}\left(\mu_{\text{ext}} - J_{\text{self}}\phi(V_I) + \sum_J p_{IJ}w_{IJ}N_J\phi(V_J)\right)\right]dt + \Xi_I(t)dt,$$

$$\rightarrow d\mathbf{V} = \mathbf{A}\left(\mathbf{A}^{-1}\left(\tau_s^{-1}\mu_{\text{ext}} + I_1\right) - \mathbf{V}\right)dt + \mathbf{\Sigma}d\mathbf{W}_t = \mathbf{A}(\mu - \mathbf{V})dt + \mathbf{\Sigma}d\mathbf{W}_t,$$

where

$$\mathbf{A}_{IJ} = \delta_{IJ}\left(\tau_m^{-1} - \tau_s^{-1}J_{\text{self}}\right) + \tau_s^{-1}p_{IJ}w_{IJ}N_J$$
$$= \delta_{IJ}\tau_m^{-1} + \tau_s^{-1}w_{IJ}^*,$$
$$w_{IJ}^* = -\delta_{IJ}J_{\text{self}} + p_{IJ}w_{IJ}N_J,$$
$$(\Sigma\Sigma^T)_{IJ} = \frac{\delta_{IJ}}{\tau_s^2 N_I}\mu_{\text{ext}}.$$

APPENDIX D: BALANCE EQUATIONS

To derive the balanced state conditions for the network, we begin with the population-averaged spiking network as derived in Appendix B:

$$\frac{d}{dt}V_I = -\frac{V_I-\varepsilon_I}{\tau_m} + I_1 + \tau_s^{-1} \times \left[\mu_{\text{ext}} + \sum_{J=0,1,2}\left(-\delta_{IJ}\frac{J_{\text{self}}}{N_I} + p_{IJ}w_{IJ}\right)\dot{m}_J(t)\right]$$

$$\dot{m}_I(t)dt \sim \text{Poiss}[N_I\phi(V_I(t))dt],$$

where $p_{IJ}w_{IJ}$ came from the population-averaged synaptic connection $\langle\langle w_{ij}\rangle\rangle_J$ and the effective spike count processes are $\dot{m}_J(t) = \sum_{j\in J}\dot{n}_j(t)$. The total external input to "neuron" I is $I_I + \tau_s^{-1}[\mu_{\text{ext}} - \frac{J_{\text{self}}}{N_I}\dot{m}_I(t) + \sum_J p_{IJ}w_{IJ}\dot{m}_J(t)]$. We want to estimate the mean and variance of this input, taken over the stochastic process. The mean is straightforward, yielding

$$\tau_s^{-1}\kappa_I \equiv I_1 + \tau_s^{-1}\left(\mu_{\text{ext}} - J_{\text{self}}\phi(V_I) + \sum_J p_{IJ}w_{IJ}N_J\phi(V_J)\right).$$

Note that the correction term $-J_{\text{self}}\dot{m}_I(t)$ is always going to be smaller than the $\sum_J p_{IJ}w_{IJ}\dot{m}_J$ term, so for the purposes of the balanced condition calculation we neglect it. For the current work, we take the injected currents I_I to be constants.

Calculating the covariance of the total input at times t and t' yields

$$\sum_{JK} p_{IJ}w_{IJ}p_{IK}w_{IK}[\langle\dot{m}_J(t)\dot{m}_K(t')\rangle - \langle\dot{m}_J(t)\rangle\langle\dot{m}_K(t')\rangle].$$

We make a Poisson approximation to replace the covariance of the $\dot{m}$ with $\langle\dot{m}_J(t)\rangle\delta_{JK}\delta(t-t')$. Hence, the covariance becomes

$$\sum_J (p_{IJ}w_{IJ})^2 N_J\phi(V_J)\delta(t-t').$$

We want the variance of the synaptic input to be $O(N^0)$, which means that to leading order we want

$$\sum_J (p_{IJ}w_{IJ})^2 N_J\phi(V_J) \approx (p_{I1}w_{I1})^2 N_1\phi(V_1) + (p_{I2}w_{I2})^2 N_2\phi(V_2) \sim O(N^0).$$

We neglect the contribution from the test neuron because it is subleading here, i.e., $N_0 = 1 \ll N_1, N_2$. In order for this expression to be order one, we see that we need w_{IJ} to scale like $1/\sqrt{N}$, as implemented in Eqs. (12) and (13).

