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 in terms 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. When analytic derivation of the relationship between model parameters and computational properties is intractable, approximate inference and simulation-based techniques are relied upon for scientific insight. We bring the use of deep generative models for probabilistic inference to bear on this problem, learning distributions of parameters that produce the specified properties of computation. By learning parameter distributions that produce computations – an emergent property, we introduce a novel methodology that is particularly well-suited to the stochastic dynamical systems models predominant in our field of theoretical neuroscience. We motivate this methodology with a worked example analyzing sensitivity in the stomatogastric ganglion. We then use it to reveal the key factors of variability in a model of primary visual cortex, gain a mechanistic understanding of rapid task switching in superior colliculus models, and scale inference of large low-rank RNN's exhibiting stable amplification. While much use of deep learning in theoretical neuroscience focuses on drawing analogies between optimized neural architectures and the brain, this work illustrates how we can further leverage the power of deep learning towards solving inverse problems in theoretical neuroscience.

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 [1]. A neural computation is systematized with a set of equations – the model – and these equations are motivated by biophysics, neurophysiology, and other conceptual considerations. 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 these particular parameter configurations? The inverse problem is crucial for reasoning about likely parameter values, uniquenesses and degeneracies, and predictions made by the model.

Consider the idealized practice: one carefully designs a model and analytically derives how model parameters govern the computation. Seminal examples of this gold standard (which often adopt approaches from statistical physics) include our field's understanding of memory capacity in associative neural networks [2], chaos and autocorrelation timescales in random neural networks [3], the paradoxical effect [4], and decision making [5]. Unfortunately, as circuit models include more biological realism, theory via analytical derivation becomes intractable. Alternatively, statistical inference can be run to obtain model parameters likely to produce some model output, and local sensitivity analyses can be performed at inferred parameter values. Since most neural circuit models stipulate a noisy system of differential equations that can only be sampled or realized through forward simulation, they lack the explicit likelihood central to the probabilistic modeling toolkit. Therefore, the most popular approaches to the inverse problem have been likelihood-free methods such as approximate Bayesian computation (ABC) [6], in which a set of reasonable parameters estimates is obtained via simulation and rejection.

Of course, the challenge of doing inference in complex models has arisen in many scientific fields. In response, the machine learning community has made remarkable progress in recent years, via the use of deep neural networks as a powerful inference engine: a flexible function family that can map observations back to probability distributions quantifying the likely parameter configurations. One celebrated example of this approach from machine learning, of which we draw key inspiration for this work, is the variational autoencoder (VAE) [7, 8], which uses a deep neural network to induce an (approximate) posterior distribution on hidden variables in a latent variable model, given data. Indeed, these tools have been used to great success in neuroscience as well, in particular for

interrogating parameters (sometimes treated as hidden states) in models of both cortical population activity [9, 10, 11, 12] and animal behavior [13, 14, 15]. These works have used deep neural networks to expand the expressivity and accuracy of statistical models of neural data [16].

Existing approaches to the inverse problem in theoretical neuroscience fall short in three key ways. First, theoretical models of neural computation aim to reflect a complex biological reality, and as a result, such models lack tractable likelihoods. Thus, standard approaches from statistical inference are unavailable. The parameter sets obtained from likelihood-free ABC lack a formalized link to Bayesian inference (except in the unrealistic 0-distance scenario), and lack parameter probabilities. Second is the undesirable trade-off between the flexibility and tractability of the approximated posterior distribution. While sampling-based approaches like ABC and Markov chain Monte Carlo (MCMC) can produce flexible posterior approximations, they must be run continually for increasing samples. While VAE approaches can result in tractable posterior sampling and sensitivity measurements post-optimization, existing approaches have relied on simplified classes of distributions, which restrict the flexibility of the posterior approximation. And third, one can never assume what inferred model parameters may predict. This is well understood when considering Box's loop and the role of posterior predictive checks in the development and critique of scientific models [17, 18]. Uncertainty about the properties of inferred model predictions introduce a conceptual degree of freedom to the inverse problem that may be unnecessary and undesirable given the scientific motivation.

To address these three challenges, we developed an inference methodology – 'emergent property inference' – which learns a distribution over parameter configurations in a theoretical model. This distribution has two critical properties: *(i)* it is chosen such that draws from the distribution (parameter configurations) correspond to systems of equations that give rise to a specified emergent property (a set of constraints); and *(ii)* it is chosen to have maximum entropy given those constraints, such that we identify all likely parameters and can use the distribution to reason about parametric sensitivity and degeneracies [19]. First, we use stochastic gradient techniques in the spirit of likelihood-free variational inference [20] to enable inference in likelihood-free models of neural computation. Second, we stipulate a bijective deep neural network that induces a flexible family of probability distributions over model parameterizations with a probability density we can calculate [21, 22, 23], which confers fast sampling and sensitivity measurements. Third, we quantify the notion of emergent properties as a set of moment constraints on datasets generated by the model. Thus, an emergent property is not a single data realization, but a phenomenon or a feature

of the model, which is ultimately the object of interest in theoretical neuroscience. Conditioning on an emergent property requires a variant of deep probabilistic inference methods, which we have previously introduced [24]. Taken together, emergent property inference (EPI) provides a methodology for inferring parameter configurations consistent with a particular emergent phenomena in theoretical models. We use a classic example of parametric degeneracy in a biological system, the stomatogastric ganglion [25], to motivate and clarify the technical details of EPI.

