Interrogating theoretical models of neural computation with deep inference

Sean R. Bittner[1], Agostina Palmigiano[1], Alex T. Piet[2,3,4], Chunyu A. Duan[5], Carlos D. Brody[2,3,6], Kenneth D. Miller[1], and John P. Cunningham[7].

[1]Department of Neuroscience, Columbia University,

[2]Princeton Neuroscience Institute,

[3]Princeton University,

[4]Allen Institute for Brain Science,

[5]Institute of Neuroscience, Chinese Academy of Sciences,

[6]Howard Hughes Medical Institute,

[7]Department of Statistics, Columbia University

1 Abstract

A cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures a hypothesized neural mechanism. Such models are valuable when they give rise to an experimentally observed phenomenon – whether behavioral or a pattern of neural activity – and thus can offer insights into neural computation. The operation of these circuits, like all models, critically depends on the choices of model parameters. A key step is then to identify the model parameters consistent with observed phenomena: to solve the inverse problem. To solve challenging inverse problems in neuroscience, statistical inference techniques have been used to infer parameters distributions most likely to have produced neural datasets. In this work, we present a novel technique, emergent property inference (EPI), that brings the power and versatility of the modern probabilistic modeling toolkit to theoretical neuroscience. When theorizing circuit models, scientists predominantly focus on reproducing computational properties rather than a particular dataset. Our method uses deep neural networks to learn parameter distributions with complex structure that produce specific computational properties in circuit models. This methodology is introduced through a motivational example inferring conductance parameters in a circuit model of the stomatogastric ganglion. Then, with recurrent neural networks of increasing size, we show that EPI allows precise control over the behavior of inferred parameters, and that EPI scales better in parameter dimension than alternative techniques. In the remainder of this work, we present novel theoretical findings gained through the examination of complex parametric structure captured by EPI. In a model of primary visual cortex, we discovered how connectivity with multiple inhibitory subtypes shapes variability in the

excitatory population. Finally, in a model of superior colliculus, we identified and characterized two distinct regimes of connectivity that facilitate switching between opposite tasks amidst interleaved trials, mechanistically characterized each regime using probabilistic tools afforded by EPI, and found conditions where these circuit models reproduce results from optogenetic silencing experiments. Beyond its scientific contribution, this work illustrates the variety of analyses possible once deep learning is harnessed towards solving theoretical inverse problems.

2 Introduction

The fundamental practice of theoretical neuroscience is to use a mathematical model to understand neural computation, whether that computation enables perception, action, or some intermediate processing. A neural circuit is systematized with a set of equations – the model – and these equations are motivated by biophysics, neurophysiology, and other conceptual considerations [1–5]. The function of this system is governed by the choice of model *parameters*, which when configured in a particular way, give rise to a measurable signature of a computation. The work of analyzing a model then requires solving the inverse problem: given a computation of interest, how can we reason about the space, manifold, or distribution of parameters that give rise to it? The inverse problem is crucial for reasoning about likely parameter values, uniquenesses and degeneracies, and predictions made by the model [6–8].

Ideally, one carefully designs a model and analytically derives how computational properties determine model parameters. Seminal examples of this gold standard include our field's understanding of memory capacity in associative neural networks [9], chaos and autocorrelation timescales in random neural networks [10], central pattern generation [11], the paradoxical effect [12], and decision making [13]. Unfortunately, as circuit models include more biological realism, theory via analytical derivation becomes intractable. Absent this analysis, statistical inference offers a toolkit by which to solve the inverse problem by identifying, at least approximately, the distribution of parameters that produce computations in a biologically realistic model [14–19].

Statistical inference, of course, requires quantification of the vague term *computation*. In neuroscience, two perspectives are dominant. First, often we directly use an *exemplar dataset*: a collection of samples that express the computation of interest, this data being gathered either experimentally in the lab or from a computer simulation. While in some sense the best choice given its connection to experiment [20], some drawbacks exist: these data are well known to have fea-

tures irrelevant to the computation of interest [21–23], confounding inferences made on such data. Related to this point, use of a conventional dataset encourages conventional data likelihoods or loss functions, which focus on some global metric like squared error or marginal evidence, rather than the computation itself.

Alternatively, researchers often quantify an *emergent property* (EP): a statistic of data that directly quantifies the computation of interest, wherein the dataset is implicit. While such a choice may seem esoteric, it is not: the above "gold standard" examples [9–13] all quantify and focus on some derived feature of the data, rather than the data drawn from the model. An emergent property is of course a dataset by another name, but it suggests different approach to solving the same inverse problem: here we directly specify the desired emergent property – a statistic of data drawn from the model – and the value we wish that property to have, and we set up an optimization program to find the distribution of parameters that produce this computation. This statistical framework is not new: it is intimately connected to the literature on approximate bayesian computation [24–26], parameter sensitivity analyses [27–30], maximum entropy modeling [31–33], and approximate bayesian inference [34, 35]; we detail these connections in Section 5.1.1.

The parameter distributions producing a computation may be thin, curved, bent, or multimodal along various parameter axes and combinations. It is by capturing and quantifying this complex structure that EPI offers scientific insight. Traditional approximation families (e.g. mean-field or mixture of gaussians) are limited in the distributional structure they may learn. To address such restrictions on expressivity, major advancements in machine learning have enabled the use of deep probability distributions as flexible approximating families for such complicated distributions [36,37] (see Section 5.1.2). However, the adaptation of deep probability distributions to the problem of theoretical circuit analysis requires recent developments in deep learning for constrained optimization [38], and architectural choices for efficient and expressive deep generative modeling [39, 40]. We detail our method, which we call emergent property inference (EPI) in Section 3.2.

Equipped with this method, we demonstrate the capabilities of EPI and present novel theoretical findings from its analysis. First, we show EPI's ability to handle biologically realistic circuit models using a five-neuron model of the stomatogastric ganglion [41]: a neural circuit whose parametric degeneracy is closely studied [42]. Then, we show EPI's scalability to high dimensional parameter distributions by inferring connectivities of recurrent neural networks (RNNs) that exhibit stable, yet amplified responses – a hallmark of neural responses throughout the brain [43–45]. In a model of primary visual cortex [46,47], EPI reveals how the recurrent processing across different neuron-type

populations shapes excitatory variability: a finding that we show is analytically intractable. Finally, we investigated the possible connectivities of superior colliculus that allow execution of different tasks on interleaved trials [48]. EPI discovered a rich distribution containing two connectivity regimes with different solution classes. We queried the deep probability distribution learned by EPI to produce a mechanistic understanding of neural responses in each regime. Intriguingly, the inferred connectivities of each regime reproduced results from optogenetic inactivation experiments in markedly different ways. These theoretical insights afforded by EPI illustrate the value of deep inference for the interrogation of neural circuit models.

3 Results

3.1 Motivating emergent property inference of theoretical models

Consideration of the typical workflow of theoretical modeling clarifies the need for emergent property inference. First, one designs or chooses an existing circuit model that, it is hypothesized, captures the computation of interest. To ground this process in a well-known example, consider the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic muscle activation patterns for digestion [49]. Despite full knowledge of STG connectivity and a precise characterization of its rhythmic pattern generation, biophysical models of the STG have complicated relationships between circuit parameters and computation [15, 42].

A subcircuit model of the STG [41] is shown schematically in Figure 1A. The fast population (f1 and f2) represents the subnetwork generating the pyloric rhythm and the slow population (s1 and s2) represents the subnetwork of the gastric mill rhythm. The two fast neurons mutually inhibit one another, and spike at a greater frequency than the mutually inhibiting slow neurons. The hub neuron couples with either the fast or slow population, or both depending on modulatory conditions. The jagged connections indicate electrical coupling having electrical conductance g_{el}, smooth connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto the hub neuron, and $g_{\text{synB}} = 5\text{nS}$ for mutual inhibitory connections. Note that the behavior of this model will be critically dependent on its parameterization – the choices of conductance parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$.

Second, once the model is selected, one must specify what the model should produce. In this STG model, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model (Fig. 1B). An emergent property studied by Gutierrez et al. is the hub neuron firing

at an intermediate frequency between the intrinsic spiking rates of the fast and slow populations. This emergent property is shown in Figure 1C at an average frequency of 0.55Hz. Our notion of intermediate hub frequency is not strictly 0.55Hz, but also moderate deviations of this frequency between the fast (.35Hz) and slow (.68Hz) frequencies.

Third, the model parameters producing the emergent property are inferred. By precisely quantifying the emergent property of interest as a statistical feature of the model, we use EPI to condition directly on this emergent property. Before presenting technical details (in the following section), let us understand emergent property inference schematically. EPI (Fig. 1D) takes, as input, the model and the specified emergent property, and as its output, returns the parameter distribution (Fig. 1E). This distribution – represented for clarity as samples from the distribution – is a parameter distribution constrained such that the circuit model produces the emergent property. Once EPI is run, the returned distribution can be used to efficiently generate additional parameter samples. Most importantly, the inferred distribution can be efficiently queried to quantify the parametric structure that it captures. By quantifying the parametric structure governing the emergent property, EPI informs the central question of this inverse problem: what aspects or combinations of model parameters have the desired emergent property?

3.2 A deep generative modeling approach to emergent property inference

Emergent property inference (EPI) formalizes the three-step procedure of the previous section with deep probability distributions [36, 37]. First, as is typical, we consider the model as a coupled set of noisy differential equations. In this STG example, the model activity (or state) $\mathbf{x} = [x_{\text{f1}}, x_{\text{f2}}, x_{\text{hub}}, x_{\text{s1}}, x_{\text{s2}}]$ is the membrane potential for each neuron, which evolves according to the biophysical conductance-based equation:

$$C_m \frac{d\mathbf{x}(t)}{dt} = -h(\mathbf{x}(t); \mathbf{z}) + d\mathbf{B} \tag{1}$$

where C_m=1nF, and $\mathbf{h}$ is a sum of the leak, calcium, potassium, hyperpolarization, electrical, and synaptic currents, all of which have their own complicated dependence on activity $\mathbf{x}$ and parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$, and $d\mathbf{B}$ is white gaussian noise [41] (see Section 5.2.1 for more detail).

Second, we determine that our model should produce the emergent property of "intermediate hub frequency" (Figure 1C). We stipulate that the hub neuron's spiking frequency – denoted by statistic $\omega_{\text{hub}}(\mathbf{x})$ – is close to a frequency of 0.55Hz, between that of the slow and fast frequencies. Mathematically, we define this emergent property with two constraints: that the mean hub frequency is

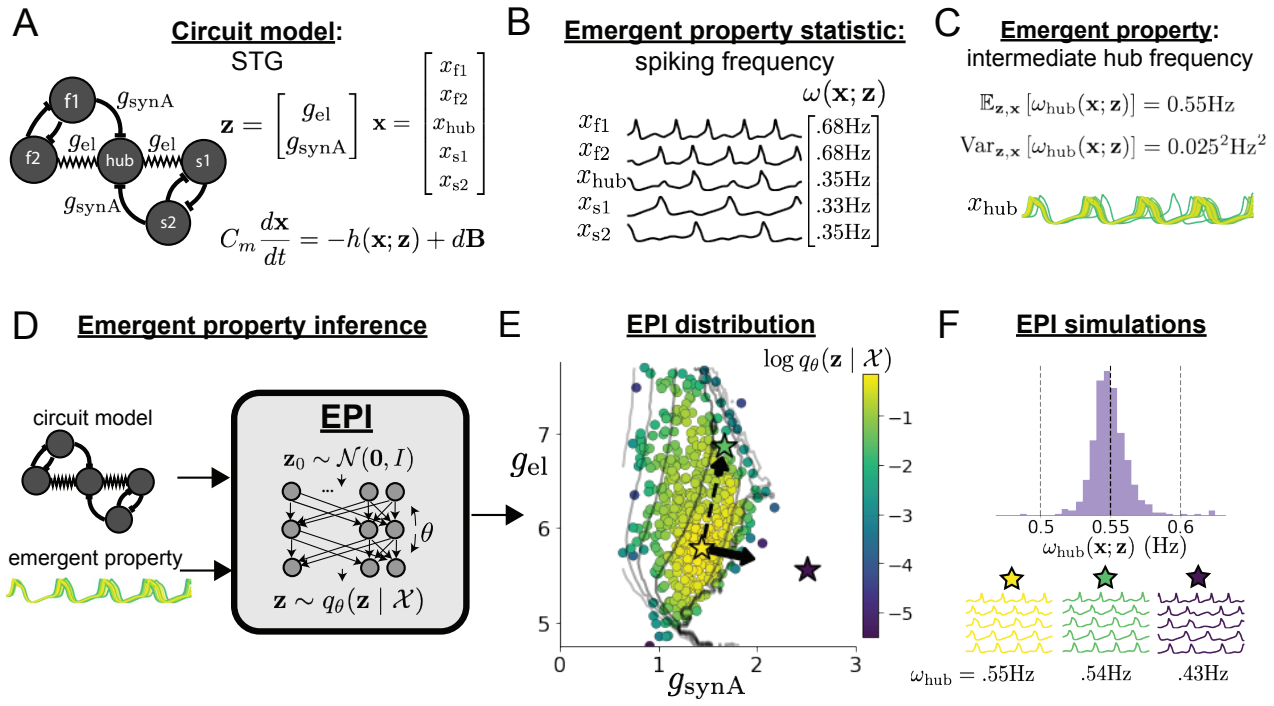

Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A**. Conductance-based subcircuit model of the STG. **B**. Spiking frequency $\omega(\mathbf{x};\mathbf{z})$ is an emergent property statistic. Simulated at $g_{\text{el}} = 4.5\text{nS}$ and $g_{\text{synA}} = 3\text{nS}$. **C**. The emergent property of intermediate hub frequency. Simulated activity traces are colored by log probability of generating parameters in the EPI distribution (Panel E). **D**. For a choice of circuit model and emergent property, emergent property inference (EPI) learns a deep probability distribution of parameters $\mathbf{z}$. **E**. The EPI distribution producing intermediate hub frequency. Samples are colored by log probability density. Contours of hub neuron frequency error are shown at levels of .525, .53, ... .575 Hz (dark to light gray away from mean). Dimension of sensitivity $\mathbf{v}_1$ (solid arrow) and robustness $\mathbf{v}_2$ (dashed arrow). **F** (Top) The predictions of the EPI distribution. The black and gray dashed lines show the mean and two standard deviations according the emergent property. (Bottom) Simulations at the starred parameter values.

0.55Hz,

$$\mathbb{E}_{\mathbf{z},\mathbf{x}}\left[\omega_{\text{hub}}(\mathbf{x};\mathbf{z})\right] \quad = \quad 0.55 \tag{2}$$

and that the variance of the hub frequency is moderate

$$\text{Var}_{\mathbf{z},\mathbf{x}}\left[\omega_{\text{hub}}(\mathbf{x};\mathbf{z})\right] \quad = \quad 0.025^2 \tag{3}$$

In the emergent property of intermediate hub frequency, the statistic of hub neuron frequency is constrained over the distribution of parameters $\mathbf{z}$ and the distribution of the data $\mathbf{x}$ that those parameters produce. This expectation over $\mathbf{x}$ is taken over any randomness in initial state ($\mathbf{x}(t=0)$) and noise (e.g. $d\mathbf{B}$ in Equation 1). Formally, the emergent property is the collection of these two constraints. In general, an emergent property is a collection of constraints on statistical moments that together define the computation.

Third, we perform emergent property inference: we find a distribution over parameter configurations $\mathbf{z}$ of models that produce the emergent property; in other words, they satisfy the constraints introduced in Equations 2 and 3. This distribution will be chosen from a family of probability distributions $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$, defined by a deep neural network [36, 37] (Figure 1D, EPI box). Deep probability distributions map a simple random variable $\mathbf{z}_0$ (e.g. an isotropic gaussian) through a deep neural network with weights and biases $\boldsymbol{\theta}$ to parameters $\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0)$ of a suitably complicated distribution (see Section 5.1.2 for more details). Many distributions in $\mathcal{Q}$ will respect the emergent property constraints, so we select the most random (highest entropy) distribution, which is the same choice commonly made in variational bayesian methods (see Section 5.1.6). In EPI optimization, stochastic gradient steps in $\boldsymbol{\theta}$ are taken such that entropy is maximized, and the emergent property $\mathcal{X}$ is produced (see Section 5.1). We then denote the inferred EPI distribution as $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$, since the structure of the learned parameter distribution is determined by weights and biases $\boldsymbol{\theta}$, and this distribution is conditioned upon emergent property $\mathcal{X}$.

The structure of the inferred parameter distributions of EPI can be analyzed to reveal key information about how the circuit model produces the emergent property. The modes of $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ indicate parameter choices emblematic of the emergent property (Fig. 1E yellow star). As probability in the EPI distribution decreases, the emergent property deteriorates. Perturbing $\mathbf{z}$ along a dimension in which $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ change little will not disturb the emergent property, making this parameter combination *robust* with respect to the emergent property. In contrast, if $\mathbf{z}$ is perturbed along a dimension with strongly decreasing $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$, that parameter combination is deemed *sensitive* [27, 30]. By querying the second order derivative (Hessian) of $\log q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ at a mode, we

can quantitatively identify how sensitive (or robust) each eigenvector is by its eigenvalue; the more negative, the more sensitive and the closer to zero, the more robust (see Section 5.2.4). Indeed, samples equidistant from the mode along these dimensions of sensitivity ($\mathbf{v}_1$, smaller eigenvalue) and robustness ($\mathbf{v}_2$, greater eigenvalue) (Fig. 1E, arrows) agree with error contours (Fig. 1E contours) and have diminished or preserved hub frequency, respectively (Fig. 1F activity traces). The directionality of $\mathbf{v}_2$ suggests that changes in conductance along this parameter combination will most preserve hub neuron firing between the intrinsic rates of the pyloric and gastric mill rhythms. Importantly, once an EPI distribution has been learned, the modes and Hessians of the distribution can be measured with trivial computation (see Section 5.1.2).

In the following sections, we demonstrate EPI on three neural circuit models across ranges of biological realism, neural system function, and network scale. First, we demonstrate the superior scalability of EPI compared to alternative techniques by inferring high-dimensional distributions of recurrent neural network connectivities that exhibit amplified, yet stable responses. Next, in a model of primary visual cortex [46,47], we show how EPI discovers parametric degeneracy, revealing how input variability across neuron types affects the excitatory population. Finally, in a model of superior colliculus [48], we used EPI to capture multiple parametric regimes of task switching, and queried the dimensions of parameter sensitivity to characterize each regime.

3.3 Scaling inference of recurrent neural network connectivity with EPI

To understand how EPI scales in comparison to existing techniques, we consider recurrent neural networks (RNNs). Transient amplification is a hallmark of neural activity throughout cortex, and is often thought to be intrinsically generated by recurrent connectivity in the responding cortical area [43–45]. It has been shown that to generate such amplified, yet stabilized responses, the connectivity of RNNs must be non-normal [43, 50], and satisfy additional constraints [51]. In theoretical neuroscience, RNNs are optimized and then examined to show how dynamical systems could execute a given computation [52, 53], but such biologically realistic constraints on connectivity [43, 50, 51] are ignored for simplicity or because constrained optimization is difficult. In general, access to distributions of connectivity that produce theoretical criteria like stable amplification, chaotic fluctuations [10], or low tangling [54] would add scientific value to existing research with RNNs. Here, we use EPI to learn RNN connectivities producing stable amplification, and demonstrate the superior scalability and efficiency of EPI to alternative approaches.

We consider a rank-2 RNN with N neurons having connectivity $W = UV^\top$ and dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}, \tag{4}$$

where $U = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} + g\chi^{(U)}$, $V = \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix} + g\chi^{(V)}$, $\mathbf{U}_1\mathbf{U}_2, \mathbf{V}_1, \mathbf{V}_2 \in [-1, 1]^N$, and $\chi^{(U)}_{i,j}, \chi^{(V)}_{i,j} \sim \mathcal{N}(0, 1)$. We infer connectivity parameters $\mathbf{z} = \left[\mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{V}_1^\top, \mathbf{V}_2^\top\right]^\top$ that produce stable amplification. Two conditions are necessary and sufficient for RNNs to exhibit stable amplification [51]: $\text{real}(\lambda_1) < 1$ and $\lambda_1^s > 1$, where λ_1 is the eigenvalue of W with greatest real part and λ^s is the maximum eigenvalue of $W^s = \frac{W+W^\top}{2}$. RNNs with $\text{real}(\lambda_1) = 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$ will be stable with modest decay rate ($\text{real}(\lambda_1)$ close to its upper bound of 1) and exhibit modest amplification (λ_1^s close to its lower bound of 1). EPI can naturally condition on this emergent property

$$\begin{aligned} \mathcal{X} \quad : \quad \mathbb{E}_{\mathbf{z},\mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z},\mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}. \end{aligned} \tag{5}$$

Variance constraints predicate that the majority of the distribution (within two standard deviations) are within the specified ranges.

For comparison, we infer the parameters $\mathbf{z}$ likely to produce stable amplification using two alternative simulation-based inference approaches. Sequential Monte Carlo approximate Bayesian computation (SMC-ABC) [26] is a rejection sampling approach that uses SMC techniques to improve efficiency, and sequential neural posterior estimation (SNPE) [35] approximates posteriors with deep probability distributions using a two-network architecture (see Section 5.1.1). Unlike EPI, these statistical inference techniques do not constrain the statistics of the predictive distribution, so they were run by conditioning on an exemplary dataset $\mathbf{x}_0 = \boldsymbol{\mu}$, following standard practice with these methods [26, 35]. To compare the efficiency of these different techniques, we measured the time and number of simulations necessary for the distance of the predictive mean to be less than 0.5 from $\boldsymbol{\mu} = \mathbf{x}_0$ (see Section 5.3).