We return to the mean input to neuron I, which we write as

$$\tau_s^{-1}\kappa_I \approx \sqrt{N}\left(\frac{I_1 + \tau_s^{-1}\mu_{\text{ext}}}{\sqrt{N}} + \tau_s^{-1}\left\{p_{I1}w_{I1}\frac{N_1}{\sqrt{N}}\phi(V_1) + p_{I2}w_{I2}\frac{N_2}{\sqrt{N}}\phi(V_2)\right\}\right).$$

For a balanced network κ_I should be $O(1)$ for all I, which means that the terms in brackets must vanish faster than $1/\sqrt{N}$. We assume that $I_I, \mu_{\text{ext}} \propto \sqrt{N}$, and because $w_{IJ} \sim 1/\sqrt{N}$ and $N_I \propto N$ (for $I \neq 0$), the terms in brackets are $O(1)$.

As $N \to \infty$, the terms in brackets must vanish in order for κ_I to be finite. This yields a linear system of equations that uniquely determines the means $\mu_I = \phi(V_I)$ and allows us to place constraints on the parameters:

$$-\begin{bmatrix} I_1 + \tau_s^{-1}\mu_{\text{ext}} \\ I_2 + \tau_s^{-1}\mu_{\text{ext}} \end{bmatrix} = \frac{1}{\tau_s}\begin{bmatrix} p_{11}w_{11}N_1 & p_{12}w_{12}N_2 \\ p_{21}w_{21}N_1 & p_{22}w_{22}N_2 \end{bmatrix}\begin{bmatrix} \phi(V_1) \\ \phi(V_2) \end{bmatrix}.$$

Solving this system of equations for the spike rates $\phi(V_I^{\text{mf}})$, we get

$$\phi(V_1) = \frac{\tau_s}{N_1}\frac{p_{12}w_{12}\left(I_2 + \tau_s^{-1}\mu_{\text{ext}}\right) - p_{22}w_{22}\left(I_1 + \tau_s^{-1}\mu_{\text{ext}}\right)}{p_{11}p_{22}w_{11}w_{22} - p_{12}p_{21}w_{21}w_{12}},$$

$$\phi(V_2) = \frac{\tau_s}{N_2}\frac{p_{21}w_{21}\left(I_1 + \tau_s^{-1}\mu_{\text{ext}}\right) - p_{11}w_{11}\left(I_2 + \tau_s^{-1}\mu_{\text{ext}}\right)}{p_{11}p_{22}w_{11}w_{22} - p_{12}p_{21}w_{21}w_{12}}.$$

In the case of our particular models, we can further reduce this expression by noting that $I_I = 0$ for $I = 1, 2$ and $p_{IJ} = p \; \forall \, I, J$:

$$\phi(V_1) = \frac{1}{pN_1}\frac{w_{12} - w_{22}}{w_{11}w_{22} - w_{21}w_{12}}\mu_{\text{ext}},$$

$$\phi(V_2) = \frac{1}{pN_2}\frac{w_{21} - w_{11}}{w_{11}w_{22} - w_{21}w_{12}}\mu_{\text{ext}}.$$

We highlight here that $\phi(V_I) > 0$ by its definition as a firing rate. Additionally, μ_{ext} is assumed to by synaptic input projected into the local network and is thus positive (i.e., excitatory) here. Taken together, these two points mean the synaptic parameters must satisfy one of the two following sets of inequalities to be in a balanced regime:

$$\begin{aligned} &w_{11}w_{22} > w_{12}w_{21}, \\ &w_{12} > w_{22}, \\ &w_{21} > w_{11}, \end{aligned} \tag{D1}$$

or

$$\begin{aligned} &w_{11}w_{22} < w_{12}w_{21}, \\ &w_{12} < w_{22}, \\ &w_{21} < w_{11}. \end{aligned} \tag{D2}$$

With this, we have derived the appropriate scaling for the various parameters in the model and found constraints for the synaptic strengths in order satisfy the necessary properties of a balanced network.