Equipped with this methodology, we then investigated three models of current importance in theoretical neuroscience. These models were chosen to demonstrate generality through ranges of biological realism (from conductance-based biophysics to recurrent neural networks), neural system function (from pattern generation to abstract cognitive function), and network scale (from four to hundreds of neurons). First, we use EPI to understand the characteristics of noise that govern Fano factor in a stochastic four neuron-type model of primary visual cortex. Second, we discover connectivity patterns in superior colliculus resilient to optogenetic perturbation by using EPI to condition on rapid task switching. The novel scientific insights offered by EPI contextualize and clarify the previous studies exploring these models [26, 27]. Third, we emphasize the methodological advancement of EPI by inferring high-dimensional distributions of RNN connectivities exhibiting stable amplification. These results point to the value of deep inference for the interrogation of biologically relevant 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 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 [28]. 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 neural activity [25, 29]. A subcircuit model of the STG [30] is shown schematically in Figure 1A, and note that the behavior of this model will be critically dependent on its parameterization – the choices of conductance parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$. Specifically, the two fast neurons ($f1$ and $f2$) mutually inhibit one another, and

oscillate at a faster frequency than the mutually inhibiting slow neurons ($s1$ and $s2$). The hub neuron (hub) couples with either the fast or slow population or both.

Second, once the model is selected, one defines the emergent phenomena of scientific interest. In the STG example, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model 1B. An interesting emergent property of this stochastic model is when the hub neuron fires 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.

Third, parameter analyses ensue: brute-force parameter sweeps, ABC sampling, and sensitivity analyses are all routinely used to reason about what parameter configurations lead to an emergent property. In this last step lies the opportunity for a precise quantification of the emergent property as a statistical feature of the model. Once we have such a methodology, we can infer a probability distribution over parameter configurations that produce 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, produces the parameter distribution EPI (Fig. 1E). This distribution – represented for clarity as samples from the distribution – is then a scientifically meaningful and mathematically tractable object. In the STG model, this distribution can be specifically queried to reveal the prototypical parameter configuration for network syncing (the mode; Figure 1E yellow star), and how network syncing decays based on changes away from the mode. The eigenvectors (of the Hessian of the distribution at the mode) quantitatively formalize the robustness of unified intermediacy (Fig. 1B solid (v_1) and dashed (v_2) black arrows). Indeed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity (v_1) and degeneracy (v_2) agree with error contours (Fig. 1B contours) and have diminished or preserved network syncing, respectively (Fig. 1F activity traces, Fig. S TODO) (see Section 5.2.1).

3.2 A deep generative modeling approach to emergent property inference

Emergent property inference (EPI) systematizes the three-step procedure of the previous section. First, we consider the model as a coupled set of differential equations [30]. In the running STG example, the model activity $\mathbf{x} = [x_{\text{f1}}, x_{\text{f2}}, x_{\text{hub}}, x_{\text{s1}}, x_{\text{s2}}]$ is the membrane potential for each neuron, which evolves according to the biophysical conductance-based equation:

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

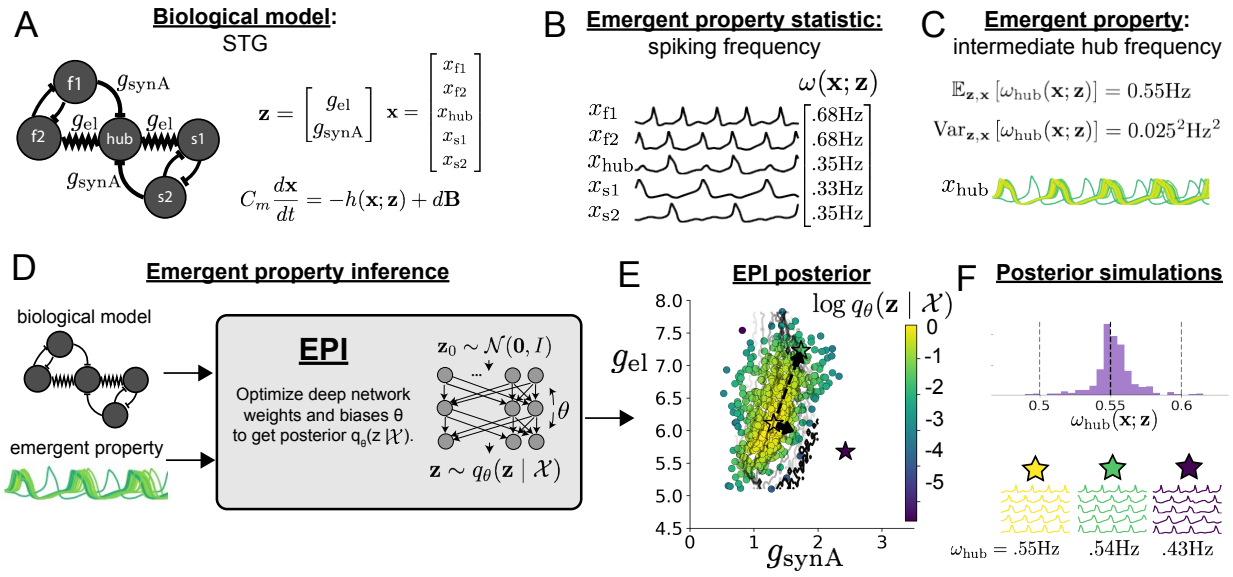

Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A**. Conductance-based biophysical model of the STG subcircuit. In the STG model, jagged connections indicate electrical coupling having electrical conductance g_{el}. Other connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto the hub neuron, and $g_{\text{synB}} = 5\text{nS}$ for mutual inhibitory connections. Parameters are represented by the vector $\mathbf{z}$ and membrane potentials by the vector $\mathbf{x}$. The evolution of this model's activity $\mathbf{x}(t)$ is predicated by differential equations. **B**. Spiking frequency $\omega(\mathbf{x}; \mathbf{z})$ is an emergent property statistic. Spiking frequency is measured from simulated activity of the STG model at parameter choices of $g_{\text{el}} = 4.5\text{nS}$ and $g_{\text{synA}} = 3\text{nS}$. **C**. The emergent property of intermediate hub frequency, in which the hub neuron fires at a rate between the fast and slow frequencies. Simulated activity traces are colored by log probability density of their generating parameters in the EPI-inferred distribution (Panel E). **D**. For a choice of model and emergent property, emergent property inference (EPI) learns a distribution of the model parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$ producing intermediate hub frequency. Deep probability distributions map a simple random variable $\mathbf{z}_0$ through a deep neural network with weights and biases $\boldsymbol{\theta}$ to parameters $\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0)$ distributed as $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$. In EPI optimization, stochastic gradient steps in $\boldsymbol{\theta}$ are taken such that entropy is maximized, and the emergent property $\mathcal{X}$ is produced. **E**. The EPI distribution of STG model parameters producing intermediate hub frequency. Samples are colored by log probability density. Distribution contours of hub neuron frequency from mean of .55 Hz are shown at levels of .525, .53, ... .575 Hz (dark to light gray away from mean). Frequencies are averages over the stochasticity of the model. Eigenvectors of the Hessian at the mode of the inferred distribution are indicated as $\mathbf{v}_1$ (solid) and $\mathbf{v}_2$ (dashed) with lengths scaled by the square root of the absolute value of their eigenvalues. Simulated activity is shown for three samples (stars). **F** Simulations from parameters in E. (Top) The predictive distribution of the posterior obeys the constraints stipulated by the emergent property. The black and gray dashed lines show the mean and two standard deviations according the emergent property, respectively. (Bottom) Simulations at the starred parameter values.

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 $\mathbf{x}$ and $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$, and $d\mathbf{B}$ is white gaussian noise (see Section 5.2.1).

Second, we define the emergent property, which as above is "intermediate hub frequency" (Figure 1C). Quantifying this phenomenon is straightforward: we stipulate that the hub neuron's spiking frequency – denoted $\omega_{\text{hub}}(\mathbf{x})$ is close to an intermediate frequency of 0.55Hz. Mathematically, we achieve this via constraints on the mean and variance of the hub neuron spiking frequency.

$$\begin{aligned} \mathcal{X} \quad : \quad \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] \quad &\triangleq \quad \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[\omega_{\text{hub}}(\mathbf{x};\mathbf{z})\right] \quad = \quad \left[0.55\right] \quad \triangleq \quad \boldsymbol{\mu} \\ \text{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] \quad &\triangleq \quad \text{Var}_{\mathbf{z},\mathbf{x}}\left[\omega_{\text{hub}}(\mathbf{x};\mathbf{z})\right] \quad = \quad \left[0.025^2\right] \quad \triangleq \quad \boldsymbol{\sigma}^2. \end{aligned} \tag{2}$$

The emergent property statistic $f(\mathbf{x};\mathbf{z}) = \omega_{\text{hub}}(\mathbf{x};\mathbf{z})$ along with its constrained mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$ define the emergent property denoted $\mathcal{X}$.