As the number of neurons N in the RNN, and thus the dimension of the parameter space $\mathbf{z} \in [-1, 1]^{4N}$, is scaled, we see that EPI converges at greater speed and at greater dimension than SMC-ABC and SNPE (Fig. 2A). It also becomes most efficient to use EPI in terms of simulation count at $N = 50$ (Fig. 2B). It is well known that ABC techniques struggle in parameter spaces of modest dimension [55], yet we were careful to assess the scalability of SNPE, which is a more closely related methodology to EPI. Between EPI and SNPE, we closely controlled the number of

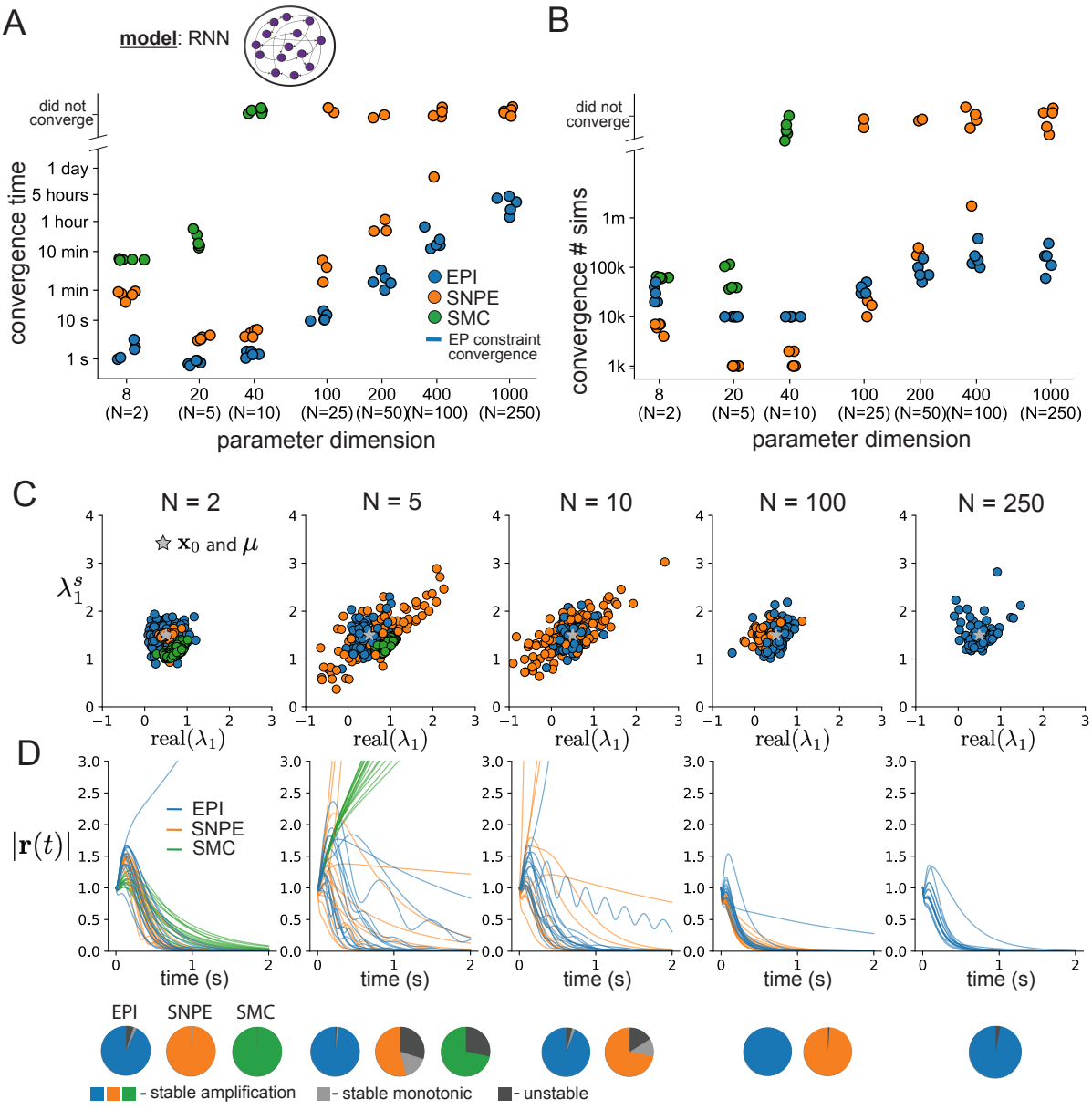

Figure 2: **A**. Wall time of EPI (blue), SNPE (orange), and SMC-ABC (green) to converge on RNN connectivities producing stable amplification. Each dot shows convergence time for an individual random seed. For reference, the mean wall time for EPI to achieve its full constraint convergence (means and variances) is shown (blue line). **B**. Simulation count of each algorithm to achieve convergence. Same conventions as A. **C**. The predictive distributions of connectivities inferred by EPI (blue), SNPE (orange), and SMC-ABC (green), with reference to $\mathbf{x}_0 = \boldsymbol{\mu}$ (gray star). **D**. Simulations of networks inferred by each method ($\tau = 100ms$). Each trace (15 per algorithm) corresponds to simulation of one z. (Below) Ratio of obtained samples producing stable amplification, monotonic decay, and instability.

parameters in deep probability distributions by dimensionality (Fig. 10), and tested more aggressive SNPE hyperparameterizations when SNPE failed to converge (Fig. 11). In this analysis, we see that deep inference techniques EPI and SNPE are far more amenable to inference of high dimensional RNN connectivities than rejection sampling techniques like SMC-ABC, and that EPI outperforms SNPE in both wall time (elapsed real time) and simulation count.

No matter the number of neurons, EPI always produces connectivity distributions with mean and variance of real(λ_1) and λ_1^s according to $\mathcal{X}$ (Fig. 2C, blue). For the dimensionalities in which SMC-ABC is tractable, the inferred parameters are concentrated and offset from the exemplary dataset $\mathbf{x}_0$ (Fig. 2C, green). When using SNPE, the predictions of the inferred parameters are highly concentrated at some RNN sizes and widely varied in others (Fig. 2C, orange). We see these properties reflected in simulations from the inferred distributions: EPI produces a consistent variety of stable, amplified activity norms $|\mathbf{x}(t)|$, SMC-ABC produces a limited variety of responses, and the changing variety of responses from SNPE emphasizes the control of EPI on parameter predictions (Fig. 2D). Even for moderate neuron counts, the predictions of the inferred distribution of SNPE are highly dependent on N and g, while EPI maintains the emergent property across choices of RNN (see Section 5.3.5).

To understand these differences, note that EPI outperforms SNPE in high dimensions by using gradient information (from $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z}) = \nabla_{\mathbf{z}}[\text{real}(\lambda_1), \lambda_1^s]^\top$). This choice agrees with recent speculation that such gradient information could improve the efficiency of simulation-based inference techniques [56], as well as reflecting the classic tradeoff between gradient and sampling efficiency (scaling and speed versus generality). Since gradients of the emergent property statistics are necessary in EPI optimization, gradient tractability is a key criteria when determining the suitability of a simulation-based inference technique. If the emergent property gradient is efficiently calculated, EPI is a clear choice for inferring high dimensional parameter distributions. In the next two sections, we use EPI for novel scientific insight by examining the structure of inferred distributions.

3.4 EPI reveals how recurrence with multiple inhibitory subtypes governs excitatory variability in a V1 model

Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output function have succeeded in explaining a host of experimentally documented phenomena in primary visual cortex (V1). In a regime characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to paradoxical responses [12], selective amplification [43, 50], surround

suppression [58] and normalization [59]. Recent theoretical work [60] shows that stabilized E-I models reproduce the effect of variability suppression [61]. Furthermore, experimental evidence shows that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) – composing 80% of GABAergic interneurons in V1 [63–65], and that these inhibitory cell types follow specific connectivity patterns (Fig. 3A) [66]. Here, we use EPI on a model of V1 with biologically realistic connectivity to show how the structure of input across neuron types affects the variability of the excitatory population – the population largely responsible for projecting to other brain areas [67].

We considered response variability of a nonlinear dynamical V1 circuit model (Fig. 3A) with a state comprised of each neuron-type population's rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population receives recurrent input $W\mathbf{x}$, where W is the effective connectivity matrix (see Section 5.4) and an external input with mean $\mathbf{h}$, which determines population rate via supralinear nonlinearity $\phi(\cdot) = [\cdot]_+^2$. The external input has an additive noisy component $\boldsymbol{\epsilon}$ with variance $\boldsymbol{\sigma}^2 = \left[\sigma_E^2, \sigma_P^2, \sigma_S^2, \sigma_V^2\right]$. This noise has a slower dynamical timescale $\tau_{\text{noise}} > \tau$ than the population rate, allowing fluctuations around a stimulus-dependent steady-state (Fig. 3B). This model is the stochastic stabilized supralinear network (SSSN) [60]

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \tag{6}$$

generalized to have multiple inhibitory neuron types. It introduces stochasticity to four neuron-type models of V1 [46]. Stochasticity and inhibitory multiplicity introduce substantial complexity to the mathematical treatment of this problem (see Section 5.4.4) motivating the analysis of this model with EPI. Here, we consider fixed weights W and input $\mathbf{h}$ [47], and study the effect of input variability $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on excitatory variability.

We quantify levels of E-population variability by studying two emergent properties

$$\begin{aligned} \mathcal{X}(5\text{ Hz}) \;:\; \mathbb{E}_{\mathbf{z}}\left[s_E(\mathbf{x};\mathbf{z})\right] &= 5\text{ Hz} \qquad \mathcal{X}(10\text{ Hz}) \;:\; \mathbb{E}_{\mathbf{z}}\left[s_E(\mathbf{x};\mathbf{z})\right] = 10\text{ Hz} \\ \text{Var}_{\mathbf{z}}\left[s_E(\mathbf{x};\mathbf{z})\right] &= 1\text{Hz}^2 \qquad\qquad \text{Var}_{\mathbf{z}}\left[s_E(\mathbf{x};\mathbf{z})\right] = 1\text{Hz}^2, \end{aligned} \tag{7}$$

where $s_E(\mathbf{x};\mathbf{z})$ is the standard deviation of the stochastic E-population response about its steady state (Fig. 3C). In the following analyses, we select 1 Hz^2 variance such that the two emergent properties do not overlap in $s_E(\mathbf{z};\mathbf{x})$.

First, we ran EPI to obtain parameter distribution $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}(5\text{ Hz}))$ producing E-population variability around 5 Hz (Fig. 3D). From the marginal distribution of σ_E and σ_P (Fig. 3D, top-left), we can see that $s_E(\mathbf{x};\mathbf{z})$ is sensitive to various combinations of σ_E and σ_P. Alternatively, both σ_S

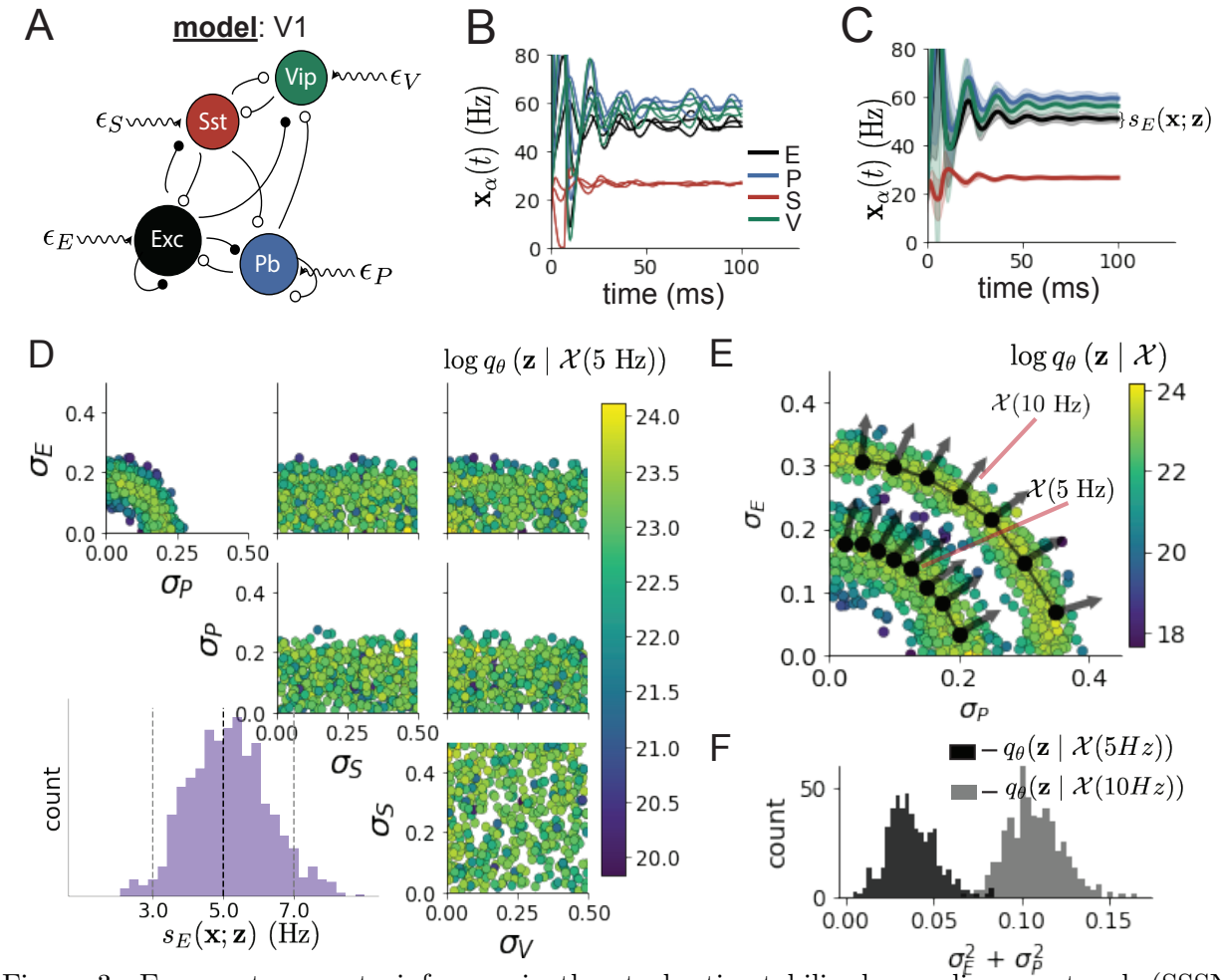

Figure 3: Emergent property inference in the stochastic stabilized supralinear network (SSSN) **A**. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). Some neuron-types largely do not form synaptic projections to others ($|W_{\alpha_1,\alpha_2}| < 0.025$). Each neural population receives a baseline input $\mathbf{h}_b$, and the E- and P-populations also receive a contrast-dependent input $\mathbf{h}_c$. Additionally, each neural population receives a slow noisy input $\boldsymbol{\epsilon}$. **B**. Transient network responses of the SSSN model. Traces are independent trials with varying initialization $\mathbf{x}(0)$ and noise $\boldsymbol{\epsilon}$. **C**. Mean (solid line) and standard deviation $s_E(\mathbf{x};\mathbf{z})$ (shading) across 100 trials. **D**. EPI distribution of noise parameters $\mathbf{z}$ conditioned on E-population variability. The EPI predictive distribution of $s_E(\mathbf{x};\mathbf{z})$ is show on the bottom-left. **E**. (Top) Enlarged visualization of the σ_E-σ_P marginal distribution of EPI $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}(5\text{ Hz})$ and $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}(10\text{ Hz})$. Each black dot shows the mode at each σ_P. The arrows show the most sensitive dimensions of the Hessian evaluated at these modes. **F**. The predictive distributions of $\sigma_E^2 + \sigma_P^2$ of each inferred distribution $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}(5\text{ Hz})$ and $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}(10\text{ Hz})$.

and σ_V are degenerate with respect to $s_E(\mathbf{x};\mathbf{z})$ evidenced by the, unexpectedly, high variability in those dimensions (Fig. 3D, bottom-right). Together, these observations imply a curved path with respect to $s_E(\mathbf{x};\mathbf{z})$ of 5 Hz, which is indicated by the modes along σ_P (Fig. 3E).

Figure 3E suggests a quadratic relationship in E-population fluctuations and the standard deviation of E- and P-population input; as the square of either σ_E or σ_P increases, the other compensates by decreasing to preserve the level of $s_E(\mathbf{x};\mathbf{z})$. This quadratic relationship is preserved at greater level of E-population variability $\mathcal{X}(10\text{ Hz})$ (Fig. 3E). Indeed, the sum of squares of σ_E and σ_P is larger in $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}(10\text{ Hz}))$ than $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}(5\text{ Hz}))$ (Fig 3F, $p < 1 \times 10^{-10}$), while the sum of squares of σ_S and σ_V are not significantly different in the two EPI distributions (Fig. 15, $p = .40$), in which parameters were bounded from 0 to 0.5. The strong interaction between E- and P-population input variability on excitatory variability is intriguing, since this circuit exhibits a paradoxical effect in the P-population (and no other inhibitory types) (Fig. 15), meaning that the E-population is P-stabilized. Future research may uncover a link between the population of network stabilization and compensatory interactions governing excitatory variability.

EPI revealed the quadratic dependence of excitatory variability on input variability to the E- and P-populations, as well as its independence to input from the other two inhibitory populations. In a simplified model ($\tau = \tau_{\text{noise}}$), it can be shown that surfaces of equal variance are ellipsoids as a function of $\boldsymbol{\sigma}$ (see Section 5.4.4). Nevertheless, the sensitive and degenerate parameters are challenging to predict mathematically, since the covariance matrix depends on the steady-state solution of the network [60,68], and terms in the covariance expression increase quadratically with each additional neuron-type population (see also Section 5.4.4). By pointing out this mathematical complexity, we emphasize the value of streamlined methods for gaining understanding about theoretical models when mathematical analysis becomes onerous or impractical. While we have just shown that EPI can be used to investigate fundamental aspects of sensory computation, in the next two sections, we use the probabilistic tools of EPI to identify and characterize two distinct parametric regimes of a neural circuit executing a computation, and then relate these insights to behavioral experiments.

3.5 EPI identifies two regimes of rapid task switching

It has been shown that rats can learn to switch from one behavioral task to the next on randomly interleaved trials [69], and an important question is what types of neural connectivity produce this computation. In this experimental setup, rats were explicitly cued on each trial to either orient towards a visual stimulus in the Pro (P) task or orient away from the stimulus in the Anti (A)

task (Fig. 4A). Neural recordings in superior colliculus (SC) exhibited two populations of neurons that represented task context (Pro or Anti). Furthermore, Pro/Anti neurons in each hemisphere were strongly correlated with the animal's decision [48]. These results motivated a model of SC that is a four-population dynamical system with functionally-defined neuron-types. Here, our goal is to understand how connectivity in this circuit model governs the ability to perform rapid task switching: to respond with satisfactory accuracy in both tasks on randomly interleaved trials.

In this SC model, there are Pro- and Anti-populations in each hemisphere (left (L) and right (R)) with activity variables $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$. The connectivity of these populations is parameterized by self sW, vertical vW, diagonal dW and horizontal hW connections (Fig. 4B). The input $\mathbf{h}$ is comprised of a positive cue-dependent signal to the Pro or Anti populations, a positive stimulus-dependent input to either the Left or Right populations, and a choice-period input to the entire network (see Section 5.5.1). Model responses are bounded from 0 to 1 as a function ϕ of an internal variable $\mathbf{u}$

$$\begin{aligned} \tau \frac{d\mathbf{u}}{dt} &= -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B} \\ \mathbf{x} &= \phi(\mathbf{u}). \end{aligned} \tag{8}$$

The model responds to the side with greater Pro neuron activation; e.g. the response is left if $x_{LP} > x_{RP}$ at the end of the trial. Here, we use EPI to determine the network connectivity $\mathbf{z} = [sW, vW, dW, hW]^\top$ that produces rapid task switching.

Rapid task switching is formalized mathematically as an emergent property with two statistics: accuracy in the Pro task $p_P(\mathbf{x}; \mathbf{z})$ and Anti task $p_A(\mathbf{x}; \mathbf{z})$. We stipulate that accuracy be on average .75 in each task with variance $.075^2$

$$\begin{aligned} \mathcal{X} \quad : \quad \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .75 \\ .75 \end{bmatrix} \\ \mathrm{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .075^2 \\ .075^2 \end{bmatrix}. \end{aligned} \tag{9}$$

75% accuracy is a realistic level of performance in each task, and with the chosen variance, inferred models will not exhibit fully random responses (50%), nor perfect performance (100%).