[1] G. D. Field and E. J. Chichilnisky, Information processing in the primate retina: Circuitry and coding., Annu. Rev. Neurosci. **30**, 1 (2007).

[2] J. Trousdale, Y. Hu, E. Shea-Brown, and K. Josić, Impact of network structure and cellular response on spike time correlations, PLoS Comput. Biol. **8**, e1002408 (2012).

[3] Y. Hu, J. Trousdale, K. Josić, and E. Shea-Brown, Motif statistics and spike correlations in neuronal networks, J. Stat. Mech. (2013) P03012.

[4] G. K. Ocker, K. Josić, E. Shea-Brown, and M. A. Buice, Linking structure and activity in nonlinear spiking networks, PLoS Comput. Biol. **13**, e1005583 (2017).

[5] G. K. Ocker, Y. Hu, M. A. Buice, B. Doiron, K. Josić, R. Rosenbaum, and E. Shea-Brown, From the statistics of connectivity to the statistics of spike times in neuronal networks, Curr. Opin. Neurobiol. **46**, 109 (2017).

[6] K. Batista-García-Ramó and C. I. Fernández-Verdecia, What we know about the brain structure–function relationship, Behav. Sci. **8**, 39 (2018).

[7] C. Curto and K. Morrison, Relating network connectivity to dynamics: Opportunities and challenges for theoretical neuroscience, Curr. Opin. Neurobiol. **58**, 11 (2019).

[8] B. E. Mizusaki and C. O'Donnell, Neural circuit function redundancy in brain disorders, Curr. Opin. Neurobiol. **70**, 74 (2021).

[9] O. Mazor and G. Laurent, Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons, Neuron **48**, 661 (2005).

[10] J. W. Pillow and E. P. Simoncelli, Dimensionality reduction in neural models: An information-theoretic generalization of spike-triggered average and covariance analysis, J. Vis. **6**, 9 (2006).

[11] S. Ganguli, J. W. Bisley, J. D. Roitman, M. N. Shadlen, M. E. Goldberg, and K. D. Miller, One-dimensional dynamics of attention and decision making in lip, Neuron **58**, 15 (2008).

[12] J. P. Cunningham and M. Y. Byron, Dimensionality reduction for large-scale neural recordings, Nat. Neurosci. **17**, 1500 (2014).

[13] P. T. Sadtler, K. M. Quick, M. D. Golub, S. M. Chase, S. I. Ryu, E. C. Tyler-Kabara, M. Y. Byron, and A. P. Batista, Neural constraints on learning, Nature (London) **512**, 423 (2014).

[14] E. W. Archer, U. Koster, J. W. Pillow, and J. H. Macke, Low-dimensional models of neural population activity in sensory cortical circuits, in *Advances in Neural Information Processing Systems 27: 28th Conference on Neural Information Processing Systems (NIPS 2014)* (2015), pp. 343–351.

[15] L. Mazzucato, A. Fontanini, and G. La Camera, Dynamics of multistable states during ongoing and evoked cortical activity, J. Neurosci. **35**, 8214 (2015).