Third, we perform emergent property inference: we find a distribution over parameter configurations $\mathbf{z}$, and insist that samples from this distribution produce the emergent property; in other words, they obey the constraints introduced in Equation 2. 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 generative distribution of the normalizing flow class [21, 22, 23] – neural networks which transform a simple distribution into a suitably complicated distribution (as is needed here). This deep distribution is represented in Figure 1C (see Section 5.1). Then, mathematically, we must solve the following optimization program:

$$\begin{aligned} q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}) \quad = \quad & \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\text{argmax}}\, H(q_{\boldsymbol{\theta}}(\mathbf{z})) \\ & \text{s.t.} \quad \mathcal{X} : \quad \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \boldsymbol{\sigma}^2 \end{aligned} \tag{3}$$

where $f(\mathbf{x},\mathbf{z}), \boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ are defined as in Equation 10. According to the emergent property of interest, $f(\mathbf{x},\mathbf{z})$ may contain multiple statistics, in which case the mean and variance vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ match this dimension. Finally, we recognize that many distributions in $\mathcal{Q}$ will respect the emergent property constraints, so we select that which has maximum entropy. This principle, captured in Equation 3 by the primal objective H, identifies parameter distributions with minimal assumptions beyond some chosen structure [31, 32, 24, 33]. Such a normative principle of maximum entropy, which is also that of Bayesian inference, naturally fits with our scientific objective of reasoning about parametric sensitivity and robustness. The recovered distribution of EPI is as variable as possible along each parametric manifold such that it produces the emergent property.

EPI optimizes the weights and biases $\boldsymbol{\theta}$ of the deep neural network (which induces the probability distribution) by iteratively solving Equation 3. The optimization is complete when the sampled models with parameters $\mathbf{z} \sim q_{\boldsymbol{\theta}}(z \mid \mathcal{X})$ produce activity consistent with the specified emergent property (Fig. S4). Such convergence is evaluated with a hypothesis test that the means and variances of each emergent property statistic are not different than their constrained values (see Section 5.1.3). Further validation of EPI is available in the supplementary materials, where we analyze a simpler model for which ground-truth statements can be made (Section 5.1.6).

In relation to broader methodology, inspection of the EPI objective reveals a natural relationship to posterior inference. Specifically, EPI executes a novel variant of Bayesian inference with a uniform prior and a gaussian likelihood on the emergent property statistic (see Section 5.1.5). A key advantage of EPI over established Bayesian inference is that the predictions made my the inferred distribution are constrained to produce the specified emergent property. Equipped with this method, we may examine structure in posterior distributions or make comparisons between posteriors conditioned at different levels of the same emergent property statistic. In Sections 3.3 and 3.4, we prove out the value of EPI by using it to investigate and produce novel insights into two prominent models in neuroscience. Subsequently in Section 3.5, we show EPI's superiority in parameter scalability and fidelity of the posterior predictive distribution by conditioning on stable amplification in low-rank RNNs.

3.3 EPI reveals how noise across neural population types governs Fano factor in a stochastic inhibition stabilized network

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 a regime characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to paradoxical responses [4], selective amplification [34, 35], surround suppression [36] and normalization [37]. Despite their strong predictive power, E-I circuit models rely on the assumption that inhibition can be studied as an indivisible unit. However, experimental evidence shows that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) – composing 80% of GABAergic interneurons in V1 [38, 39, 40], and that these inhibitory cell types follow specific connectivity patterns (Fig. 2A) [41]. Recent theoretical advances [26, 42, 43], have only started to address the consequences of this multiplicity in the dynamics of V1, strongly relying on linear theoretical tools. Here, we use EPI to characterize the properties of slow noise in a stochastic