The EPI inferred distribution (Fig. 4C) produces Pro and Anti task accuracies (Fig. 4C, middle-left) consistent with rapid task switching (Equation 9). This parameter distribution has rich structure, that is not captured well by simple linear correlations (Fig. 17). Specifically, the shape of the EPI distribution is sharply bent, matching ground truth structure indicated by brute-force

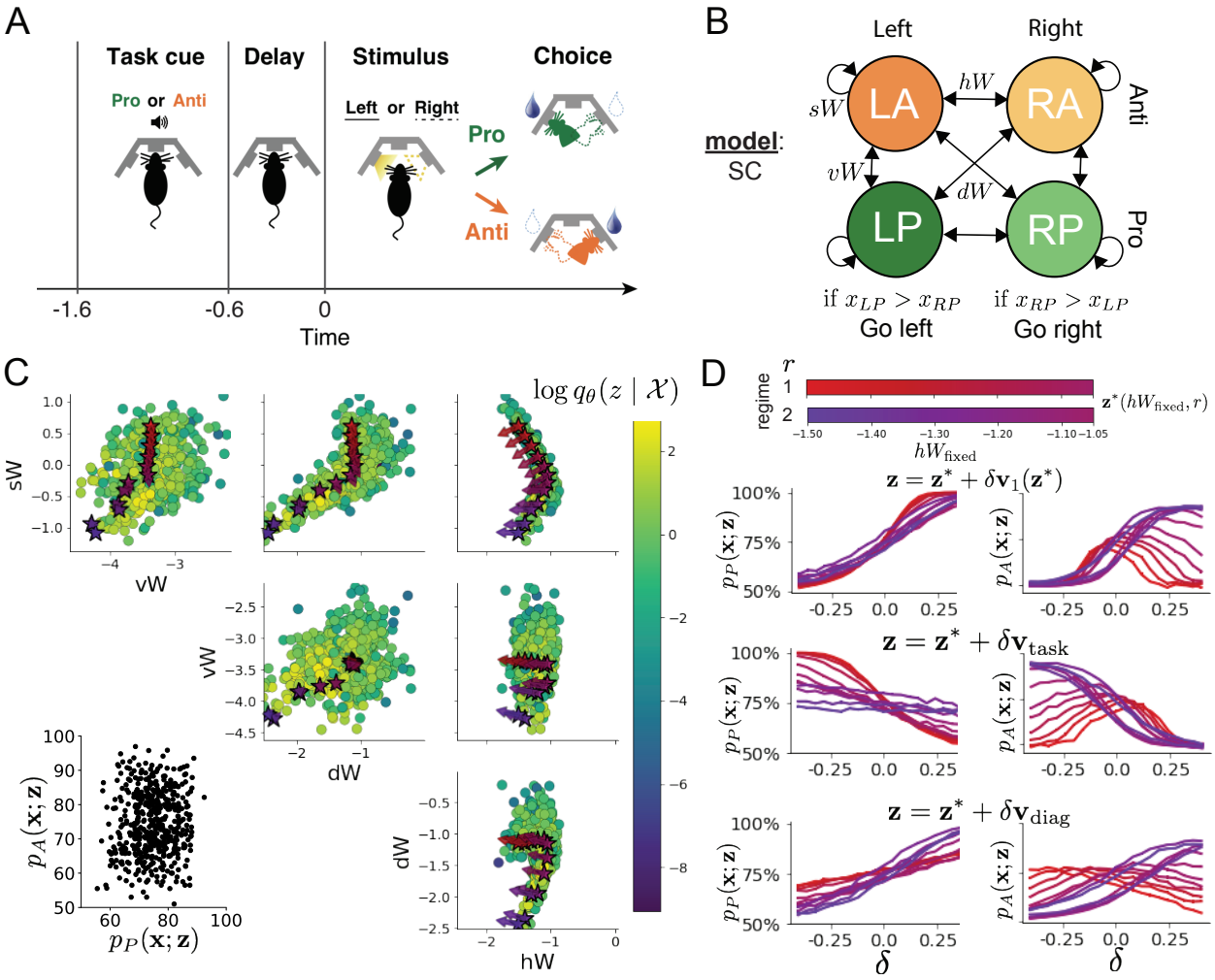

Figure 4: **A**. Rapid task switching behavioral paradigm (see text). **B**. Model of superior colliculus (SC). Neurons: LP - Left Pro, RP - Right Pro, LA - Left Anti, RA - Right Anti. Parameters: sW - self, hW - horizontal, vW -vertical, dW - diagonal weights. **C**. The EPI inferred distribution of rapid task switching networks. Stars indicate modes $\mathbf{z}^*$ whose color indicates connectivity regime (see legend Fig 4D). Sensitivity vectors $\mathbf{v}_1(\mathbf{z}^*)$ are shown by arrows. (Bottom-left) EPI predictive distribution of task accuracies. **D**. The connectivity regimes have different responses to perturbation. (Top) Mean and standard error ($N_{\text{test}} = 25$) of accuracy with respect to perturbation along the sensitivity dimension of each mode $\mathbf{z}^*$. (Middle) Same with perturbation in the dimension of increasing λ_{task} ($\mathbf{v}_{\text{task}}$). (Bottom) Same with perturbation in the dimension of increasing λ_{diag} ($\mathbf{v}_{\text{diag}}$).

sampling (Fig. 24) This is most saliently pointed out in the marginal distribution of sW-hW (Fig. 4C top-right), where anticorrelation between sW and hW switches to correlation with decreasing sW. The two regimes produce different types of responses in the Pro and Anti tasks (Fig. SC2). Not only has EPI captured this complicated distribution of connectivities producing rapid task switching, we can query the EPI distribution $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ to understand these two parametric regimes of SC connectivity.

To distinguish these two regimes, we use the EPI distribution to identify two sets of modes. By fixing hW to different values and doing gradient ascent on $\log q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$, we arrive at two solutions $\mathbf{z}^*(hW_{\text{fixed}}, r)$ where regime $r \in [1, 2]$, and regime 1 is that of greater sW (see Section 5.5.4). As hW_{fixed} increases, the modes coalesce to intermediate parameters reflecting a transition between the two sets of modes (Fig. 20 top). By using EPI to connect these two regimes through this transitionary region of parameter space, we can explore what distinguishes the two regimes by stepping from the prototypical connectivity of regime 1 to that of regime 2.

Although the connectivities gradually coalesce to the transitionary part of parameter space, the sensitivity dimensions $\mathbf{v}_1(\mathbf{z})$ are categorically different across regimes (Fig. 20 bottom). The sensitivity dimension identifies the parameter combination which causes the emergent property to diminish with the shortest perturbation. Since the two regimes have different $\mathbf{v}_1(\mathbf{z})$, this suggests they have different pathologies in their connectivity. By perturbing connectivity in each regime along the sensitivity dimension, we can get a sense of the differing nature of these pathologies.

When perturbing connectivity along the sensitivity dimension, Pro accuracy monotonically increases in both regimes (Fig. 4D, top-left). However, there is a stark difference between regimes in Anti accuracy. Anti accuracy falls in either direction of $\mathbf{v}_1$ in regime 1, yet monotonically increases along with Pro accuracy in regime 2 (Fig. 4D, top-right). These distinct pathologies of rapid task switching are caused by distinct connectivity changes ($\mathbf{v}_1(\mathbf{z}^*(r = 1))$ vs $\mathbf{v}_1(\mathbf{z}^*(r = 2))$) and explain the sharp change in local structure of the EPI distribution.

To further examine the two regimes, we can perturb $\mathbf{z}$ in the same way along dimensions that independently change the eigenvalues of the connectivity matrix (which has constant eigenvectors with respect to $\mathbf{z}$). These eigenvalues λ_{all}, λ_{side}, λ_{task}, and λ_{diag} correspond to connectivity eigenmodes with intuitive roles in processing in this task (Fig. 19A). For example, greater λ_{task} will strengthen internal representations of task, while greater λ_{diag} will amplify dominance of Pro and Anti pairs in opposite hemispheres (Section 5.5.6). Perturbation analyses reveal that decreasing λ_{task} has a very similar effect on Anti accuracy as perturbations along the sensitivity dimension (Fig. 4D, middle).

This suggests that there is a carefully tuned strength of task representation in connectivity regime 1, which if disturbed results in random Anti trial responses. Finally, we recognize that increasing λ_{diag} has opposite effects on Anti accuracy in each regime (Fig. 4D, bottom). In the next section, we build on these mechanistic characterizations of each regime by examining their resilience to optogenetic silencing.

3.6 EPI inferred SC connectivities reproduce results from optogenetic inactivation experiments

During the delay period of this task, the circuit must prepare to execute the correct task according to the presented cue. Experimental results from Duan et al. found that bilateral optogenetic inactivation of SC during the delay period consistently decreased performance in the Anti task, but had no effect on the Pro task (Fig. 5A). This suggests that SC maintains a representation of task throughout the delay period, which is important for correct execution of the Anti task. Network connectivities inferred by EPI exhibited this same effect in simulation at high optogenetic strengths γ (Fig. 5B) (see Section 5.5.7).

The mean increase in Anti error is closest to the experimentally measured value of 7% at $\gamma = 0.675$ (Fig. 5B, black dot). At this level of optogenetic strength, only regime 1 exhibits an increase in Anti error with delay period silencing (Fig. 5C, left). The connectivities in regime 2 are thus more resilient to delay period silencing during Anti trials than regime 1. In regime 1, greater λ_{task} and λ_{diag} decrease Anti error (Fig. 5C, right). In other words, these anticorrelations show that stronger task representations and diagonal amplification make the SC model more resilient to delay period silencing in the Anti task. All correlations of connectivity eigenvalue with Anti error degrade in regime 2, where there is no effect of delay period silencing on Anti error (Fig. 5C, right).

At about $\gamma = 0.85$ (Fig. 5B, gray dot), the Anti error saturates, while Pro error remains at zero Following delay period inactivation at this optogenetic strength, there are strong similarities in the responses of Pro and Anti trials during the choice period (Fig. 5D, left). We interpreted these similarities to suggest that delay period inactivation at this saturated level flips the internal representation of task (from Anti to Pro) in the circuit model. This would explain why the Anti error saturates at 50%: the average Anti accuracy in EPI inferred connectivities is 75%, but is 25% when the internal representation is flipped during delay period silencing. This hypothesis prescribes a model of Anti accuracy during delay period silencing of $p_{A,\text{opto}} = 100\% - p_P$, which is fit closely across both regimes of the EPI inferred connectivities (Fig. 5D, right). Similarities between Pro

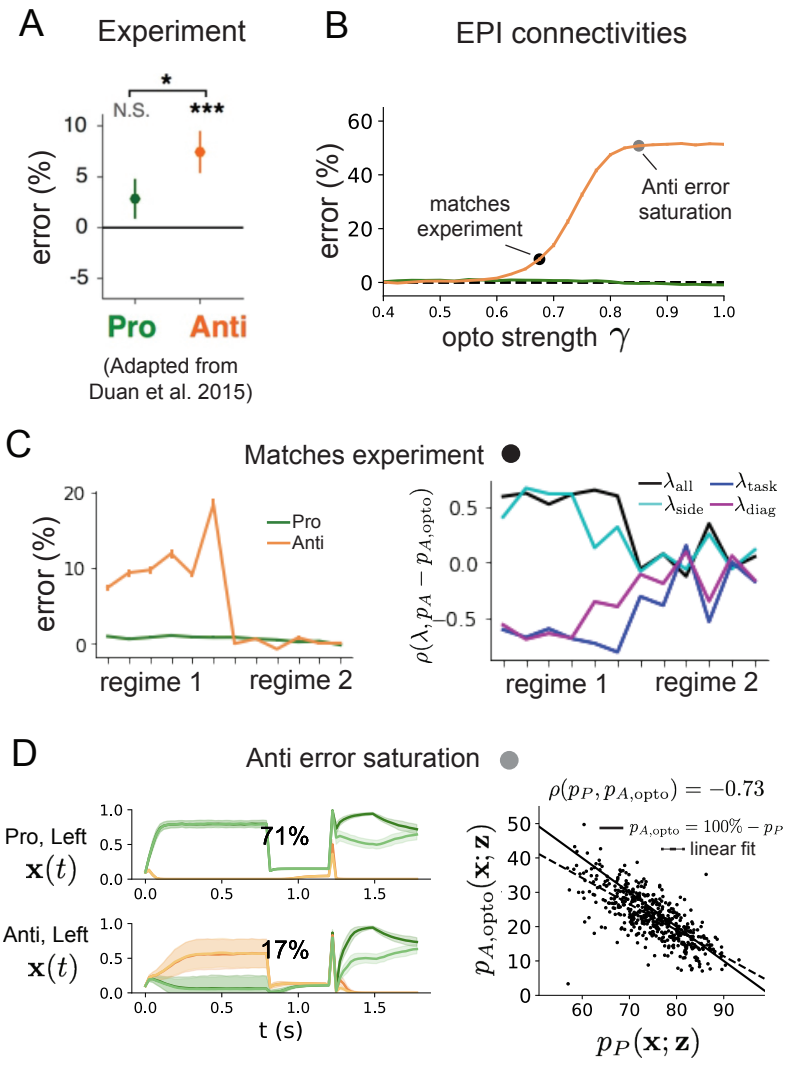

Figure 5: **A**. Experimental effect of delay period silencing on Pro and Anti task accuracy in rats. **B**. Mean and standard error (bars) of task error induced by delay period inactivation of varying optogenetic strength across the EPI distribution. **C**. (Left) Mean and standard error of Pro and Anti error from regime 1 to regime 2 at $\gamma = 0.675$. (Right) Correlations of connectivity eigenvalues with Anti error from regime 1 to regime 2 at $\gamma = 0.675$. **D**. (Left) Mean and standard deviation (shading) of responses of the SC model at the mode of the EPI distribution to delay period inactivation at $\gamma = 0.85$. (Right) Anti accuracy following delay period inactivation at $\gamma = 0.85$ versus accuracy in the Pro task across connectivities in the EPI distribution.

and Anti trial responses were not present at the experiment-matching level of $\gamma = 0.675$ (Fig. 22 left) and neither was anti-correlation in p_P and $p_{A,\text{opto}}$ (Fig. 22 right).

In summary, the connectivity inferred by EPI to perform rapid task switching replicated results from optogenetic silencing experiments. We found that at levels of optogenetic strength matching experimental levels of Anti error, only one regime actually exhibited the effect. This suggests that one regime is less resilient to optogenetic perturbation, and perhaps more biologically realistic. Finally, we mechanistically characterized the pathology in Anti error that occurs in both regimes when optogenetic strength is increased to high levels. The probabilistic tools afforded by EPI yielded this insight: we identified two regimes and the continuum of connectivities between them by taking gradients of parameter probabilities in the EPI distribution, we identified sensitivity dimensions by measuring the Hessian of the EPI distribution, and we obtained many parameter samples at each step along the continuum (in 7.36 seconds with the EPI distribution rather than 4.2 days with brute force methods, see Section 5.5).

4 Discussion

In neuroscience, machine learning has primarily been used to reveal structure in neural datasets [20]. Careful inference procedures are developed for these statistical models allowing precise, quantitative reasoning, which clarifies the way data informs beliefs about the model parameters. However, these statistical models often lack resemblance to the underlying biology, making it unclear how to go from the structure revealed by these methods, to the neural mechanisms giving rise to it. In contrast, theoretical neuroscience has primarily focused on careful models of neural circuits and the production of emergent properties of computation, rather than measuring structure in neural datasets. In this work, we improve upon parameter inference techniques in theoretical neuroscience with emergent property inference, harnessing deep learning towards parameter inference with respect to computation in neural circuit models (see Section 5.1.1).

Methodology for statistical inference in circuit models has evolved considerably in recent years. Early work used rejection sampling techniques [24–26], but EPI and other recently developed methodology (SNPE [35]) employ deep learning to improve efficiency and provide deep, flexible distribution approximations. SNPE has been used for posterior inference of parameters in circuit models conditioned upon exemplar data used to represent computation, but it does not infer parameter distributions that only produce the computation of interest like EPI (see Section 3.3). Finally,

we show that EPI has better scaling properties than SNPE when emergent property gradients are tractable (Section 3.3). In summary, choice of deep inference technique should consider emergent property complexity and differentiability, dimensionality of parameter space, and the importance of constraining the model behavior predicted by the inferred parameter distribution.

In this paper, we prove out the value of deep inference for parameter sensitivity analyses at both the local and global level. With these techniques, flexible deep probability distributions are optimized to capture global structure by approximating the full distribution of suitable parameters. Importantly, the local structure of this deep probability distribution can be quantified at any parameter choice, offering instant sensitivity measurements after fitting. For example, the global structure captured by EPI revealed two distinct parameter regimes, which had different local structure quantified by the deep probability distribution (see Section 5.5). In comparison, Bayesian MCMC is considered a popular approach for capturing global parameter structure [70], but there is no variational approximation (the deep probability distribution in EPI), so sensitivity information is not queryable and sampling remains slow after convergence. Local sensitivity analyses (e.g. [27]) may be performed independently at individual parameter samples, but these methods alone do not capture the full picture in nonlinear, complex distributions. Therefore, deep inference is the only approach yielding an object – the deep probability distribution – which produces a wholistic assessment of parameter sensitivity at the local and global level, which were used in this study to make novel insights into theoretical models of neural computation. Together, the abilities to condition upon emergent properties, the efficient inference algorithm, and the capacity for parameter sensitivity analyses make EPI a powerful new method for addressing inverse problems in theoretical neuroscience.

Even with a high degree of biophysical realism and expensive emergent property gradients, EPI was run successfully on intermediate hub frequency in a 5-neuron subcircuit model of the STG (Section 3.1). However, conditioning on the pyloric rhythm [57] in a model of the pyloric subnetwork model [15] proved to be prohibitive with EPI. The pyloric subnetwork requires many time steps for simulation and many key emergent property statistics (e.g. burst duration and phase gap) are not calculable or easily approximated with differentiable functions. In such cases, SNPE, which does not require differentiability of the emergent property has proved to be a powerful approach [35].

Acknowledgements:

This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endowment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-

toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

Data availability statement:

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Code availability statement:

All software written for the current study is available at https://github.com/cunningham-lab/epi.

References

[1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.

[2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*, 21(1):25–45, 1998.

[3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.

[4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological reviews*, 90(3):1195–1268, 2010.

[5] Timothy O'Leary, Alexander C Sutton, and Eve Marder. Computational models in the age of large datasets. *Current opinion in neurobiology*, 32:87–94, 2015.

[6] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):e189, 2007.

[7] Kamil Erguler and Michael PH Stumpf. Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Molecular BioSystems*, 7(5):1593–1602, 2011.

[8] Brian K Mannakee, Aaron P Ragsdale, Mark K Transtrum, and Ryan N Gutenkunst. Sloppiness and the geometry of parameter space. In *Uncertainty in Biology*, pages 271–299. Springer, 2016.

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

[10] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[11] Andrey V Olypher and Ronald L Calabrese. Using constraints on neuronal activity to reveal compensatory changes in neuronal parameters. *Journal of Neurophysiology*, 98(6):3749–3758, 2007.

[12] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17(11):4382–4388, 1997.

[13] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.

[14] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.

[15] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.

[16] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron model. *PLoS computational biology*, 2(7):e94, 2006.

[17] Dimitry Fisher, Itsaso Olasagasti, David W Tank, Emre RF Aksay, and Mark S Goldman. A modeling framework for deriving the structural and functional architecture of a short-term memory microcircuit. *Neuron*, 79(5):987–1000, 2013.

[18] Timothy O'Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network homeostasis, and pathological compensation from a biologically plausible ion channel expression model. *Neuron*, 82(4):809–821, 2014.

[19] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar behavior and different conductance densities. *Elife*, 8:e42722, 2019.

[20] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241, 2018.

[21] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.

[22] Aman B Saleem, Aslı Ayaz, Kathryn J Jeffery, Kenneth D Harris, and Matteo Carandini. Integration of visual motion and locomotion in mouse visual cortex. *Nature neuroscience*, 16(12):1864–1869, 2013.

[23] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22(10):1677–1686, 2019.

[24] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[25] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[26] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[27] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[28] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946, 2012.

[29] Keegan E Hines, Thomas R Middendorf, and Richard W Aldrich. Determination of parameter identifiability in nonlinear biophysical models: A bayesian approach. *Journal of General Physiology*, 143(3):401–416, 2014.

[30] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.

[31] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.

[32] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.

[33] Wiktor Młynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.

[34] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

[35] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural dynamics. *bioRxiv*, page 838383, 2019.

[36] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 2015.

[37] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

[38] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow networks. *International Conference on Learning Representations*, 2017.

[39] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[40] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.

[41] Gabrielle J Gutierrez, Timothy O'Leary, and Eve Marder. Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. *Neuron*, 77(5):845–858, 2013.

[42] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.

[43] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

[44] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–1406, 2014.

[45] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.

[46] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*, 115(3):1399–1409, 2016.

[47] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the operating regime of cortex. *bioRxiv*, 2020.

[48] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor routing. *bioRxiv*, page 245613, 2018.

[49] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodulation. *Neural Networks*, 15(4-6):479–493, 2002.

[50] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634, 2009.

[51] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural networks. *PLoS computational biology*, 16(2):e1007655, 2020.

[52] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in neurobiology*, 25:156–163, 2014.

[53] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, 46:1–6, 2017.

[54] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966, 2018.

[55] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[56] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2020.

[57] Eve Marder and Allen I Selverston. *Dynamic biological networks: the stomatogastric nervous system*. MIT press, 1992.

[58] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–592, 2009.

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

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

[61] Mark M. Churchland, Byron M. Yu, John P. Cunningham, Leo P. Sugrue, Marlene R. Cohen, Greg S. Corrado, William T. Newsome, Andrew M. Clark, Paymon Hosseini, Benjamin B. Scott, David C. Bradley, Matthew A. Smith, Adam Kohn, J. Anthony Movshon, Katherine M. Armstrong, Tirin Moore, Steve W. Chang, Lawrence H. Snyder, Stephen G. Lisberger, Nicholas J. Priebe, Ian M. Finn, David Ferster, Stephen I. Ryu, Gopal Santhanam, Maneesh Sahani, and Krishna V. Shenoy. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.*, 13(3):369–378, 2010.