[16] L. Mazzucato, A. Fontanini, and G. La Camera, Stimuli reduce the dimensionality of cortical activity, Front. Syst. Neurosci. **10**, 11 (2016).
[17] P. Gao, E. Trautmann, B. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, A theory of multineuronal dimensionality, dynamics and measurement, bioRxiv, 214262 (2017), doi:10.1101/214262.
[18] J. D. Murray, A. Bernacchia, N. A. Roy, C. Constantinidis, R. Romo, and X.-J. Wang, Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex, Proc. Natl. Acad. Sci. USA **114**, 394 (2017).
[19] E. Wärnberg and A. Kumar, Perturbing low dimensional activity manifolds in spiking neuronal networks, PLoS Comput. Biol. **15**, e1007074 (2019).
[20] C. J. MacDowell and T. J. Buschman, Low-dimensional spatiotemporal dynamics underlie cortex-wide neural activity, Curr. Biol. **30**, 2665 (2020).
[21] A. Prinz, D. Bucher, and E. Marder, Similar network activity from disparate circuit parameters, Nat. Neurosci. **7**, 1345 (2004).
[22] E. Marder and D. Bucher, Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs, Annu. Rev. Physiol. **69**, 291 (2007).
[23] E. Marder, M. Goeritz, and A. Otopalik, Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms, Curr. Opin. Neurobiol. **31**, 156 (2015).
[24] E. Cropper, A. Dacks, and K. Weiss, Consequences of degeneracy in network function, Curr. Opin. Neurobiol. **41**, 62 (2016).
[25] H. Nakahara and S.-i. Amari, Information-geometric measure for neural spikes, Neural Comput. **14**, 2269 (2002).
[26] S. Wu, S.-i. Amari, and H. Nakahara, Population coding and decoding in a neural field: A computational study, Neural Comput. **14**, 999 (2002).
[27] S.-i. Amari, H. Nakahara, S. Wu, and Y. Sakai, Synchronous firing and higher-order interactions in neuron pool, Neural Comput. **15**, 127 (2003).
[28] S.-i. Amari, H. Park, and T. Ozeki, Singularities affect dynamics of learning in neuromanifolds, Neural Comput. **18**, 1007 (2006).
[29] H. Shimazaki, S.-i. Amari, E. N. Brown, and S. Grün, State-space analysis of time-varying higher-order spike correlation for multiple neural spike train data, PLoS Comput. Biol. **8**, e1002385 (2012).
[30] S.-i. Amari, R. Karakida, and M. Oizumi, Statistical neurodynamics of deep networks: Geometry of signal spaces, Nonlinear Theory Appl. IEICE **10**, 322 (2019).
[31] S.-i. Amari, R. Karakida, and M. Oizumi, Fisher information and natural gradient learning in random deep networks, in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019), pp. 694–702.
[32] R. Karakida, S. Akaho, and S.-i. Amari, Universal statistics of Fisher information in deep neural networks: Mean field approach, J. Stat. Mech. (2020) 124005.
[33] M. K. Transtrum, B. B. Machta, and J. P. Sethna, Why are nonlinear fits to data so challenging? Phys. Rev. Lett. **104**, 060201 (2010).
[34] M. K. Transtrum, B. B. Machta, and J. P. Sethna, Geometry of nonlinear least squares with applications to sloppy models and optimization, Phys. Rev. E **83**, 036701 (2011).
[35] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, Parameter space compression underlies emergent theories and predictive models, Science **342**, 604 (2013).
[36] M. K. Transtrum and P. Qiu, Model reduction by manifold boundaries, Phys. Rev. Lett. **113**, 098701 (2014).
[37] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, Perspective: Sloppiness and emergent theories in physics, biology, and beyond, J. Chem. Phys. **143**, 010901 (2015).
[38] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, Universally sloppy parameter sensitivities in systems biology models, PLoS Comput. Biol. **3**, e189 (2007).
[39] K. N. Quinn, C. B. Clement, F. De Bernardis, M. D. Niemack, and J. P. Sethna, Visualizing probabilistic models and data with intensive principal component analysis, Proc. Natl. Acad. Sci. USA **116**, 13762 (2019).
[40] H. K. Teoh, K. N. Quinn, J. Kent-Dobias, C. B. Clement, Q. Xu, and J. P. Sethna, Visualizing probabilistic models in Minkowski space with intensive symmetrized Kullback-Leibler embedding, Phys. Rev. Res. **2**, 033221 (2020).
[41] H. B. Barlow, Possible principles underlying the transformations of sensory messages, *Sensory Communication* (MIT Press, Cambridge, MA, 2012), Chap. 13, pp. 217–234.
[42] S. Laughlin, A simple coding procedure enhances a neuron's information capacity, Z. Naturforsch., C: J. Biosci. **36**, 910 (1981).
[43] J. H. van Hateren, A theory of maximizing sensory information, Biol. Cybern. **68**, 23 (1992).
[44] J. J. Atick, Could information theory provide an ecological theory of sensory processing? Netw.: Comput. Neural Syst. **3**, 213 (1992).
[45] F. Rieke, D. Warland, R. d. R. Van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA, 1999).
[46] B. B. Averbeck, P. E. Latham, and A. Pouget, Neural correlations, population coding and computation, Nat. Rev. Neurosci. **7**, 358 (2006).
[47] Z. Wang, A. A. Stocker, and D. D. Lee, Optimal neural tuning curves for arbitrary stimulus distributions: discrimax, infomax and minimum Lp loss, in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Incorporated, Redhook, NY, 2012).
[48] R. Moreno-Bote, J. Beck, I. Kanitscheider, X. Pitkow, P. Latham, and A. Pouget, Information-limiting correlations, Nat. Neurosci. **17**, 1410 (2014).
[49] J. Gjorgjieva, H. Sompolinsky, and M. Meister, Benefits of pathway splitting in sensory coding, J. Neurosci. **34**, 12127 (2014).
[50] D. B. Kastner, S. A. Baccus, and T. O. Sharpee, Critical and maximally informative encoding between neural populations in the retina, Proc. Natl. Acad. Sci. USA **112**, 2533 (2015).
[51] B. A. Brinkman, A. I. Weber, F. Rieke, and E. Shea-Brown, How do efficient coding strategies depend on origins of noise in neural circuits? PLoS Comput. Biol. **12**, e1005150 (2016).
[52] J. Zylberberg, A. Pouget, P. E. Latham, and E. Shea-Brown, Robust information propagation through noisy neural circuits, PLoS Comput. Biol. **13**, e1005497 (2017).