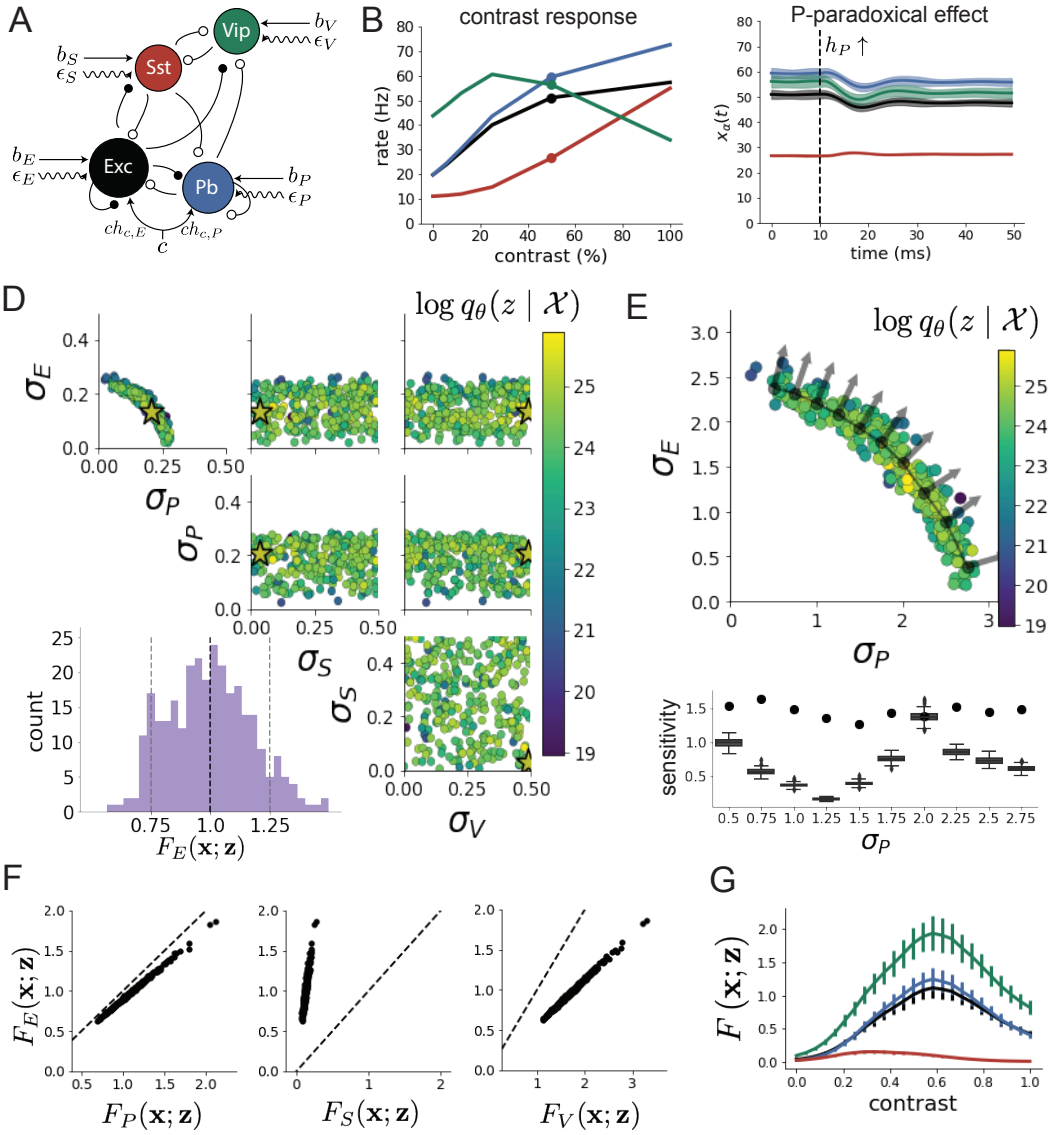

Figure 2: Emergent property inference of a stochastic stabilized supralinear network. **A**. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). 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 $\mathbf{h}_b$. Additionally, each neural population receives a slow noisy input $\boldsymbol{\epsilon}$. **B**. Responses of the deterministic smodel ($\boldsymbol{\epsilon} = \mathbf{0}$) to varying contrasts. The response at 50% contrast (dots) is the focus of our analysis. **C**. Paradoxical response of the stochastic model to a small increase in input to the P-population. **D**. EPI posterior of noise parameteres $\mathbf{z}$ conditioned on realistic E-population Fano factors. The posterior predictive distribution is show on the bottom-left. and the mode of the distribution is starred. **E**. (Top) Enlarged visualization of the σ_E-σ_P marginal distribution of the posterior. Each gray dot is a choice of σ_P, for which a constrained mode $z^*(\sigma_P, P)$ is chosen. The arrows show the most sensitive dimensions of the Hessian evaluated at these modes. (Bottom) Such sensitive dimensions of the Hessian (dots) are significantly more sensitive than randomly chosen dimensions (box and whiskers). **F**. The Fano factor of the E-population is strongly correlated with each other neuron-type population. **G**. Mean and standard deviation (across EPI posterior) of Fano factor of each neuron-type population at each level of contrast.

version of this model, which result in biologically realistic responses.

We considered the contrast response of a nonlinear dynamical V1 circuit model (Fig. 2A) 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 $W\mathbf{x}$ from synaptic projections of effective connectivity W and an external input $\mathbf{h}$, which determine the population rate via nonlinearity $\phi = []_+^2$ (see Section 5.2.2). The circuit model evolves form an initial condition $\mathbf{x}(0) \sim \mathcal{U}([10, 25])$ with time constant $\tau = 1$ms according to a contrast-dependent input $\mathbf{h}$ and slow noise $\boldsymbol{\epsilon}$ of time constant $\tau_{\text{noise}} = 5$ms. This model is the stochastic stabilized supralinear network (SSSN) [44] generalized to have inhibitory multiplicity

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}). \tag{4}$$

As contrast increases, input to the E- and P-populations increases relative to a baseline input $\mathbf{h}_b$ via $\mathbf{h}_c$

$$\mathbf{h} = \mathbf{h}_b + c\mathbf{h}_c, \tag{5}$$

where $h_{c,E}, h_{c,P} > 0$ and $h_{c,S}, h_{c,V} = 0$. In this analysis, we fixed W,$\mathbf{h}_b$, and $\mathbf{h}_c$ to values fit to mean contrast responses in mice with the deterministic model [45] ($\boldsymbol{\epsilon} = \mathbf{0}$, Fig. 2B, see Section 5.2.2). At all contrasts, the E-population of this SSSN is unstable without recurrent inhibitory feedback. At 50% contrast, only the P-population exhibits the paradoxical effect (2C, Fig. 9), so the network is P-stabilized.