[62] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.

[63] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*, 5(10):793, 2004.

[64] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neurobiology*, 71(1):45–61, 2011.

[65] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.

[66] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature Neuroscience*, 16(8):1068, 2013.

[67] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[68] C Gardiner. Stochastic methods: A Handbook for the Natural and Social Sciences, 2009.

[69] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.

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

[71] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.

[72] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[73] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[74] Ben Calderhead and Mark Girolami. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface focus*, 1(6):821–835, 2011.

[75] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–820, 2011.

[76] Oksana A Chkrebtii, David A Campbell, Ben Calderhead, Mark A Girolami, et al. Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–1267, 2016.

[77] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.

[78] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degenerate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon, Portugal*, 2019.

[79] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller, Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.

[80] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In *Bernstein Conference 2018, Berlin, Germany*, 2018.

[81] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical inference for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin, Germany*, 2019.

[82] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.

[83] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[84] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–4248. PMLR, 2020.

[85] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[86] Sean R Bittner and John P Cunningham. Approximating exponential family models (not single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.

[87] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

[88] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.

[89] Maria Pia Saccomani, Stefania Audoly, and Leontina D'Angiò. Parameter identifiability of nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.

[90] Stefan Hengl, Clemens Kreutz, Jens Timmer, and Thomas Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618, 2007.

[91] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

[92] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751, 2016.

[93] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[94] Emmanuel Klinger, Dennis Rickert, and Jan Hasenauer. pyabc: distributed, likelihood-free inference. *Bioinformatics*, 34(20):3591–3593, 2018.

[95] David S Greenberg, Marcel Nonnenmacher, and Jakob H Macke. Automatic posterior transformation for likelihood-free inference. *International Conference on Machine Learning*, 2019.

[96] Daniel P Mossing, Julia Veit, Agostina Palmigiano, Kenneth D. Miller, and Hillel Adesnik. Antagonistic inhibitory subnetworks control cooperation and competition across cortical space. *bioRxiv*, 2021.

5 Methods

5.1 Emergent property inference (EPI)

Determining the combinations of model parameters that can produce a desired output is a key part of scientific practice. Solving inverse problems is especially important in neuroscience, since we require detailed circuit models to produce computation of varying levels of complexity. While much machine learning research has focused on how to find latent structure in large-scale neural datasets, less has focused on inverting theoretical circuit models conditioned upon the emergent properties of computation. Here, we introduce a novel method for statistical inference, which finds distributions of parameter solutions that are constrained to produce the desired emergent property. This method seamlessly handles neural circuit models with stochastic nonlinear dynamical generative processes, which are predominant in theoretical neuroscience.

Consider model parameterization $\mathbf{z}$, which is a collection of scientifically meaningful variables that govern the complex simulation of data $\mathbf{x}$. For example (see Section 3.1), $\mathbf{z}$ may be the electrical conductance parameters of an STG subcircuit, and $\mathbf{x}$ the evolving membrane potentials (the state) of the five neurons. In terms of statistical modeling, this circuit model has an intractable likelihood $p(\mathbf{x} \mid \mathbf{z})$, which is predicated by the stochastic differential equations that define the model. From a theoretical perspective, we are less concerned about the likelihood of an exemplary dataset $\mathbf{x}$, but rather the emergent property of intermediate hub frequency (which implies a consistent dataset $\mathbf{x}$).

In the STG example, the statistic $f(\mathbf{x}; \mathbf{z})$ measures hub neuron frequency from the evolution of $\mathbf{x}$ governed by parameters $\mathbf{z}$. With EPI, we learn distributions of $\mathbf{z}$ constrained to produce intermediate hub frequency: to obey the constraints placed on the mean and variance of $f(\mathbf{x}; \mathbf{z})$. In general, an emergent property $\mathcal{X}$ is defined through the choice of $f(\mathbf{x}; \mathbf{z})$ (which may be one or multiple statistics), and its means $\boldsymbol{\mu}$, and variances $\boldsymbol{\sigma}^2$:

$$\mathcal{X} : \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x}; \mathbf{z})\right] = \boldsymbol{\mu}, \mathrm{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x}; \mathbf{z})\right] = \boldsymbol{\sigma}^2. \tag{10}$$

Precisely, the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ must have means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ over the EPI distribution of parameters and the data produced by those parameters. Technically, an emergent property may be a combination of first-, second-, or higher-order moments, but this study focuses on the case written in Equation 10.

In EPI, deep probability distributions are optimized to learn the inferred distribution. In deep probability distributions, a simple random variable $\mathbf{z}_0 \sim q_0(\mathbf{z}_0)$ (we choose an isotropic gaussian)

is mapped deterministically via a sequence of deep neural network layers (g_1, .. g_l) parameterized by weights and biases $\boldsymbol{\theta}$ to the support of the distribution of interest:

$$\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0) = g_l(..g_1(\mathbf{z}_0)) \sim q_{\boldsymbol{\theta}}(\mathbf{z}). \tag{11}$$

Such deep probability distributions embed the inferred distribution in a deep network. Once optimized, this deep network representation has remarkably useful properties: fast sampling and probability evaluations Importantly, fast probability evaluations confer fast gradient and Hessian calculations as well.

Given this choice of circuit model and emergent property $\mathcal{X}$, $q_{\boldsymbol{\theta}}(\mathbf{z})$ is optimized via the neural network parameters $\boldsymbol{\theta}$ to find a maximally entropic distribution $q^*_{\boldsymbol{\theta}}$ within the deep variational family $\mathcal{Q}$ producing the emergent property $\mathcal{X}$:

$$\begin{aligned} q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}) = q^*_{\boldsymbol{\theta}}(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in Q}{\operatorname{argmax}}\, & H(q_{\boldsymbol{\theta}}(\mathbf{z})) \\ \text{s.t. } \mathcal{X} \quad : \quad & \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \boldsymbol{\sigma}^2. \end{aligned} \tag{12}$$

Entropy is chosen as the normative selection principle to match that of variational Bayesian methods (see Section 5.1.3). However, a key difference is that variational Bayesian methods do not constrain the predictions of their inferred parameter distribution. This optimization is executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [38].

In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method. First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Second, we describe the special class of deep probability distributions used in EPI called normalizing flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equation 12 (Section 5.1.4). Then, we demonstrate the details of this optimization in a toy example (Section 5.1.5). Finally, we establish the known relationship between maximum entropy distributions and exponential families (Section 5.1.3), which is used to explain how EPI is a form of variational inference (Section 5.1.6).

5.1.1 Related approaches

When bayesian inference problems lack conjugacy, scientists use approximate inference methods like variational inference (VI) [71] and Markov chain Monte Carlo (MCMC) [72,73]. After optimization, variational methods return a parameterized posterior distribution, which we can analyze. Also, the variational approximating distribution class is often chosen such that it permits fast sampling. In

contrast MCMC methods only produce samples from the approximated posterior distribution. No parameterized distribution is estimated, and additional samples are always generated with the same sampling complexity. Inference in models defined by systems of differential has been demonstrated with MCMC [70], although this approach requires tractable likelihoods. Advancements have introduced sampling [74], likelihood approximation [75], and uncertainty quantification techniques [76] to make MCMC approaches more efficient and expand the class of applicable models.

Simulation-based inference [56] is model parameter inference in the absence of a tractable likelihood function. The most prevalent approach to simulation-based inference is approximate Bayesian computation (ABC) [24], in which satisfactory parameter samples are kept from random prior sampling according to a rejection heuristic. The obtained set of parameters do not have a probabilities, and further insight about the model must be gained from examination of the parameter set and their generated activity. Methodological advances to ABC methods have come through the use of Markov chain Monte Carlo (MCMC-ABC) [25] and sequential Monte Carlo (SMC-ABC) [26] sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still struggles to scale in dimensionality [55] (cf. Fig. 2). Still, this method has enjoyed much success in systems biology [77]. Furthermore, once a parameter set has been obtained by SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again from scratch with a new population of initialized particles to obtain additional samples.

For scientific model analysis, we seek a parameter distribution represented by an approximating distribution as in variational inference [71]: a variational approximation that once optimized yields fast analytic calculations and samples. For the reasons described above, ABC and MCMC techniques are unattractive, since they only produce a set of parameter samples lacking probabilities and have unchanging sampling rate. EPI infers parameters in circuit models using the MEFN [38] algorithm with a deep variational approximation. The deep neural network of EPI (Fig. 1E) defines the parametric form (with variational parameters $\boldsymbol{\theta}$) of the deep variational approximation of circuit parameters $\mathbf{z}$.

Since EPI is not conditioning upon exemplary data as in variational bayesian methodology, EPI is not doing established variational inference. In contrast, the EPI distribution is constrained to produce an emergent property. EPI optimization is enabled using stochastic gradient techniques in the spirit of likelihood-free variational inference [34]. The analytic relationship between EPI and variational inference is explained in Section 5.1.6.

We note that, during our preparation and early presentation of this work [78, 79], another work

has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural circuits ([35, 80, 81]).We are encouraged by this general problem being recognized by others in the community, and we emphasize that these works offer complementary neuroscientific contributions (different theoretical models of focus) and use different technical methodologies (ours is built on our prior work [38], theirs similarly [82]).

The method EPI differs from SNPE in some key ways. SNPE belongs to a "sequential" class of recently developed simulation-based inference methods in which two neural networks are used for posterior inference. This first neural network is a deep probability distribution (normalizing flow) used to estimate the posterior $p(\mathbf{z} \mid \mathbf{x})$ (SNPE) or the likelihood $p(\mathbf{x} \mid \mathbf{z})$ (sequential neural likelihood (SNL [83])). A recent advance uses an unconstrained neural network to estimate the likelihood ratio (sequential neural ratio estimation (SNRE [84]). In SNL and SNRE, MCMC sampling techniques are used to obtain samples from the approximated posterior. This contrasts with EPI and SNPE, which use deep probability distributions to model parameters, which facilitates immediate measurements of sample probability, gradient, or Hessian for system analysis. The second neural network in this sequential class of methods is the amortizer. This unconstrained deep network maps data $\mathbf{x}$ (or statistics $f(\mathbf{x}; \mathbf{z})$ or model parameters $\mathbf{z}$ to the weights and biases of the first neural network. These methods are optimized on a conditional density (or ratio) estimation objective. The data used to optimize this objective are generated via an adaptive procedure, in which training data pairs ($\mathbf{x}_i$, $\mathbf{z}_i$) become sequentially closer to the true data and posterior.

The approximating fidelity of the deep probability distribution in sequential approaches is optimized to generalize across the training distribution of the conditioning variable. This generalization property of the sequential methods can reduce the accuracy at the singular posterior of interest. Whereas in EPI, the entire expressivity of the deep probability distribution is dedicated to learning a single distribution as well as possible. Amortization is not possible in EPI, since EPI learns an exponential family distribution parameterized by its mean (see Section 5.1.3). Since EPI distributions are defined by the mean $\boldsymbol{\mu}$ of their statistics, there is the well-known inverse mapping problem of exponential families [85] that prohibits an amortization based approach. However, we have shown that the same two-network architecture of the sequential simulation-based inference methods can be used for amortized inference in intractable exponential family posteriors using their natural parameterization [86].

Finally, one important differentiating factor between EPI and sequential simulation-based inference methods is that EPI leverages gradients $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ during optimization. These gradients can

improve convergence time and scalability, as we have shown on an example conditioning low-rank RNN connectivity on the property of stable amplification (see Section 3.3). With EPI, we prove out the suggestion that a deep inference technique can improve efficiency by leveraging these model gradients when they are tractable. Sequential simulation-based inference techniques may be better suited for scientific problems where $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ is intractable or unavailable, like when there is a non-differentiable model or it requires lengthy simulations. However, the sequential simulation-based inference techniques cannot constrain the predictions of the inferred distribution in the manner of EPI.

Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in scientific models. Around a single parameter choice, one can measure the Jacobian. One approach for this calculation that scales well is EAR [28]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [87] and its stochastic adaptation [88]. Casting identifiability as a statistical estimation problem, the profile likelihood works via iterated optimization while holding parameters fixed [27]. An exciting recent method is capable of recovering the functional form of such unidentifiabilities away from a point by following degenerate dimensions of the fisher information matrix [30]. Global structural non-identifiabilities can be found for models with polynomial or rational dynamics equations using DAISY [89], or through mean optimal transformations [90]. With EPI, we have all the benefits given by a statistical inference method plus the ability to query the first- or second-order gradient of the probability of the inferred distribution at any chosen parameter value. The second-order gradient of the log probability (the Hessian), which is directly afforded by EPI distributions, produces quantified information about parametric sensitivity of the emergent property in parameter space (see Section 3.2).

5.1.2 Deep probability distributions and normalizing flows

Deep probability distributions are comprised of multiple layers of fully connected neural networks (Equation 11). When each neural network layer is restricted to be a bijective function, the sample density can be calculated using the change of variables formula at each layer of the network. For $\mathbf{z}_i = g_i(\mathbf{z_{i-1}})$,

$$p(\mathbf{z}_i) = p(g_i^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial g_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} \right| = p(\mathbf{z}_{i-1}) \left| \det \frac{\partial g_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right|^{-1}. \tag{13}$$

However, this computation has cubic complexity in dimensionality for fully connected layers. By restricting our layers to normalizing flows [36, 37] – bijective functions with fast log determinant

Jacobian computations, which confer a fast calculation of the sample log probability. Fast log probability calculation confers efficient optimization of the maximum entropy objective (see Section 5.1.4).

We use the Real NVP [39] normalizing flow class, because its coupling architecture confers both fast sampling (forward) and fast log probability evaluation (backward). Fast probability evaluation facilitates fast gradient and Hessian evaluation of log probability throughout parameter space. Glow permutations were used in between coupling stages [40]. This is in contrast to autoregressive architectures [91, 92], in which only one of the forward or backward passes can be efficient. In this work, normalizing flows are used as flexible parameter distribution approximations $q_{\boldsymbol{\theta}}(\mathbf{z})$ having weights and biases $\boldsymbol{\theta}$. We specify the architecture used in each application by the number of Real-NVP affine coupling stages, and the number of neural network layers and units per layer of the conditioning functions.

When calculating Hessians of log probabilities in deep probability distributions, it is important to consider the normalizing flow architecture. With autoregressive architectures [91, 92], fast sampling and fast log probability evaluations are mutually exclusive. That makes these architectures undesirable for EPI, where efficient sampling is important for optimization, and log probability evaluation speed predicates the efficiency of gradient and Hessian calculations. With Real NVP coupling architectures, we get both fast sampling and fast Hessians making both optimization and scientific analysis efficient.

5.1.3 Maximum entropy distributions and exponential families

EPI is a maximum entropy distribution, which have fundamental links to exponential family distributions. A maximum entropy distribution of form:

$$\begin{aligned} p^*(\mathbf{z}) = &\underset{p \in \mathcal{P}}{\operatorname{argmax}}\, H(p(\mathbf{z})) \\ &\text{s.t. } \mathbb{E}_{\mathbf{z}\sim p}\left[T(\mathbf{z})\right] = \boldsymbol{\mu}_{\text{opt}}. \end{aligned} \tag{14}$$

will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \tag{15}$$

The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are formally hard to identify except in special cases [85].

In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equations 2 and 3. The variance constraint is a second moment constraint on $f(\mathbf{x};\mathbf{z})$

$$\mathrm{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[\left(f(\mathbf{x};\mathbf{z}) - \boldsymbol{\mu}\right)^2\right] \tag{16}$$

As a general maximum entropy distribution (Equation 14), the sufficient statistics vector contains both first and second order moments of $f(\mathbf{x};\mathbf{z})$

$$T(\mathbf{x};\mathbf{z}) = \begin{bmatrix} f(\mathbf{x};\mathbf{z}) \\ \left(f(\mathbf{x};\mathbf{z}) - \boldsymbol{\mu}\right)^2 \end{bmatrix}, \tag{17}$$

which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\mathrm{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma}^2 \end{bmatrix}. \tag{18}$$

5.1.4 Augmented Lagrangian optimization

To optimize $q_{\boldsymbol{\theta}}(\mathbf{z})$ in Equation 12, the constrained maximum entropy optimization is executed using the augmented Lagrangian method. The following objective is minimized:

$$L(\boldsymbol{\theta};\boldsymbol{\eta}_{\mathrm{opt}},c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\mathrm{opt}}^{\top} R(\boldsymbol{\theta}) + \frac{c}{2}||R(\boldsymbol{\theta})||^2 \tag{19}$$

where average constraint violations $R(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}\left[\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\mathbf{z})}\left[T(\mathbf{x};\mathbf{z}) - \boldsymbol{\mu}_{\mathrm{opt}}\right]\right]$, $\boldsymbol{\eta}_{\mathrm{opt}} \in \mathbb{R}^m$ are the Lagrange multipliers where $m = |\boldsymbol{\mu}_{\mathrm{opt}}| = |T(\mathbf{x};\mathbf{z})| = 2|f(\mathbf{x};\mathbf{z})|$, and c is the penalty coefficient. The sufficient statistics $T(\mathbf{x};\mathbf{z})$ and mean parameter $\boldsymbol{\mu}_{\mathrm{opt}}$ are determined by the means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ of emergent property statistics $f(\mathbf{x};\mathbf{z})$ defined in Equation 12 (see Section 5.1.6). Specifically, $T(\mathbf{x};\mathbf{z})$ is a concatenation of the first and second moments, $\boldsymbol{\mu}_{\mathrm{opt}}$ is a concatenation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ (see section 5.1.3), and the Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of exponential families (see Section 5.1.3). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution are optimized according to Equation 19 using the Adam optimizer with learning rate 10^{-3} [93].

The gradient with respect to entropy $H(q_{\boldsymbol{\theta}}(\mathbf{z}))$ can be expressed using the reparameterization trick as an expectation of the negative log density of parameter samples $\mathbf{z}$ over the randomness in the parameterless initial distribution $q_0(\mathbf{z}_0)$:

$$H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \int -q_{\boldsymbol{\theta}}(\mathbf{z})\log(q_{\boldsymbol{\theta}}(\mathbf{z}))d\mathbf{z} = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}\left[-\log(q_{\boldsymbol{\theta}}(\mathbf{z}))\right] = \mathbb{E}_{\mathbf{z}_0\sim q_0}\left[-\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))\right]. \tag{20}$$

Thus, the gradient of the entropy of the deep probability distribution can be estimated as an average with respect to the base distribution $\mathbf{z}_0$:

$$\nabla_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0\sim q_0}\left[-\nabla_{\boldsymbol{\theta}}\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))\right]. \tag{21}$$

The lagrangian parameters $\boldsymbol{\eta}_{\text{opt}}$ are initialized to zero and adapted following each augmented Lagrangian epoch, which is a period of optimization with fixed ($\boldsymbol{\eta}_{\text{opt}}$, c) for a given number of stochastic optimization iterations. A low value of c is used initially, and conditionally increased after each epoch based on constraint error reduction. The penalty coefficient is updated based on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of $\mathbb{E}[||R(\boldsymbol{\theta}_{k+1})||] > \gamma\mathbb{E}\left[||R(\boldsymbol{\theta}_k)||\right]$ is computed, and c_{k+1} is updated to βc_k with probability $1-p$. The other update rule is $\boldsymbol{\eta}_{\text{opt},k+1} = \boldsymbol{\eta}_{\text{opt},k} + c_k \frac{1}{n}\sum_{i=1}^{n}(T(\mathbf{x}^{(i)}) - \boldsymbol{\mu}_{\text{opt}})$ given a batch size n. Throughout the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied according to application.

The intention is that c and $\boldsymbol{\eta}_{\text{opt}}$ start at values encouraging entropic growth early in optimization. With each training epoch in which the update rule for c is invoked by unsatisfactory constraint error reduction, the constraint satisfaction terms are increasingly weighted, resulting in a decreased entropy. This encourages the discovery of suitable regions of parameter space, and the subsequent refinement of the distribution to produce the emergent property (see example in Section 5.1.5). The momentum parameters of the Adam optimizer are reset at the end of each augmented Lagrangian epoch.

Rather than starting optimization from some $\boldsymbol{\theta}$ drawn from a randomized distribution, we found that initializing $q_{\boldsymbol{\theta}}(\mathbf{z})$ to approximate an isotropic Gaussian distribution conferred more stable, consistent optimization. The parameters of the Gaussian initialization were chosen on an application-specific basis. Throughout the study, we chose isotropic Gaussian initializations with mean $\boldsymbol{\mu}_{\text{init}}$ at the center of the distribution support and some standard deviation $\boldsymbol{\sigma}_{\text{init}}$, except for one case, where an initialization informed by random search was used (see Section 5.2).

To assess whether the EPI distribution $q_{\boldsymbol{\theta}}(\mathbf{z})$ produces the emergent property, we assess whether each individual constraint on the means and variances of $f(\mathbf{x};\mathbf{z})$ is satisfied. We consider the EPI to have converged when a null hypothesis test of constraint violations $R(\boldsymbol{\theta})_i$ being zero is accepted for all constraints $i \in \{1, ..., m\}$ at a significance threshold $\alpha = 0.05$. This significance threshold is adjusted through Bonferroni correction according to the number of constraints m. The p-values for each constraint are calculated according to a two-tailed nonparametric test, where 200 estimations of the sample mean $R(\boldsymbol{\theta})^i$ are made using N_{test} samples of $\mathbf{z} \sim q_{\boldsymbol{\theta}}(\mathbf{z})$ at the end of the augmented Lagrangian epoch.