[53] J. A. Pruszynski and J. Zylberberg, The language of the brain: Real-world neural population codes, Curr. Opin. Neurobiol. **58**, 30 (2019).

[54] Y. Zhou, B. H. Smith, and T. O. Sharpee, Hyperbolic geometry of the olfactory space, Sci. Adv. **4**, eaaq1458 (2018).

[55] Y. Zhou and T. O. Sharpee, Hyperbolic geometry of gene expression, iScience **24**, 102225 (2021).

[56] H. Zhang, P. D. Rich, A. K. Lee, and T. O. Sharpee, Hippocampal spatial representations exhibit a hyperbolic geometry that expands with experience, Nat. Neurosci. **26**, 131 (2023).

[57] B. A. W. Brinkman, F. Rieke, E. Shea-Brown, and M. A. Buice, Predicting how and when hidden neurons skew measured synaptic interactions, PLoS Comput. Biol. **14**, e1006490 (2018).

[58] F. Nadim and D. Bucher, Neuromodulation of neurons and synapses, Curr. Opin. Neurobiol. **29**, 48 (2014).

[59] O. Swanson, R. Semaan, and A. Maffei, Reduced dopamine signaling impacts pyramidal neuron excitability in mouse motor cortex, eNeuro **8**, ENEURO.0548-19.2021 (2021).

[60] O. K. Swanson, P. E. Yevoo, D. Richard, and A. Maffei, Altered thalamocortical signaling in a mouse model of parkinson's disease, J. Neurosci. **43**, 6021 (2023).

[61] C. Chow and M. Buice, Path integral methods for stochastic differential equations, J. Math. Neurosci. **5**, 8 (2015).

[62] G. K. Ocker, Dynamics of stochastic integrate-and-fire networks, Phys. Rev. X **12**, 041007 (2022).

[63] B. A. W. Brinkman, Non-perturbative renormalization group analysis of nonlinear spiking networks, arXiv:2301.09600.

[64] P. Vatiwutipong and N. Phewchean, Alternative way to derive the distribution of the multivariate Ornstein-Uhlenbeck process, Adv. Contin. Discrete Models **2019**, 276 (2019).

[65] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná, Hyperbolic geometry of complex networks, Phys. Rev. E **82**, 036106 (2010).

[66] A. Clauset, C. Moore, and M. E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature (London) **453**, 98 (2008).

[67] M. Advani, S. Lahiri, and S. Ganguli, Statistical mechanics of complex neural systems and high dimensional data, J. Stat. Mech. (2013) P03014.