The slow noise of the SSSN is an an Ornstein-Uhlenbeck process

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}}\sigma_\alpha dB, \tag{6}$$

parameterized by σ_α, which can be different for each neuron type,

$$\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top. \tag{7}$$

For this SSSN, we are interested in the parameters of slow noise that produce realistic stochastic fluctuations. Here, we quantify this emergent property as having an excitatory population Fano factor near 1:

$$\begin{aligned} \mathcal{X} \quad : \quad \mathbb{E}_{\mathbf{z}}\left[F_E(\mathbf{x};\mathbf{z})\right] &= 1 \\ \text{Var}_{\mathbf{z}}\left[F_E(\mathbf{x};\mathbf{z})\right] &= 0.125^2, \end{aligned} \tag{8}$$

where $F_\alpha(\mathbf{x};\mathbf{z})$ is the Fano factor of the α-population.

We ran EPI to obtain a posterior $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$, where each parameter $\mathbf{z}$ produces biologically realistic levels of E-population variability (Fig. 2D). From the marginal distribution of σ_E and σ_P (Fig.

2D, top-left), we can see that $F_E(\mathbf{x};\mathbf{z})$ is sensitive to the combination of σ_E and σ_P. In fact, the posterior obtained through EPI offers exactly how this sensitivity changes along this ridge of the posterior (Fig. 2E). σ_S and σ_V are degenerate with respect to $F_E(\mathbf{x};\mathbf{z})$ evidenced by the uniform distribution in those dimensions of the posterior (Fig. 2D, bottom-right). Together, this posterior indicates a parametric manifold of degeneracy with respect to Fano factor: the ridge visualized in the σ_E-σ_P marginal (Fig. 10) and the dimensions of σ_S and σ_V.

Greater σ_E and σ_P confer greater Fano factor, and the Fano factors of each neuron-type are strongly correlated across the posterior (Fig 2F), showing that Fano factor of each neuron-type can be modulated globally via σ_E and σ_P. Furthermore, across the entire posterior distribution of noise parameterizations, we find that when contrast is increased above 50%, variability is quenched for all neuron types (Fig 2G). In summary, we used EPI to obtain a posterior of SSSNs producing realistic Fano factors, which allowed degenerate manifold identification via sample visualization, fast sensitivity measurements via Hessian evaluation, and predictions of variability quenching.

3.4 EPI identifies neural mechanisms of flexible task switching

In a rapid task switching experiment [46], rats were explicitly cued on each trial to either orient towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A) task (Fig. 3A). Neural recordings in the midbrain superior colliculus (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): the Pro/Contra and Anti/Ipsi neurons [27]. Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is a four-population dynamical system. We analyzed this model, where the neuron-type populations are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right (R)), their connectivity is parameterized geometrically (Fig. 3B). The input-output function of this model is chosen such that the population responses $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$ are bounded from 0 to 1 as a function ϕ of a dynamically evolving internal variable $\mathbf{u}$. The model responds to the side with greater Pro neuron activation; e.g. the reponse is left if $x_{LP} > x_{RP}$ at the end of the trial. The dynamics evolve with timescale $\tau = 0.09$ governed by connectivity weights W

$$\begin{aligned} \tau \frac{d\mathbf{u}}{dt} &= -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B} \\ \mathbf{x} &= \phi(\mathbf{u}) \end{aligned} \tag{9}$$

with white noise of variance 0.2^2. The input $\mathbf{h}$ is comprised of a cue-dependent input to the Pro or Anti populations, a stimulus orientation input to either the Left or Right populations, and a choice-period input to the entire network (see Section 5.2.3). Here, we use EPI to determine the changes in network connectivity $\mathbf{z} = [sW, vW, dW, hW]^\top$ resulting in execution of rapid task switching behavior.

We define rapid task switching behavior as accurate execution of each task. Inferred models should not exhibit fully random responses (50%), or perfect performance (100%), since perfection is never attained by even the best trained rats. We formulate rapid task switching as an emergent property by stipulating that the average accuracy in the Pro task $p_P(\mathbf{x}, \mathbf{z})$ and Anti task $p_A(\mathbf{x}, \mathbf{z})$ be 75% with variance $5\%^2$.

$$\begin{aligned} \mathcal{X} \quad : \quad \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x};\mathbf{z}) \\ p_A(\mathbf{x};\mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 75\% \\ 75\% \end{bmatrix} \\ \mathrm{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x};\mathbf{z}) \\ p_A(\mathbf{x};\mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 5\%^2 \\ 5\%^2 \end{bmatrix} \end{aligned} \tag{10}$$

A variance of $5\%^2$ performance in each task will confer a posterior producing performances ranging from about $65\% - 85\%$, allowing us to examine the properties of connectivity that yield better performance.