When assessing the suitability of EPI for a particular modeling question, there are some important technical considerations. First and foremost, as in any optimization problem, the defined emergent

property should always be appropriately conditioned (constraints should not have wildly different units). Furthermore, if the program is underconstrained (not enough constraints), the distribution grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no parameter set producing the emergent property, and EPI optimization will fail (appropriately). Next, one should consider the computational cost of the gradient calculations. In the best circumstance, there is a simple, closed form expression (e.g. Section 5.3) for the emergent property statistic given the model parameters. On the other end of the spectrum, many forward simulation iterations may be required before a high quality measurement of the emergent property statistic is available (e.g. Section 5.2). In such cases, backpropagating gradients through the SDE evolution will be expensive.

5.1.5 Example: 2D LDS

To gain intuition for EPI, consider a two-dimensional linear dynamical system (2D LDS) model (Fig. S1A):

$$\tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \tag{22}$$

with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \tag{23}$$

To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fixing $\tau = 1$s), the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ were chosen to contain the oscillatory frequency, $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscillating system. λ_1 is the eigenvalue of greatest real part when the imaginary component is zero, and alternatively of positive imaginary component when the eigenvalues are complex conjugate pairs. To learn the distribution of real entries of A that produce a band of oscillating systems around 1Hz, we formalized this emergent property as $\text{real}(\lambda_1)$ having mean zero with variance 0.25^2, and the oscillation frequency $2\pi\text{imag}(\lambda_1)$ having mean 1Hz with variance $(0.1\text{Hz})^2$:

$$\mathbb{E}\left[T(\mathbf{x})\right]_{\mathbf{z},\mathbf{x}} \triangleq \mathbb{E}\begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x};\mathbf{z}) \\ \text{imag}(\lambda_1)(\mathbf{x};\mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x};\mathbf{z}) - 0)^2 \\ (\text{imag}(\lambda_1)(\mathbf{x};\mathbf{z}) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}_{\text{opt}}. \tag{24}$$

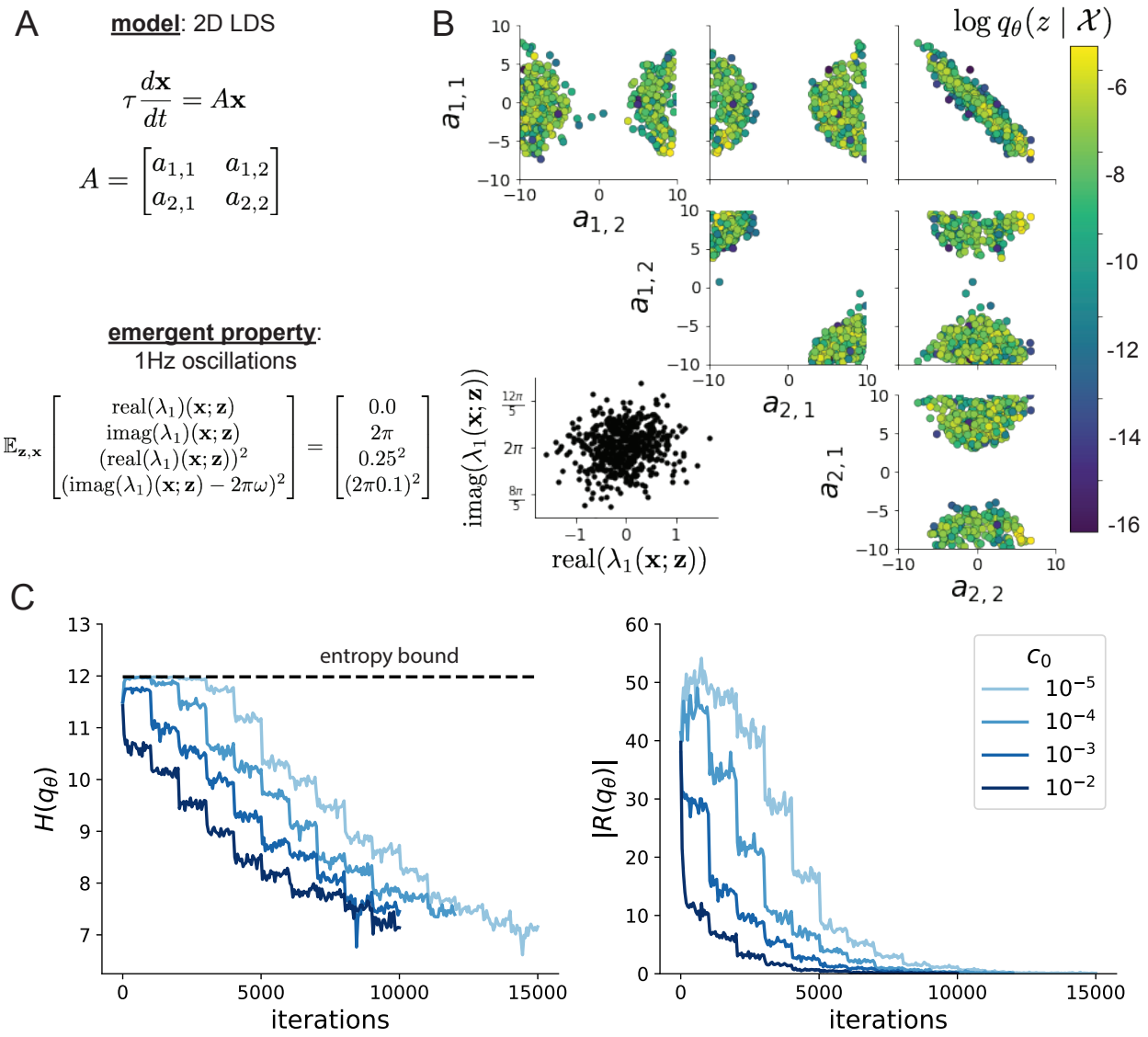

Figure 6: (LDS1): **A**. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. **B**. The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed lines indicate the parameter axes. **C**. Entropy throughout the optimization. At the beginning of each augmented Lagrangian epoch (2,000 iterations), the entropy dipped due to the shifted optimization manifold where emergent property constraint satisfaction is increasingly weighted. **D**. Emergent property moments throughout optimization. At the beginning of each augmented Lagrangian epoch, the emergent property moments adjust closer to their constraints.

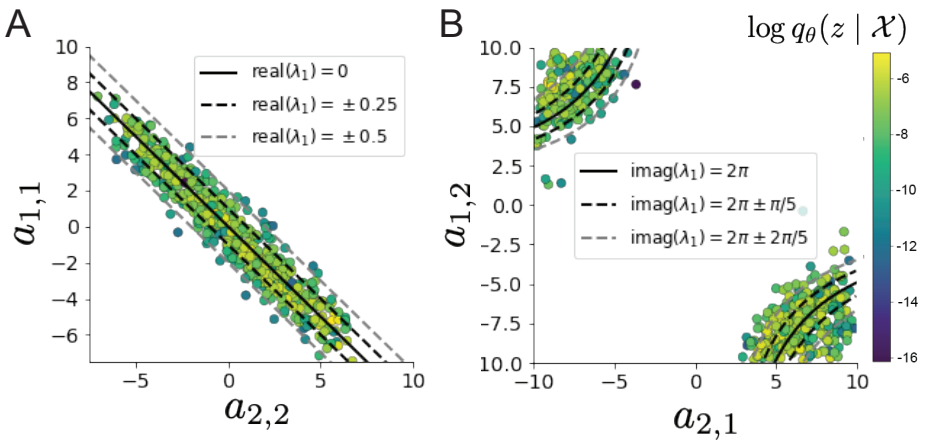

Figure 7: (LDS2): **A**. Probability contours in the a_1-a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor real(λ_1). **B**. Probability contours in the a_2-a_3 plane were derived from the emergent property statistic of oscillation frequency 2πimag(λ_1).

Unlike the models we presented in the main text, this model admits an analytical form for the mean emergent property statistics given parameter $\mathbf{z}$, since the eigenvalues can be calculated using the quadratic formula:

$$\lambda = \frac{(\frac{a_1+a_4}{\tau}) \pm \sqrt{(\frac{a_1+a_4}{\tau})^2 + 4(\frac{a_2 a_3 - a_1 a_4}{\tau})}}{2}. \tag{25}$$

Importantly, even though $\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\mathbf{z})}[T(\mathbf{x})]$ is calculable directly via a closed form function and does not require simulation, we cannot derive the distribution $q^*_{\boldsymbol{\theta}}$ directly. This fact is due to the formally hard problem of the backward mapping: finding the natural parameters η from the mean parameters $\boldsymbol{\mu}$ of an exponential family distribution [85]. Instead, we used EPI to approximate this distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

Even this relatively simple system has nontrivial (though intuitively sensible) structure in the parameter distribution. To validate our method, we analytically derived the contours of the probability density from the emergent property statistics and values. In the a_1-a_4 plane, the black line at real(λ_1) = $\frac{a_1+a_4}{2} = 0$, dotted black line at the standard deviation real(λ_1) = $\frac{a_1+a_4}{2} \pm 0.25$, and the dotted gray line at twice the standard deviation real(λ_1) = $\frac{a_1+a_4}{2} \pm 0.5$ follow the contour of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired statistical constraints and model degeneracy in the sum of a_1 and a_4. Intuitively, the parameters equivalent with respect to emergent property statistic real(λ_1) have similar log densities.

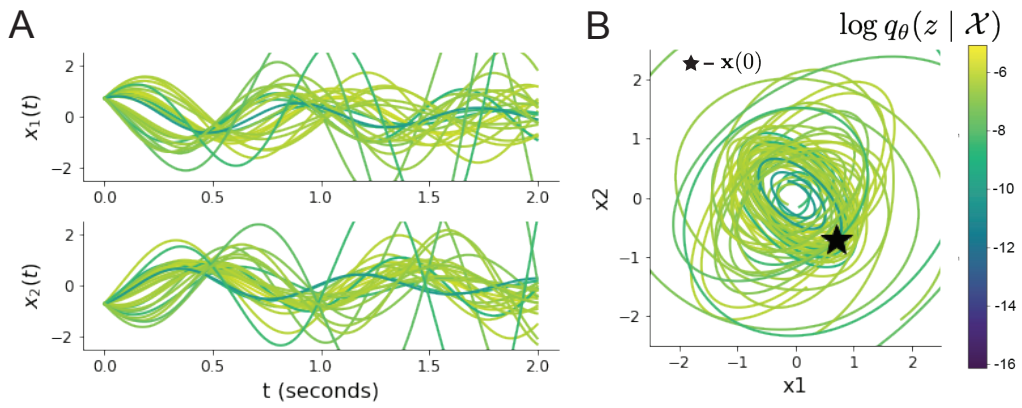

Figure 8: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\boldsymbol{\theta}}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. **A**. Each dimension of the simulated trajectories throughout time. **B**. The simulated trajectories in phase space.

To explain the bimodality of the EPI distribution, we examined the imaginary component of λ_1. When $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, we have

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1 a_4 - a_2 a_3}{\tau}}, & \text{if } a_1 a_4 < a_2 a_3 \\ 0 & \text{otherwise} \end{cases}. \tag{26}$$

When $\tau = 1$ and $a_1 a_4 > a_2 a_3$ (center of distribution above), we have the following equation for the other two dimensions:

$$\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 \tag{27}$$

Since we constrained $\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}[\text{imag}(\lambda)] = 2\pi$, we can plot contours of the equation $\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 = (2\pi)^2$ for various $a_1 a_4$ (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}(|a_1 a_4 - E_{q_{\boldsymbol{\theta}}}[a_1 a_4]|)$, we show the contours as $a_1 a_4 = 0$ (black), $a_1 a_4 = -\sigma_{1,4}$ (black dotted), and $a_1 a_4 = -2\sigma_{1,4}$ (grey dotted). This validates the curved structure of the inferred distribution learned through EPI. We took steps in negative standard deviation of $a_1 a_4$ (dotted and gray lines), since there are few positive values $a_1 a_4$ in the learned distribution. Subtler combinations of model and emergent property will have more complexity, further motivating the use of EPI for understanding these systems. As we expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz (Fig. S3).

5.1.6 EPI as variational inference

In Bayesian inference a prior belief about model parameters $\mathbf{z}$ is stated in a prior distribution $p(\mathbf{z})$, and the statistical model capturing the effect of $\mathbf{z}$ on observed data points $\mathbf{x}$ is formalized in the likelihood distribution $p(\mathbf{x} \mid \mathbf{z})$. In Bayesian inference, we obtain a posterior distribution $p(\mathbf{z} \mid \mathbf{x})$, which captures how the data inform our knowledge of model parameters using Bayes' rule:

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \tag{28}$$

The posterior distribution is analytically available when the prior is conjugate with the likelihood. However, conjugacy is rare in practice, and alternative methods, such as variational inference [71], are utilized.

In variational inference, a posterior approximation $q^*_{\boldsymbol{\theta}}$ is chosen from within some variational family $\mathcal{Q}$

$$q^*_{\boldsymbol{\theta}}(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in Q}{\operatorname{argmin}}\, KL(q_{\boldsymbol{\theta}}(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{x})). \tag{29}$$

The KL divergence can be written in terms of entropy of the variational approximation:

$$KL(q_{\boldsymbol{\theta}}(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{x})) = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}\left[\log(q_{\boldsymbol{\theta}}(\mathbf{z}))\right] - \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}\left[\log(p(\mathbf{z} \mid \mathbf{x}))\right] \tag{30}$$

$$= -H(q_{\boldsymbol{\theta}}) - \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}\left[\log(p(\mathbf{x} \mid \mathbf{z})) + \log(p(\mathbf{z})) - \log(p(\mathbf{x}))\right] \tag{31}$$

Since the marginal distribution of the data $p(\mathbf{x})$ (or "evidence") is independent of $\boldsymbol{\theta}$, variational inference is executed by optimizing the remaining expression. This is usually framed as maximizing the evidence lower bound (ELBO)

$$\underset{q_{\boldsymbol{\theta}} \in Q}{\operatorname{argmin}}\, KL(q_{\boldsymbol{\theta}} \mid\mid p(\mathbf{z} \mid \mathbf{x})) = \underset{q_{\boldsymbol{\theta}} \in Q}{\operatorname{argmax}}\, H(q_{\boldsymbol{\theta}}) + \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}\left[\log(p(\mathbf{x} \mid \mathbf{z})) + \log(p(\mathbf{z}))\right]. \tag{32}$$

Now, consider the setting where we have chosen a uniform prior, and stipulate a mean-field gaussian likelihood on a chosen statistic of the data $f(\mathbf{x}; \mathbf{z})$

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) \mid \boldsymbol{\mu}_f, \Sigma_f), \tag{33}$$

where $\Sigma_f = \operatorname{diag}(\boldsymbol{\sigma}_f^2)$. The log likelihood is then proportional to a dot product of the natural parameter of this mean-field gaussian distribution and the first and second moment statistics.

$$\log p(\mathbf{x} \mid \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \tag{34}$$

where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\boldsymbol{\mu}_f}{\boldsymbol{\sigma}_f^2} \\ \frac{-1}{2\boldsymbol{\sigma}_f^2} \end{bmatrix}, \text{ and} \tag{35}$$

$$T(\mathbf{x};\mathbf{z}) = \begin{bmatrix} f(\mathbf{x};\mathbf{z}) \\ (f(\mathbf{x};\mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \tag{36}$$

The variational objective is then

$$\underset{q_{\boldsymbol{\theta}} \in Q}{\operatorname{argmax}}\, H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_f^\top \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}\left[T(\mathbf{x};\mathbf{z})\right] \tag{37}$$

Comparing this to the Lagrangian objective (without augmentation) of EPI, we see they are the same

$$\begin{aligned} q^*_{\boldsymbol{\theta}}(\mathbf{z}) &= \underset{q_{\boldsymbol{\theta}} \in Q}{\operatorname{argmin}} -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top \left(\mathbb{E}_{\mathbf{z},\mathbf{x}}\left[T(\mathbf{x};\mathbf{z})\right] - \boldsymbol{\mu}_{\text{opt}}\right) \\ &= \underset{q_{\boldsymbol{\theta}} \in Q}{\operatorname{argmin}} -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[T(\mathbf{x};\mathbf{z})\right]. \end{aligned} \tag{38}$$

where $T(\mathbf{x};\mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x};\mathbf{z})$ (Equation 17). Thus, EPI is implicitly executing variational inference with a uniform prior and a mean-field gaussian likelihood on the emergent property statistics. The mean and variances of the mean-field gaussian likelihood are predicated by $\boldsymbol{\eta}_{\text{opt}}$ (Equations 35 and 37), which is adapted after each EPI optimization epoch based on $\mathcal{X}$ (see Section 5.1.4). In EPI, the inferred distribution is not conditioned on a finite dataset as in variational inference, but rather the emergent property $\mathcal{X}$ dictates the likelihood parameterization such that the inferred distribution will produce the emergent property. As a note, we could not simply choose $\boldsymbol{\mu}_f$ and $\boldsymbol{\sigma}_f$ directly from the outset, since we do not know which of these choices will produce the emergent property $\mathcal{X}$, which necessitates the EPI optimization routine that adapts $\boldsymbol{\eta}_{\text{opt}}$. Accordingly, we replace the notation of $p(\mathbf{z} \mid \mathbf{x})$ with $p(\mathbf{z} \mid \mathcal{X})$ conceptualizing an inferred distribution that obeys emergent property $\mathcal{X}$ (see Section 5.1).

5.2 Stomatogastric ganglion

In Section 3.1 and 3.2, we used EPI to infer conductance parameters in a model of the stomatogastric ganglion (STG) [41]. This 5-neuron circuit model represents two subcircuits: that generating the pyloric rhythm (fast population) and that generating the gastric mill rhythm (slow population). The additional neuron (the IC neuron of the STG) receives inhibitory synaptic input from both subcircuits, and can couple to either rhythm dependent on modulatory conditions. There is also a parametric regime in which this neuron fires at an intermediate frequency between that of the fast and slow populations [41], which we infer with EPI as a motivational example. This model

is not to be confused with an STG subcircuit model of the pyloric rhythm [57], which has been statistically inferred in other studies [15, 35].

5.2.1 STG model

We analyze how the parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$ govern the emergent phenomena of intermediate hub frequency in a model of the stomatogastric ganglion (STG) [41] shown in Figure 1A with activity $\mathbf{x} = [x_{\text{f1}}, x_{\text{f2}}, x_{\text{hub}}, x_{\text{s1}}, x_{\text{s2}}]$, using the same hyperparameter choices as Gutierrez et al. Each neuron's membrane potential $x_\alpha(t)$ for $\alpha \in \{\text{f1}, \text{f2}, \text{hub}, \text{s1}, \text{s2}\}$ is the solution of the following stochastic differential equation:

$$C_m \frac{dx_\alpha}{dt} = -\left[h_{leak}(\mathbf{x};\mathbf{z}) + h_{Ca}(\mathbf{x};\mathbf{z}) + h_K(\mathbf{x};\mathbf{z}) + h_{hyp}(\mathbf{x};\mathbf{z}) + h_{elec}(\mathbf{x};\mathbf{z}) + h_{syn}(\mathbf{x};\mathbf{z})\right] + dB. \quad (39)$$

The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization, electrical and synaptic currents. Each current component is a function of all membrane potentials and the conductance parameters $\mathbf{z}$. Finally, we include gaussian noise dB to the model of Gutierrez et al. so that the model stochastic, although this is not required by EPI.

The capacitance of the cell membrane was set to $C_m = 1nF$. Specifically, the currents are the difference in the neuron's membrane potential and that current type's reversal potential multiplied by a conductance:

$$h_{leak}(\mathbf{x};\mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (40)$$

$$h_{elec}(\mathbf{x};\mathbf{z}) = g_{\text{el}}(x_\alpha^{post} - x_\alpha^{pre}) \quad (41)$$

$$h_{syn}(\mathbf{x};\mathbf{z}) = g_{syn} S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (42)$$

$$h_{Ca}(\mathbf{x};\mathbf{z}) = g_{Ca} M_\infty(x_\alpha - V_{Ca}) \quad (43)$$

$$h_K(\mathbf{x};\mathbf{z}) = g_K N(x_\alpha - V_K) \quad (44)$$

$$h_{hyp}(\mathbf{x};\mathbf{z}) = g_h H(x_\alpha - V_{hyp}). \quad (45)$$

The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$, and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca}, g_K, and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast conductances had values $g_{Ca} = 1.9\times10^{-2}$, $g_K = 3.9\times10^{-2}$, and $g_{hyp} = 2.5\times10^{-2}$. The intermediate conductances had values $g_{Ca} = 1.7 \times 10^{-2}$, $g_K = 1.9 \times 10^{-2}$, and $g_{hyp} = 8.0 \times 10^{-3}$. Finally, the slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating dynamics dependent on steady-state gating variables M_∞, N_∞ and H_∞, respectively:

$$M_\infty = 0.5\left(1 + \tanh\left(\frac{x_\alpha - v_1}{v_2}\right)\right) \tag{46}$$

$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \tag{47}$$

$$N_\infty = 0.5\left(1 + \tanh\left(\frac{x_\alpha - v_3}{v_4}\right)\right) \tag{48}$$

$$\lambda_N = \phi_N \cosh\left(\frac{x_\alpha - v_3}{2v_4}\right) \tag{49}$$

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \tag{50}$$

$$H_\infty = \frac{1}{1 + \exp\left(\frac{x_\alpha + v_5}{v_6}\right)} \tag{51}$$

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp\left(\frac{-x_\alpha + v_7}{v_8}\right)}\right). \tag{52}$$

where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$, $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

Finally, there is a synaptic gating variable as well:

$$S_\infty = \frac{1}{1 + \exp\left(\frac{v_{th} - x_\alpha}{v_9}\right)}. \tag{53}$$

When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear dynamical system. The gaussian noise $d\mathbf{B}$ has variance $(1 \times 10^{-12})^2$ A^2, and introduces variability in frequency at each parameterization $\mathbf{z}$.

5.2.2 Hub frequency calculation

In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for $T = 300$ time steps of $dt = 25$ms. The chosen dt and T were the most computationally convenient choices yielding accurate frequency measurement. We used a basis of complex exponentials with frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

$$\Phi = [0.0, 0.01, ..., 1.0]^\top ... \tag{54}$$

To measure spiking frequency, we processed simulated membrane potentials with a relu (spike extraction) and low-pass filter with averaging window of size 20, then took the frequency with the

maximum absolute value of the complex exponential basis coefficients of the processed time-series. The first 20 temporal samples of the simulation are ignored to account for initial transients.

To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in \mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in \{\mathrm{f1}, \mathrm{f2}, \mathrm{hub}, \mathrm{s1}, \mathrm{s2}\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

$$\psi_\alpha = \mathrm{softmax}(\beta |X_\alpha| \odot i), \tag{55}$$

where $i = [0, 1, ..., 100]$. The frequency is then calculated as

$$\omega_\alpha = 0.01 \psi_\alpha \mathrm{Hz}. \tag{56}$$

Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean and variance of the emergent property statistics. In this case, we have one statistic, hub neuron frequency, where the mean was chosen to be 0.55Hz,(Equation 2) and variance was chosen to be $(0.025\mathrm{Hz})^2$ (Equation 3).

5.2.3 EPI details for the STG model

As a maximum entropy distribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments of the hub neuron frequency (as in Equations 17 and 18)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\mathrm{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\mathrm{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \tag{57}$$

$$\boldsymbol{\mu}_{\mathrm{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \tag{58}$$

Throughout optimization, the augmented Lagrangian parameters η and c, were updated after each epoch of 5,000 iterations(see Section 5.1.4). The optimization converged after five epochs (Fig. S4).

For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-layer neural networks of 25 units per layer. The normalizing flow architecture mapped $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$ to a support of $\mathbf{z} = [g_{\mathrm{el}}, g_{\mathrm{synA}}] \in [4, 8] \times [0.01, 4]$, initialized to a gaussian approximation of samples returned by a preliminary ABC search. We did not include $g_{\mathrm{synA}} < 0.01$, for numerical stability. EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The architecture converged with criteria $N_{\mathrm{test}} = 100$.

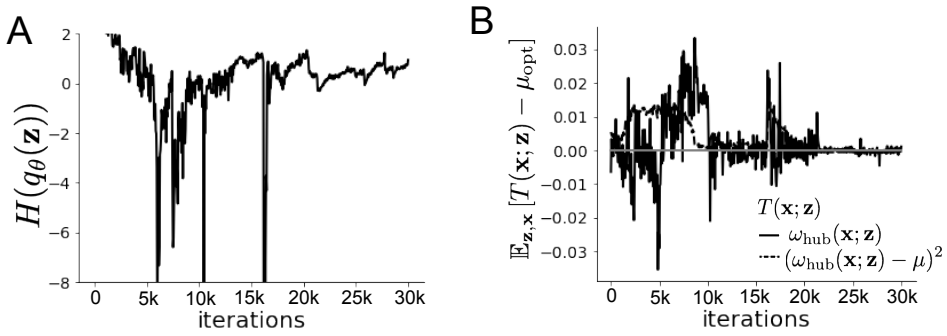

Figure 9: (STG1): EPI optimization of the STG model producing network syncing. **A**. Entropy throughout optimization. **B**. The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

5.2.4 Hessian sensitivity vectors

To quantify the second-order structure of the EPI distribution, we evaluated the Hessian of the log probability $\frac{\partial^2 \log q(\mathbf{z}|\mathcal{X})}{\partial \mathbf{z}\mathbf{z}^\top}$. The eigenvector of this Hessian with most negative eigenvalue is defined as the sensitivity dimension $\mathbf{v}_1$, and all subsequent eigenvectors are ordered by increasing eigenvalue. These eigenvalues are quantifications of how fast the emergent property deteriorates via the parameter combination of their associated eigenvector. In Figure 1D, the sensitivity dimension v_1 (solid) and the second eigenvector of the Hessian v_2 (dashed) are shown evaluated at the mode of the distribution. Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D parameter space were chosen to have positive g_{synA}. The length of the arrows is inversely proportional to the square root of the absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$. For the same magnitude perturbation away from the mode, intermediate hub frequency only diminishes along the sensitivity dimension $\mathbf{v}_1$ (Fig. 1E-F).

5.3 Scaling EPI for stable amplification in RNNs

5.3.1 Rank-2 RNN model

We examined the scaling properties of EPI by learning connectivities of RNNs of increasing size that exhibit stable amplification. Rank-2 RNN connectivity was modeled as $W = UV^\top$, where $U = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} + g\chi^{(W)}$, $V = \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix} + g\chi^{(V)}$, and $\chi^{(W)}_{i,j}, \chi^{(V)}_{i,j} \sim \mathcal{N}(0,1)$. This RNN model has

dynamics

$$\tau\dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}. \tag{59}$$

In this analysis, we inferred connectivity parameterizations $\mathbf{z} = \left[\mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{V}_1^\top, \mathbf{V}_2^\top\right]^\top \in [-1, 1]^{(4N)}$ that produced stable amplification using EPI, SMC-ABC [26], and SNPE [35] (see Section Related Methods).

5.3.2 Stable amplification

For this RNN model to be stable, all real eigenvalues of W must be less than 1: $\text{real}(\lambda_1) < 1$, where λ_1 denotes the greatest real eigenvalue of W. For a stable RNN to amplify at least one input pattern, the symmetric connectivity $W^s = \frac{W+W^\top}{2}$ must have an eigenvalue greater than 1: $\lambda_1^s > 1$, where λ^s is the maximum eigenvalue of W^s. These two conditions are necessary and sufficient for stable amplification in RNNs [51].

5.3.3 EPI details for RNNs

We defined the emergent property of stable amplification with means of these eigenvalues (0.5 and 1.5, respectively) that satisfy these conditions. To complete the emergent property definition, we chose variances (0.25^2) about those means such that samples rarely violate the eigenvalue constraints. In terms of the EPI optimization variables, this is written as

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \lambda_1^s(\mathbf{x}; \mathbf{z}) \\ \left(\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0.5\right)^2 \\ \left(\lambda_1^s(\mathbf{x}; \mathbf{z}) - 1.5\right)^2 \end{bmatrix}, \tag{60}$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.25^2 \\ 0.25^2 \end{bmatrix}. \tag{61}$$

Gradients of maximum eigenvalues of Hermitian matrices like W^s are available with modern automatic differentiation tools. To differentiate through the $\text{real}(\lambda_1)$, we solved the following equation for eigenvalues of rank-2 matrices using the rank reduced matrix $W^r = V^\top U$

$$\lambda_\pm = \frac{\text{Tr}(W^r) \pm \sqrt{\text{Tr}(W^r)^2 - 4\text{Det}(W^r)}}{2}. \tag{62}$$

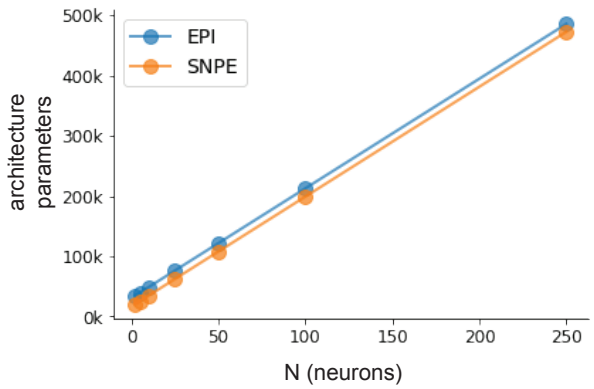

Figure 10: (RNN1): Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size (N).

For EPI in Fig. 2, we used a real NVP architecture with three coupling layers of affine transformations parameterized by two-layer neural networks of 100 units per layer. The initial distribution was a standard isotropic gaussian $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to the support of $\mathbf{z}_i \in [-1, 1]$. We used an augmented Lagrangian coefficient of $c_0 = 10^3$, a batch size $n = 200$, $\beta = 4$, and chose to use 500 iterations per augmented Lagrangian epoch and emergent property constraint convergence was evaluated at $N_{\text{test}} = 200$ (Fig. 2B blue line, and Fig. 2C-D blue).

5.3.4 Methodological comparison

We compared EPI to two alternative simulation-based inference techniques, since the likelihood of these eigenvalues given $\mathbf{z}$ is not available. Approximate Bayesian computation (ABC) [24] is a rejection sampling technique for obtaining sets of parameters $\mathbf{z}$ that produce activity $\mathbf{x}$ close to some observed data $\mathbf{x}_0$. Sequential Monte Carlo approximate Bayesian computation (SMC-ABC) is the state-of-the-art ABC method, which leverages SMC techniques to improve sampling speed. We ran SMC-ABC with the pyABC package [94] to infer RNNs with stable amplification: connectivities having eigenvalues within an ϵ-defined l-2 distance of

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}. \tag{63}$$

SMC-ABC was run with a uniform prior over $\mathbf{z} \in [-1, 1]^{(4N)}$, a population size of 1,000 particles with simulations parallelized over 32 cores, and a multivariate normal transition model.

SNPE, the next approach in our comparison, is far more similar to EPI. Like EPI, SNPE treats pa-

rameters in mechanistic models with deep probability distributions, yet the two learning algorithms are categorically different. SNPE uses a two-network architecture to approximate the posterior distribution of the model conditioned on observed data $\mathbf{x}_0$. The amortizing network maps observations $\mathbf{x}_i$ to the parameters of the deep probability distribution. The weights and biases of the parameter network are optimized by sequentially augmenting the training data with additional pairs ($\mathbf{z}_i$, $\mathbf{x}_i$) based on the most recent posterior approximation. This sequential procedure is important to get training data $\mathbf{z}_i$ to be closer to the true posterior, and $\mathbf{x}_i$ to be closer to the observed data. For the deep probability distribution architecture, we chose a masked autoregressive flow with affine couplings (the default choice), three transforms, 50 hidden units, and a normalizing flow mapping to the support as in EPI. This architectural choice closely tracked the size of the architecture used by EPI (Fig. 10). As in SMC-ABC, we ran SNPE with $\mathbf{x}_0 = \mu$. All SNPE optimizations were run for a limit of 1.5 days on a Tesla V100 GPU, or until two consecutive rounds resulted in a validation log probability lower than the maximum observed for that random seed.

To compare the efficiency of these algorithms for inferring RNN connectivity distributions producing stable amplification, we develop a convergence criteria that can be used across methods. While EPI has its own hypothesis testing convergence criteria for the emergent property, it would not make sense to use this criteria on SNPE and SMC-ABC which do not constrain the means and variances of their predictions. Instead, we consider EPI and SNPE to have converged after completing its most recent optimization epoch (EPI) or round (SNPE) in which the distance

$$d(q_\theta(z)) = |\mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] - \boldsymbol{\mu}|_2 \tag{64}$$

is less than 0.5. We consider SMC-ABC to have converged once the population produces samples within the $\epsilon = 0.5$ ball ensuring stable amplification.

When assessing the scalability of SNPE, it is important to check that alternative hyperparamterizations could not yield better performance. Key hyperparameters of the SNPE optimization are the number of simulations per round n_{round}, the number of atoms used in the atomic proposals of the SNPE-C algorithm [95], and the batch size n. To match EPI, we used a batch size of $n = 200$ for $N <= 25$, however we found $n = 1,000$ to be helpful for SNPE in higher dimensions. While $n_{\text{round}} = 1,000$ yielded SNPE convergence for $N <= 25$, we found that a substantial increase to $n_{\text{round}} = 25,000$ yielded more consistent convergence at $N = 50$ (Fig. 11A). By increasing n_{round}, we also necessarily increase the duration of each round. At $N = 100$, we tried two hyperparameter modifications. As suggested in [95], we increased n_{atom} by an order of magnitude to improve gradient quality, but this had little effect on the optimization (much overlap between same random

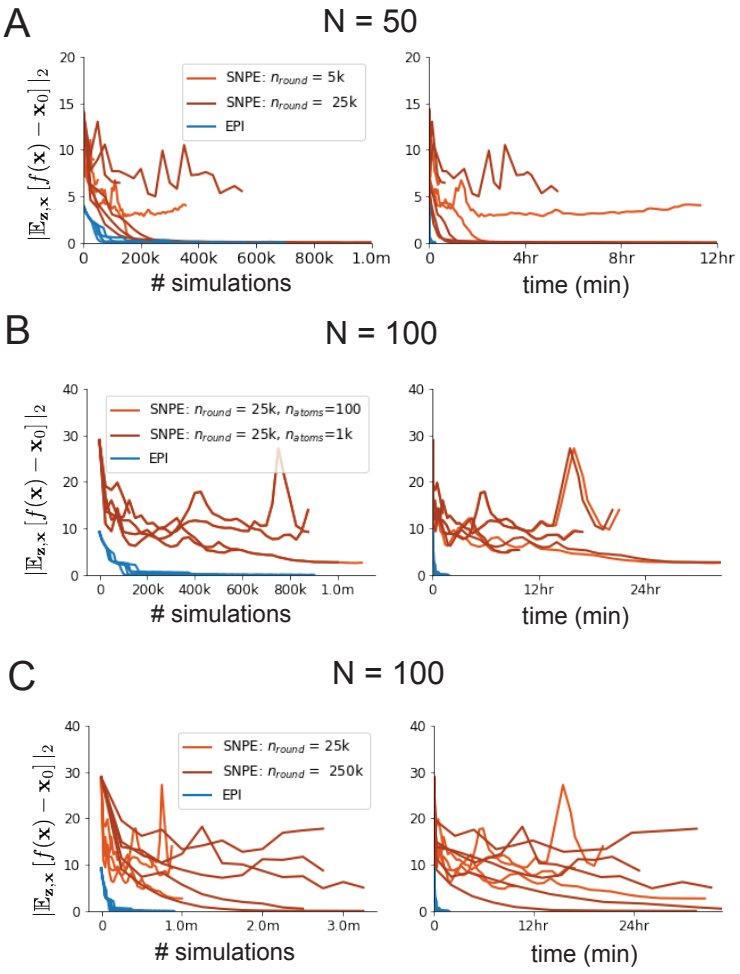

Figure 11: (RNN3): SNPE convergence was enabled by increasing n_{round}, not n_{atom}. **A**. Difference of mean predictions $\mathbf{x}_0$ throughout optimization at $N = 50$ with by simulation count (left) and wall time (right) of SNPE with $n_{\text{round}} = 5,000$ (light orange), SNPE with $n_{\text{round}} = 25,000$ (dark orange), and EPI (blue). Each line shows an individual random seed. **B**. Same conventions as A at $N = 100$ of SNPE with $n_{\text{atom}} = 100$ (light orange) and $n_{\text{atom}} = 1,000$ (dark orange). **C**. Same conventions as A at $N = 100$ of SNPE with $n_{\text{round}} = 25,000$ (light orange) and $n_{\text{round}} = 250,000$ (dark orange).

seeds) (Fig. 11B). Finally, we increased n_{round} by an order of magnitude, which yielded convergence in one case, but no others. We found no way to improve the convergence rate of SNPE without making more aggressive hyperparameter choices requiring high numbers of simulations. In Figure 2C-D, we show samples from the random seed resulting in emergent property convergence at greatest entropy (EPI), the random seed resulting in greatest validation log probability (SNPE), and the result of all converged random seeds (SMC).

5.3.5 Effect of RNN parameters on EPI and SNPE inferred distributions

To clarify the difference in objectives of EPI and SNPE, we show their results on RNN models with different numbers of neurons N and random strength g. The parameters inferred by EPI consistently produces the same mean and variance of real(λ_1) and λ_1^s, while those inferred by SNPE change according to the model definition (Fig. 12A). For $N = 2$ and $g = 0.01$, the SNPE posterior has greater concentration in eigenvalues around $\mathbf{x}_0$ than at $g = 0.1$, where the model has greater randomness (Fig. 12B top, orange). At both levels of g when $N = 2$, the posterior of SNPE has lower entropy than EPI at convergence (Fig. 12B top). However at $N = 10$, SNPE results in a predictive distribution of more widely dispersed eigenvalues (Fig. 12A bottom), and an inferred posterior with greater entropy than EPI (Fig. 12B bottom). We highlight these differences not to focus on an insightful trend, but to emphasize that these methods optimize different objectives with different implications.

Note that SNPE converges when it's validation log probability has saturated after several rounds of optimization (Fig. 12C), and that EPI converges after several epochs of its own optimization to enforce the emergent property constraints (Fig. 12D blue). Importantly, as SNPE optimizes its posterior approximation, the predictive means change, and at convergence may be different than $\mathbf{x}_0$ (Fig. 12D orange, left). It is sensible to assume that predictions of a well-approximated SNPE posterior should closely reflect the data on average (especially given a uniform prior and a low degree of stochasticity), however this is not a given. Furthermore, no aspect of the SNPE optimization controls the variance of the predictions (Fig. 12D orange, right).

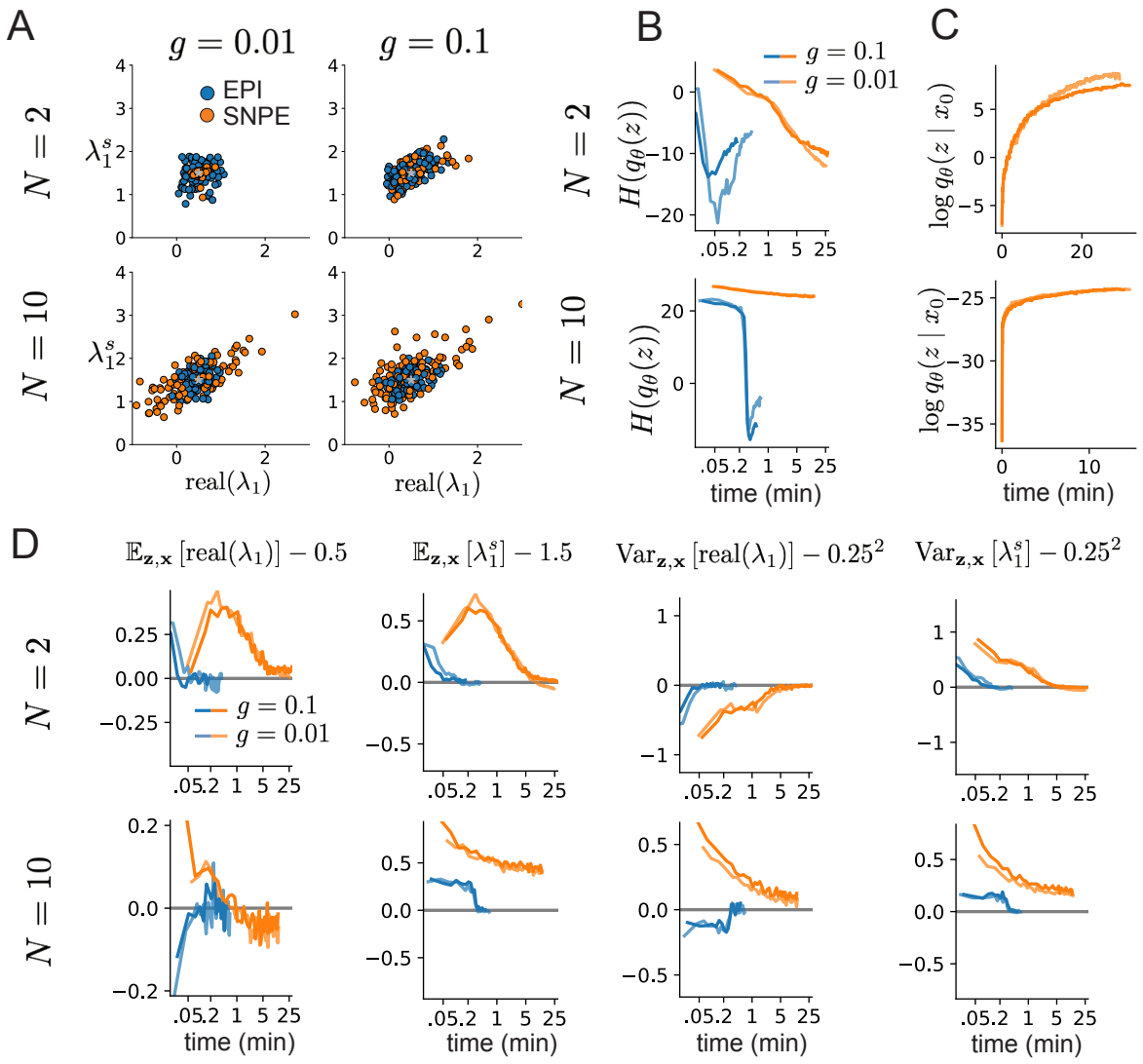

Figure 12: (RNN2): Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A**. Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with $N = 2$ (top), $N = 10$ (bottom), $g = 0.01$ (left), and $g = 0.1$ (right). **B**. Entropy of parameter distribution approximations throughout optimization with $N = 2$ (top), $N = 10$ (bottom), $g = 0.1$ (dark shade), and $g = 0.01$ (light shade). **C**. Validation log probabilities throughout SNPE optimization. Same conventions as B. **D**. Adherence to EPI constraints. Same conventions as B.