We ran EPI to obtain SC model connectivity parameters $\mathbf{z}$ producing rapid task switching (Fig. 3C). Some parameters were predictive of accuracy while others were not (Fig. 11), and often had different effects on p_P and p_A. To make sense of this inferred distribution, we took the eigendecomposition of the symmetric connectivity matrices $W = V\Lambda V^{-1}$, which results in the same basis vectors $\mathbf{v}_i$ for all W parameterized by $\mathbf{z}$ (Fig. 12A). These basis vectors have intuitive roles in processing for this task, and are accordingly named the *all* mode - all neurons co-fluctuate, *side* mode - one side dominates the other, *task* mode - the Pro or Anti populations dominate the other, and *diag* mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair.

Greater λ_{task}, λ_{side}, and λ_{diag} all produce greater Pro accuracy. This shows that strong task representations and hemispheral dominance in the dynamics result in better execution of the Pro task. By visualizing these four variables together by p_A (Fig. 13B), we see that low λ_{task} and λ_{diag} producing strong Anti accuracy also have high λ_{side} and λ_{all}. Thus, stronger hemispheral dominance, relaxed task and diag mode dynamics, and slower circuit-wide decay result in greater Anti accuracy.

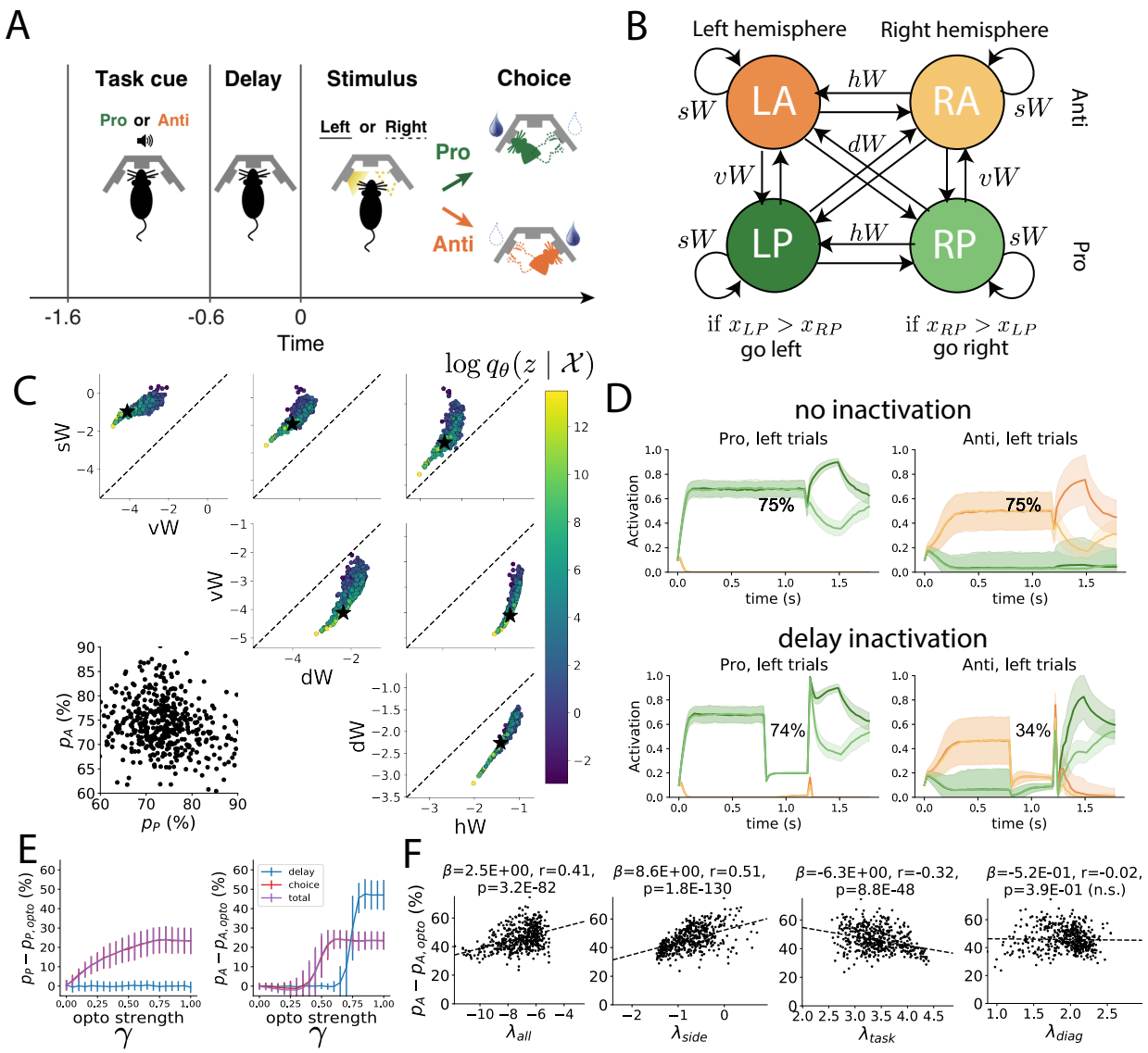

Figure 3: **A**. Rapid task switching behavioral paradigm (see text). **B**. Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW -vertical, dW - diagonal weights. Subscripts P and A of connectivity weights indicate Pro or Anti populations. **C**. The EPI parameter distribution of rapid task switching networks. Black star indicates parameter choice of simulations (D). **D**. Simulations of an SC network from the EPI distribution with 75% accuracy in each task. Top row shows no inactivation during Pro and Anti trials, and bottom row shows simulations with delay period inactivation (optogenetic strength $\gamma = 0.7$). Shading indicates standard deviation across trials. **E**. Difference in performance of each task during inactivation. Inactivation level "opto strength" scales from no inactivation (0) to full inactivation (1). We compare delay period inactivation $1.2 < t < 1.5$ (blue), choice period inactivation $1.5 < t$ (red), and total inactivation $0 \leq t \leq 1.8$ (purple). **F**. The effect of delay period inactivation on Anti accuracy versus dynamics eigenvalues.

In agreement with experimental results form Duan et al., we found that inactivation above nominal strength during the delay period consistently decreased performance in the Anti task, but had no consistent effect on the Pro task (Fig. 3E) e.g. (Fig. 3D, bottom). This difference in resiliency across tasks to delay perturbation is a prediction made by the inferred EPI distribution, rather than an emergent property that was conditioned upon. Even though p_P and p_A are anticorrelated in the EPI posterior ($r = -0.15$, $p = 3.68 \times 10^{-}12$), greater p_P and p_A both result in decreased resiliency to delay perturbation in the Anti task (Fig. 14). Ultimately, lower λ_{side} and λ_{all} and greater λ_{task} produce networks more robust to delay perturbation in the Anti task (Fig. 3F)).