5.4 Primary visual cortex

5.4.1 V1 model

E-I circuit models, rely on the assumption that inhibition can be studied as an indivisible unit, despite ample experimental evidence showing that inhibition is instead composed of distinct elements [65]. In particular three types of genetically identified inhibitory cell-types – parvalbumin (P), somatostatin (S), VIP (V) – compose 80% of GABAergic interneurons in V1 [63–65], and follow specific connectivity patterns (Fig. 3A) [66], which lead to cell-type specific computations [47, 96]. Currently, how the subdivision of inhibitory cell-types, shapes correlated variability by reconfiguring recurrent network dynamics is not understood.

In the stochastic stabilized supralinear network [60], population rate responses $\mathbf{x}$ to mean input $\mathbf{h}$, recurrent input $W\mathbf{x}$ and slow noise $\boldsymbol{\epsilon}$ are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \tag{65}$$

where the noise is an Ornstein-Uhlenbeck process $\boldsymbol{\epsilon} \sim OU(\tau_{\text{noise}}, \boldsymbol{\sigma})$

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}}\tilde{\sigma}_\alpha dB \tag{66}$$

with $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$. The noisy process is parameterized as

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \tag{67}$$

so that $\boldsymbol{\sigma}$ parameterizes the variance of the noisy input in the absence of recurrent connectivity ($W = \mathbf{0}$). As contrast $c \in [0, 1]$ increases, input to the E- and P-populations increases relative to a baseline input $\mathbf{h} = \mathbf{h}_b + c\mathbf{h}_c$. Connectivity ($W_{\text{fit}}$) and input ($\mathbf{h}_{b,\text{fit}}$ and $\mathbf{h}_{c,\text{fit}}$) parameters were fit using the deterministic V1 circuit model [47]

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \tag{68}$$

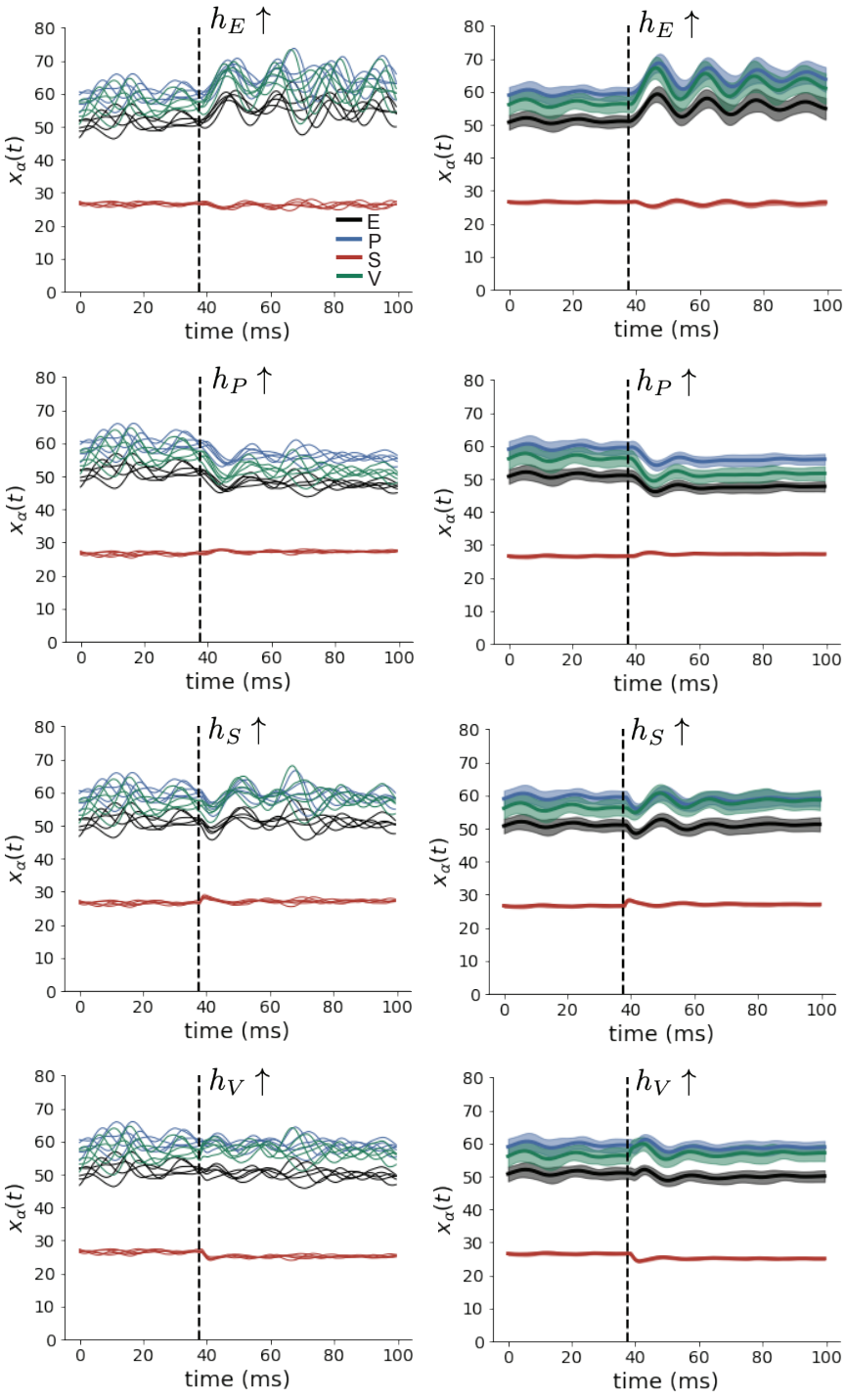

Figure 13: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

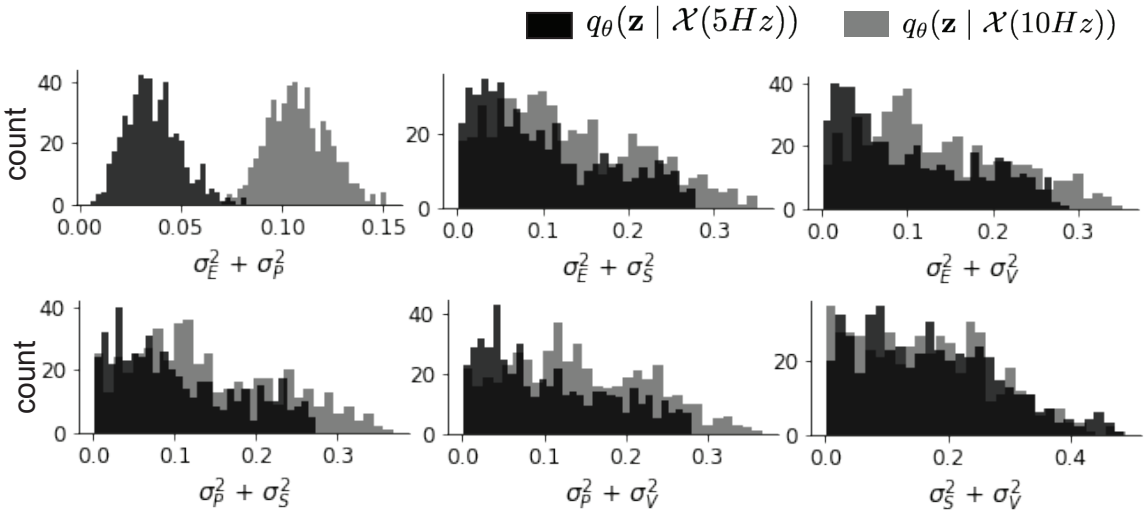

Figure 14: (V1 2) EPI predictive distributions of the sum of squares of each pair of noise parameters.

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \tag{69}$$

and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \tag{70}$$

To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equivalence class

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \tag{71}$$

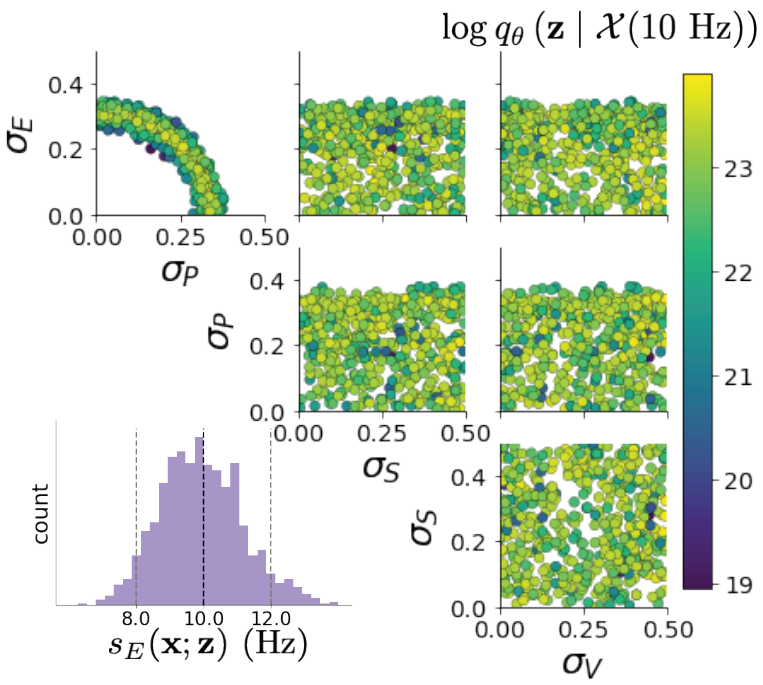

Figure 15: (V1 3) EPI inferred distribution for $\mathcal{X}(10\text{ Hz})$.

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \tag{72}$$

and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \tag{73}$$

Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5$ms from an initial condition drawn from $\mathbf{x}(0) \sim U\,[10\text{ Hz}, 25\text{ Hz}]$. Standard deviation of the E-population $s_E(\mathbf{x};\mathbf{z})$ is calculated as the square root of the temporal variance from $t_{ss} = 75$ms to $Tdt = 100$ms averaged over 100 independent trials.

$$s_E(\mathbf{x};\mathbf{z}) = \mathbb{E}_x\left[\sqrt{\mathbb{E}_{t>t_{ss}}\left[(x_E(t) - \mathbb{E}_{t>t_{ss}}\left[x_E(t)\right])^2\right]}\right] \tag{74}$$

5.4.2 EPI details for the V1 model

For EPI in Fig 3D-E, we used a real NVP architecture with three Real NVP coupling layers and two-layer neural networks of 50 units per layer. The normalizing flow architecture mapped $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ to a support of $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V] \in [0.0, 0.5]^4$. EPI optimization was run using three different random seeds for architecture initialization $\boldsymbol{\theta}$ with an augmented Lagrangian coefficient of $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the architectures converging with criteria $N_{\text{test}} = 100$ at greatest entropy across three random seeds.

5.4.3 Sensitivity analyses

In Fig. 3E, we visualize the modes of $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ throughout the σ_E-σ_P marginal. Specifically, we calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) = &\operatorname*{argmax}_{\mathbf{z}} \log q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}) \\ &\text{s.t. } \sigma_P = \sigma_{P,\text{fixed}} \end{aligned} \tag{75}$$

At each mode $\mathbf{z}^*$, we calculated the Hessian and visualized the sensitivity dimension in the direction of positive σ_E.

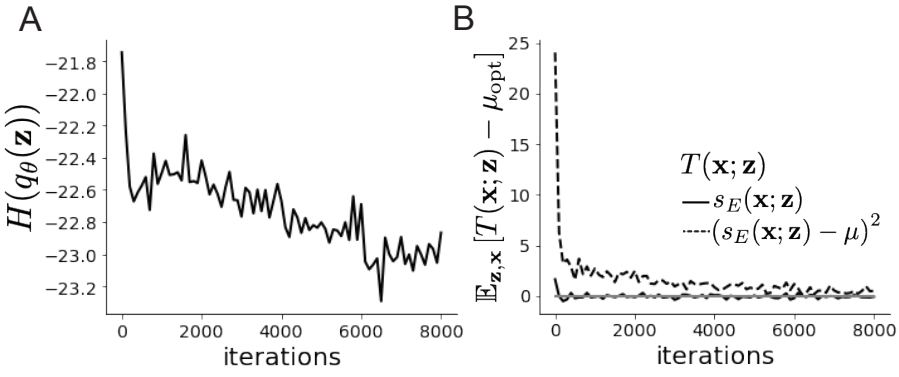

Figure 16: (V1 4) Optimization for V1

5.4.4 Primary visual cortex: Mathematical intuition and challenges

The dynamical system that we are working with can be written as

$$
\begin{aligned}
dx &= \frac{1}{\tau}(-x + f(Wx + h + \epsilon))dt \\
d\epsilon &= -\frac{dt}{\tau_{\text{noise}}}\epsilon + \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}}\Sigma_\epsilon dW
\end{aligned}
\tag{76}
$$

Where in this paper we chose

$$
\Sigma_\epsilon = \tau_{\text{noise}} \begin{bmatrix} \tilde{\sigma}_E & 0 & 0 & 0 \\ 0 & \tilde{\sigma}_P & 0 & 0 \\ 0 & 0 & \tilde{\sigma}_S & 0 \\ 0 & 0 & 0 & \tilde{\sigma}_V \end{bmatrix} \tag{77}
$$

where $\tilde{\sigma}_\alpha$ is the reparameterized standard deviation of the noise for population α from Equation 67.

In order to compute this covariance, we define $v = \omega x + h + \epsilon$ and $S = I - \omega f'(v))$, to re-write Eq. (76) as an 8-dimensional system:

$$
d\begin{pmatrix} \delta v \\ \epsilon \end{pmatrix} = -\begin{pmatrix} S & -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I \\ 0 & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} \begin{pmatrix} \delta v \\ \epsilon \end{pmatrix} dt + \begin{pmatrix} 0 & \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}}\Sigma_\epsilon \\ 0 & \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}}\Sigma_\epsilon \end{pmatrix} d\mathbf{W} \tag{78}
$$

Where $d\mathbf{W}$ is a vector with the private noise of each variable. The $d\mathbf{W}$ term is multiplied by a non-diagonal matrix is because the noise that the voltage receives is the exact same than the one that comes from the OU process and not another process. The solution of this problem is given by the Lyapunov Equation [60, 68]:

$$
\begin{pmatrix} S & -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I \\ 0 & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} \begin{pmatrix} \Lambda_v & \Lambda_c \\ \Lambda_c^T & \Lambda_\epsilon \end{pmatrix} + \begin{pmatrix} \Lambda_v & \Lambda_c \\ \Lambda_c^T & \Lambda_\epsilon \end{pmatrix} \begin{pmatrix} S^T & 0 \\ -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} = \begin{pmatrix} \frac{2}{\tau_{\text{noise}}}\Lambda_\epsilon & \frac{2}{\tau_{\text{noise}}}\Lambda_\epsilon \\ \frac{2}{\tau_{\text{noise}}}\Lambda_\epsilon & \frac{2}{\tau_{\text{noise}}}\Lambda_\epsilon \end{pmatrix} \tag{79}
$$

To obtain an equation for Λ_v, we solve this block matrix multiplication:

$$
S\Lambda_v + \Lambda_v S^T = \frac{2\Lambda_\epsilon}{\tau_{\text{noise}}} + \frac{\tau_{\text{noise}}^2 - \tau^2}{(\tau \tau_{\text{noise}})^2} \left((\frac{1}{\tau_{\text{noise}}} I + S)^{-1} \Lambda_\epsilon + \Lambda_\epsilon (\frac{1}{\tau_{\text{noise}}} I + S^T)^{-1} \right) \tag{80}
$$

Which is another Lyapunov Equation, now in 4 dimensions. In the simplest case in which $\tau_{\text{noise}} = \tau$, the voltage is directly driven by white noise, and Λ_v can be expressed in powers of S and

S^T. Because S satisfies its own polynomial equation (Cayley Hamilton theorem), there will be 4 coefficients for the expansion of S and 4 for S^T, resulting in 16 coefficients that define Λ_v for a given S. Due to symmetry arguments [68], in this case the diagonal elements of the covariance matrix of the voltage will have the form:

$$\Lambda_{v_{ii}} = \sum_{i=\{E,P,S,V\}} g_i(S)\sigma_{ii}^2 \tag{81}$$

These coefficients $g_i(S)$ are complicated functions of the Jacobian of the system. Although expressions for these coefficients can be found explicitly, only numerical evaluation of those expressions determine which components of the noisy input are going to strongly influence the variability of excitatory population. Showing the generality of this dependence in more complicated noise scenarios (e.g. $\tau_{\text{noise}} > \tau$ as in Section 3.4), is the focus of current research.

5.5 Superior colliculus

5.5.1 SC model

The ability to switch between two separate tasks throughout randomly interleaved trials, or "rapid task switching," has been studied in rats, and midbrain superior colliculus (SC) has been show to play an important in this computation [69]. Neural recordings in SC exhibited two populations of neurons that simultaneously represented both task context (Pro or Anti) and motor response (contralateral or ipsilateral to the recorded side), which led to the distinction of two functional classes: the Pro/Contra and Anti/Ipsi neurons [48]. Given this evidence, Duan et al. proposed a model with four functionally-defined neuron-type populations: two in each hemisphere corresponding to the Pro/Contra and Anti/Ipsi populations. We study how the connectivity of this neural circuit governs rapid task switching ability.

The four populations of this model are denoted as left Pro (LP), left Anti (LA), right Pro (RP) and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2}\tanh\left(\frac{u_\alpha - a}{b}\right) + \frac{1}{2}\right), \tag{82}$$

where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-

1289 earity. We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \tag{83}$$

1290 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \tag{84}$$

1291 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2. These
1292 hyperparameter values are motivated by modeling choices and results from [48].

1293 The weight matrix has 4 parameters for self sW, vertical vW, horizontal hW, and diagonal dW
1294 connections:

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \tag{85}$$

1295 We study the role of parameters $\mathbf{z} = [sW, vW, hW, dW]^\top$ in rapid task switching.

1296 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \tag{86}$$

1297 There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]\top, \tag{87}$$

1298 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]\top, \tag{88}$$

1299 rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \tag{89}$$

1300

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \tag{90}$$

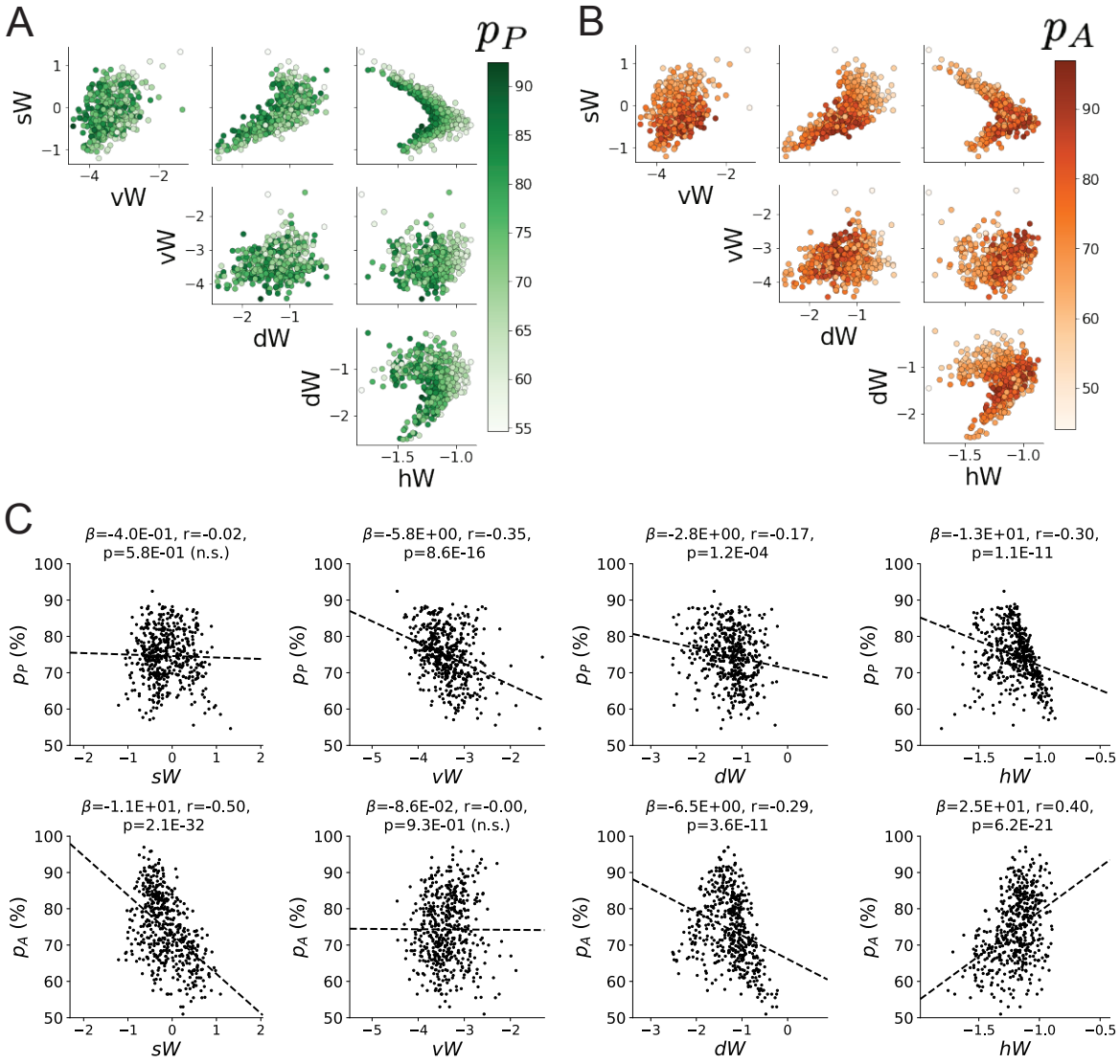

Figure 17: (SC1): **A**. Same pairplot as Fig. 4C colored by Pro task accuracy. **B**. Same as A colored by Anti task accuracy. **C**. Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

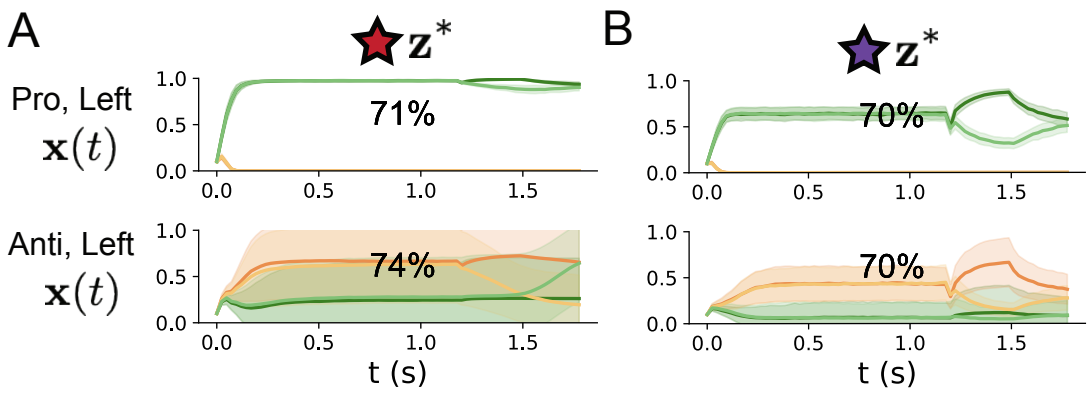

Figure 18: (SC2): **A**. Simulations in network regime 1 ($hW_{\text{fixed}} = -1.5$). **B**. Simulations in network regime 2 ($hW_{\text{fixed}} = -1.5$) .

a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1,1,1,1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \tag{91}$$

and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1,1,0,0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0,0,1,1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases}. \tag{92}$$

The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{\text{P,bias}} = 0.5$, $I_{\text{P,rule}} = 0.6$, $I_{\text{A,rule}} = 0.6$, $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

5.5.2 Task accuracy calculation

The accuracies of each task p_P and p_A are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} \left[\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]\right] \tag{93}$$

and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} \left[\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]\right] \tag{94}$$

given that the stimulus is on the left side, where Θ is the Heaviside step function, and the accuracy is averaged over 200 independent trials. The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \tag{95}$$

where $\beta = 100$.

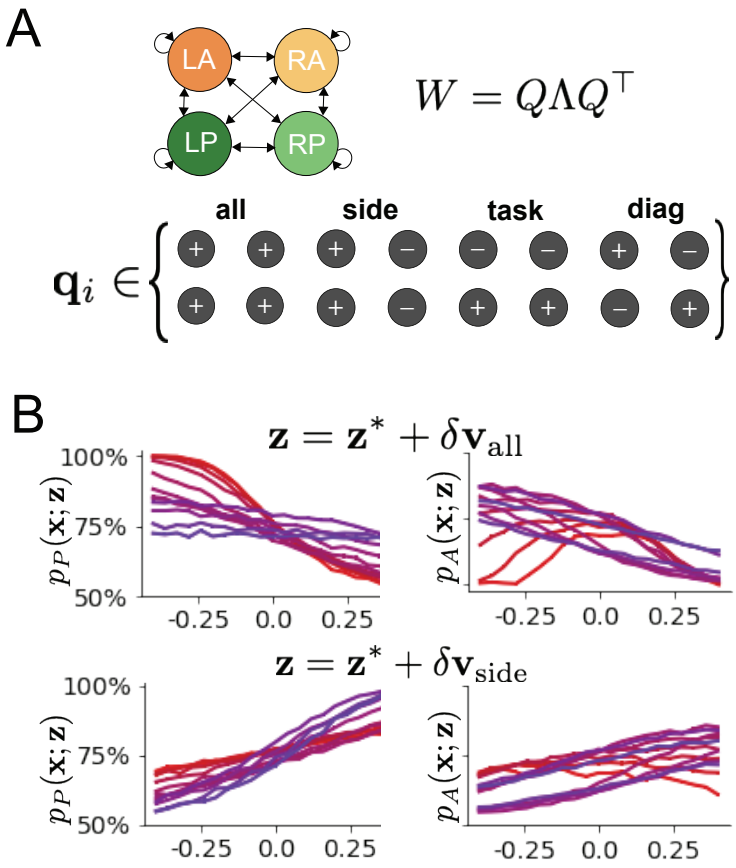

Figure 19: (SC3): **A**. Invariant eigenvectors of connectivity matrix W. **B**. Accuracies for connectivity perturbations for increasing λ_{all} and λ_{side} (rest shown in Fig. 4D).

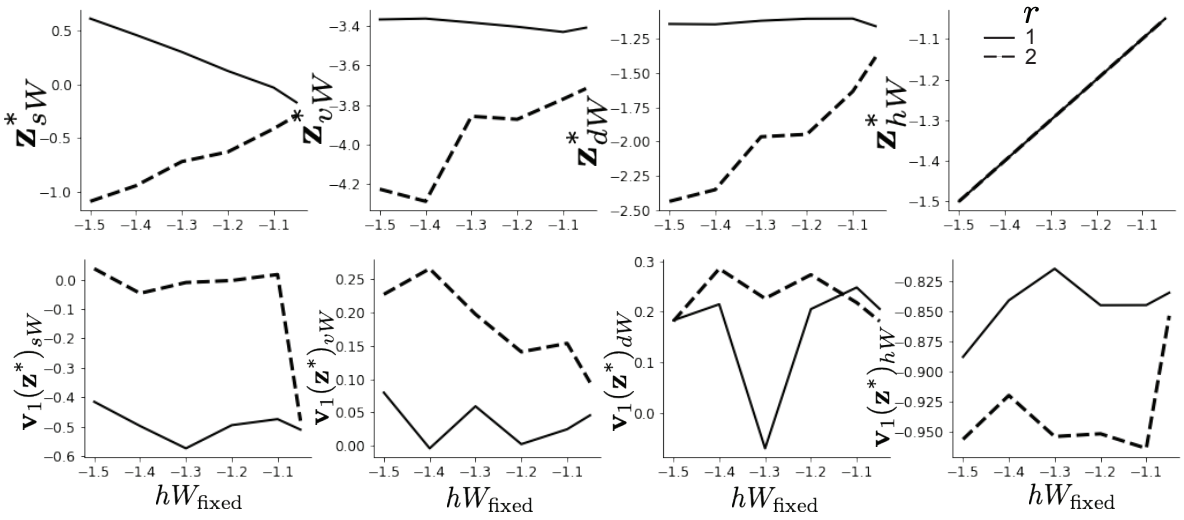

Figure 20: (SC4): **A**. The individual parameters of each mode throughout the two regimes. **B**. The individual sensitivities of parameters of each mode throughout the two regimes.

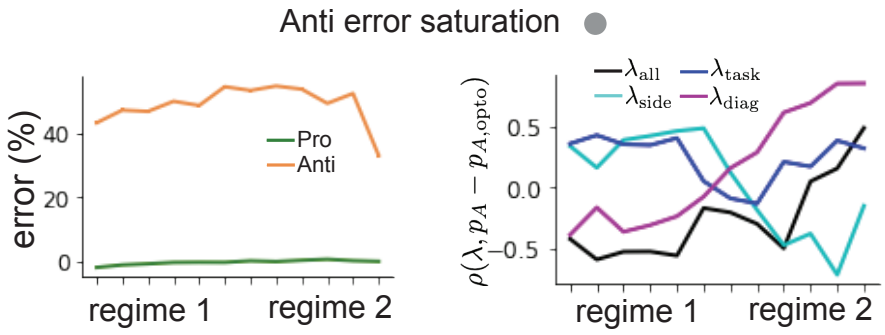

Figure 21: (SC5): (Left) Mean and standard error of Pro and Anti error from regime 1 to regime 2 at $\gamma = 0.85$. (Right) Correlations of connectivity eigenvalues with Anti error from regime 1 to regime 2 at $\gamma = 0.85$.

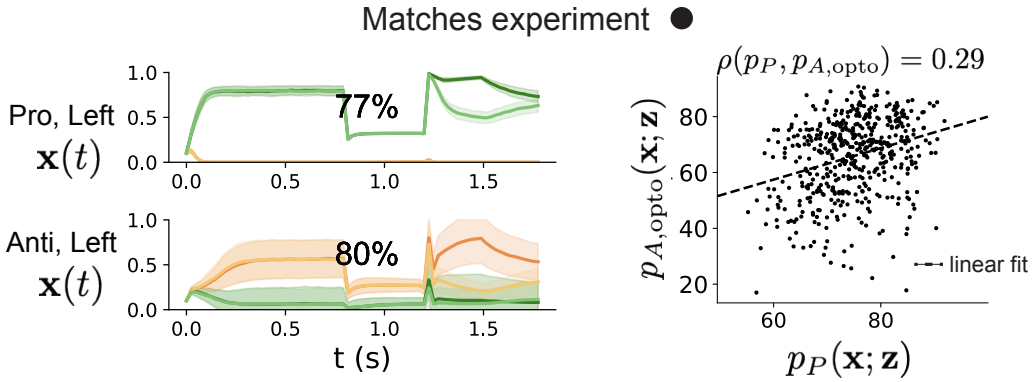

Figure 22: (SC6): (Left) Mean and standard deviation (shading) of responses of the SC model at the mode of the EPI distribution to delay period inactivation at $\gamma = 0.675$. (Right) Anti accuracy following delay period inactivation at $\gamma = 0.675$ versus accuracy in the Pro task across connectivities in the EPI distribution.

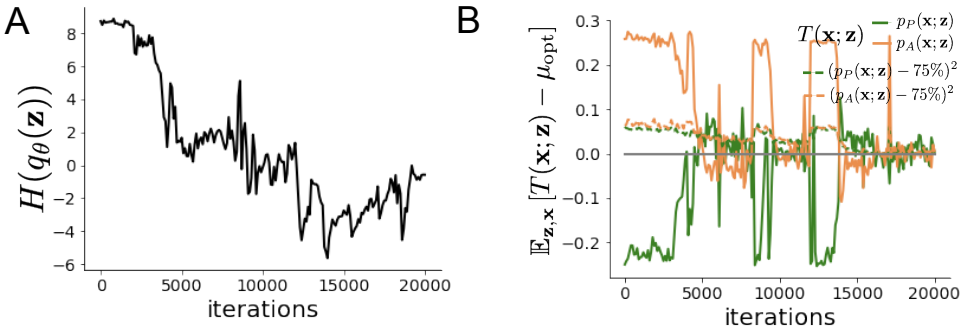

Figure 23: (SC7): **A**. Entropy throughout optimization. **B**. The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

5.5.3 EPI details for the SC model

Writing the EPI distribution as a maximum entropy distribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments of the accuracy in each task (as in Equations 17 and 18)

$$T(\mathbf{x};\mathbf{z}) = \begin{bmatrix} p_P(\mathbf{x};\mathbf{z}) \\ p_A(\mathbf{x};\mathbf{z}) \\ (p_P(\mathbf{x};\mathbf{z}) - .75)^2 \\ (p_A(\mathbf{x};\mathbf{z}) - .75)^2 \end{bmatrix}, \tag{96}$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} .75 \\ .75 \\ .075^2 \\ .075^2 \end{bmatrix}. \tag{97}$$

Throughout optimization, the augmented Lagrangian parameters η and c, were updated after each epoch of 2,000 iterations (see Section 5.1.4). The optimization converged after ten epochs (Fig. 22).

For EPI in Fig. 4C, we used a real NVP architecture with three coupling layers of affine transformations parameterized by two-layer neural networks of 50 units per layer. The initial distribution was a standard isotropic gaussian $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to a support of $\mathbf{z}_i \in [-5, 5]$. We used an augmented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution was the greatest EPI distribution to converge across 5 random seeds with criteria $N_{\text{test}} = 25$.

The bend in the EPI distribution is not a spurious result of the EPI optimization. The structure discovered by EPI matches the shape of the set of points returned from brute-force random sampling

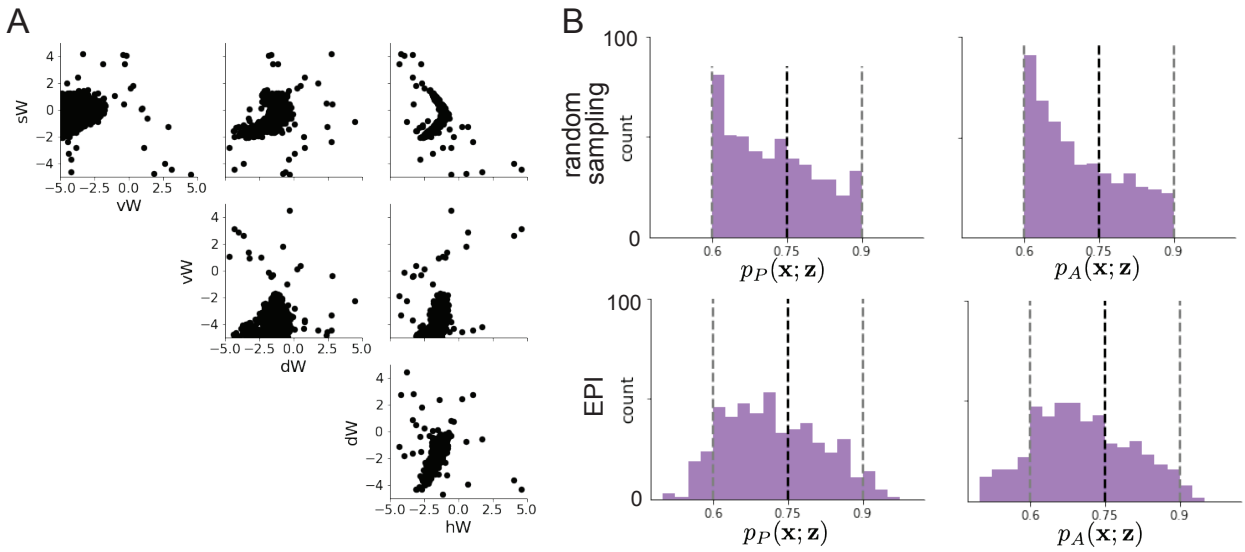

Figure 24: (SC8): **A**. Entropy throughout optimization. **B**. The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

(Fig. 24A) These connectivities were sampled from a uniform distribution over the range of each connectivity parameter, and all parameters producing accuracy in each task within the range of 60% to 90% were kept. This set of connectivities will not match the distribution of EPI exactly, since it is not conditioned on the emergent property. For example the parameter set returned by the brute-force search is biased towards lower accuracies (Fig. 24B).

5.5.4 Regime identification with EPI

We sought two sets of parameters from $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ that were representative of each regime, so that we could assess their implications on computation. For fixed values of hW, we hypothesized that there are two modes: one in each regime of greater and lesser sW. To begin, we found one mode for each regime at $hW_{\text{fixed}} = -1.5$ using 200 steps of gradient ascent of the deep probability distribution $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$. In regime 1, the initialization had positive sW, and the initialization had negative sW in regime 2, which led to disparate modes (Fig. 20 top). These modes were then used as the initialization to find the next mode at $hW_{\text{fixed}} = -1.4$ and so on. 200 steps of gradient ascent were always taken, and learning rates of $2.5 \times 10^{-}4$ and $5 \times 10^{-}4$ were used for regimes 1 and 2, respectively. Each of these modes is denoted $\mathbf{z}^*(hW_{\text{fixed}}, r)$ for regime $r \in \{1, 2\}$.

For the analyses in Figure 5C and Figure 21, we obtained parameters for each step along the continuum between regimes 1 and 2 by sampling from the EPI distribution. Each sample was

assigned to the closest mode $\mathbf{z}^*(hW_{\text{fixed}}, r)$. Sampling continued until 500 samples were assigned to each mode, which took 7.36 seconds. To obtain this many samples for each mode with brute force sampling over the chosen prior, this would take 4.20 days.

5.5.5 Sensitivity analysis

At each mode, we measure the sensitivity dimension (that of most negative eigenvalue in the Hessian of the EPI distribution) $\mathbf{v}_1(\mathbf{z}^*)$. To resolve sign degeneracy in eigenvectors, we chose $\mathbf{v}_1(\mathbf{z}^*)$ to have negative element in hW. This tells us what parameter combination rapid task switching is most sensitive to at this parameter choice in the regime. We see that while the modes of each regime gradually converge to similar connectivities at $hW_{\text{fixed}} = -1.05$ (Fig. 20 top), the sensitivity dimensions remain categorically different throughout the two regimes (Fig. 20 bottom) Only at $hW_{\text{fixed}} = -1.05$ is there a flip in sensitivity from regime 2 to regime 1 (in $\mathbf{v}_1(\mathbf{z}^*)_{sW}$ and $\mathbf{v}_1(\mathbf{z}^*)_{hW}$). There is thus some ambiguity regarding the "regime" of $\mathbf{z}^*(-1.05, 2)$, since the mode is derived from an initialization in regime 2, but has sensitivity like regime 1. We can consider this as an intermediate transitionary region of parameter space between the two regimes. To emphasize this, $\mathbf{z}^*(-1.05, 1)$ and $\mathbf{z}^*(-1.05, 2)$ have the same color.

5.5.6 Connectivity eigendecomposition and processing modes

To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposition of the symmetric connectivity matrices $W = Q\Lambda Q^{-1}$, which results in the same basis vectors $\mathbf{q}_i$ for all W parameterized by $\mathbf{z}$ (Fig. 19A). These basis vectors have intuitive roles in processing for this task, and are accordingly named the *all* eigenmode - all neurons co-fluctuate, *side* eigenmode - one side dominates the other, *task* eigenmode - the Pro or Anti populations dominate the other, and *diag* mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair. Due to the parametric structure of the connectivity matrix, the parameters $\mathbf{z}$ are a linear function of the eigenvalues $\boldsymbol{\lambda} = [\lambda_{\text{all}}, \lambda_{\text{side}}, \lambda_{\text{task}}\lambda_{\text{diag}}]^\top$ associated with these eigenmodes.

$$\mathbf{z} = A\boldsymbol{\lambda} \tag{98}$$

$$A = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}. \tag{99}$$

We are interested in the effect of raising or lowering the amplification of each eigenmode in the connectivity matrix. To test this, we calculate the unit vector of changes in the connectivity $\mathbf{z}$ that result from a change in the associated eigenvalues

$$\mathbf{v}_{\mathrm{a}} = \frac{\frac{\partial \mathbf{z}}{\partial \lambda_{\mathrm{a}}}}{|\frac{\partial \mathbf{z}}{\partial \lambda_{\mathrm{a}}}|_2}, \tag{100}$$

where

$$\frac{\partial \mathbf{z}}{\partial \lambda_{\mathrm{a}}} = A\mathbf{e}_{\mathrm{a}}, \tag{101}$$

and e.g. $\mathbf{e}_{\mathrm{all}} = [1, 0, 0, 0]^\top$. So $\mathbf{v}_{\mathrm{a}}$ is the normalized column of A corresponding to eigenmode a. While perturbations in the sensitivity dimension $\mathbf{v}_1(\mathbf{z}^*)$ adapt with the mode $\mathbf{z}^*$ chosen, perturbations in $\mathbf{v}_a$ for $a \in \{\mathrm{all}, \mathrm{side}, \mathrm{text}, \mathrm{diag}\}$ are invariant to $\mathbf{z}$ (Equation 101).

To understand the connectivity mechanisms that distinguish these two regimes, we perturb connectivity at each mode in dimensions that have well defined roles in processing for the Pro and Anti tasks. A convenient property of this connectivity parameterization is that there are $\mathbf{z}$-invariant eigenmodes of connectivity, whose eigenvalues (or degree of amplification) change with $\mathbf{z}$. These eigenmodes have intuitive roles in processing in each task, and are accordingly named the *all*, *side*, *task*, and *diag* eigenmodes (see Section 5.5). Furthermore, the parameter dimension $\mathbf{v}_a$ ($a \in \{\mathrm{all}, \mathrm{side}, \mathrm{task}, \text{ and } \mathrm{diag}\}$) that increases the eigenvalue of connectivity λ_a is $\mathbf{z}$-invariant (unlike the sensitivity dimension $\mathbf{v}_1(\mathbf{z})$) and $\mathbf{v}_a \perp \mathbf{v}_{b \neq a}$. Thus, by changing the degree of amplification of each processing mode by perturbing $\mathbf{z}$ along $\mathbf{v}_a$, we can elicit the differentiating properties of the two regimes.

5.5.7 Modeling optogenetic silencing.

We tested whether the inferred SC model connectivities could reproduce experimental effects of optogenetic inactivation in rats [69]. During periods of simulated optogenetic inactivation, activity was decreased proportional to the optogenetic strength $\gamma \in [0, 1]$

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \tag{102}$$

Delay period inactivation was from $0.8 < t < 1.2$.