In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task switching. This posterior revealed the mechanisms leading to greater accuarcy in each task as well as those increasing resiliency to perturbation in the Anti task. Importantly, every connectivity from this inferred distribution predicts fragility and robustness of performance in the Anti and Pro tasks, respectively. EPI allows us to conclude that since *all* parameters of this model producing rapid task switching make such an experimentally verified prediction, we have a well chosen model.

3.5 EPI scales well to high-dimensional parameter spaces

Here, we are interested in the scalability of EPI in number of parameters ($|\mathbf{z}|$). We consider rank-2 RNN with N neurons of connectivity

$$W = UV^\top + g\chi \tag{11}$$

and dynamics

$$\tau\dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x} \tag{12}$$

where $U = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix}$, $V = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2\end{bmatrix}$, $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2 \in [-1, 1]^N$, and $g = 0.01$.

We want to learn distributions of connectivity that produce stable amplification. Two conditions are both necessary and sufficient for RNNs to exhibit stable amplification [**?**]. These conditions are inequalities on real(λ_1) and λ_1^s the maximal real eigenvalue of W and the maximum eigenvalue of $W^s = \frac{W+W^\top}{2}$, respectively.

In our analysis, we seek to condition rank-2 networks of increasing size on a regime of stable amplification. Networks with real(λ_1) $= 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$ will yield moderate amplification.

EPI can naturally condition on this emergent property

$$\begin{aligned} \mathcal{X} \quad : \quad \mathbb{E}_{\mathbf{z},\mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z},\mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}. \end{aligned} \tag{13}$$

In contrast, SNPE cannot condition on the variance of observations across posterior. ' Thus, we condition on an observation x_0 located at the mean of our desired emergent property.

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \tag{14}$$

ABC methods define tolerance ϵ and distance for observed data x_0. Here, we chose $\epsilon = 0.5$, an $l-2$ distance, and the same choice for x_0 as in Equation 14.

EPI is capable of scaling to higher dimensional parameter spaces than ABC and SNPE. EPI consistently produces the same posterior predictive distribution independent of the dimensionality. SMC produces a limited variety of parameters due to the nature of its proposal generation algorithm, yet all parameters obtained produce stable amplification. SNPE's posterior predictive distribution is not necessarily close to the conditioning point, and is very dependent on dimensionality.

4 Discussion

NOTE: **This is the old discussion section. I will rewrite this based on our discussion of the rest of the draft.**

In neuroscience, machine learning has primarily been used to reveal structure in large-scale neural datasets [50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60] (see review, [16]). Such 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 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 focused on careful mechanistic modeling and the production of emergent properties of computation. The careful steps of *i.*) model design and *ii.*) emergent property definition, are followed by *iii.*) practical inference methods resulting in an opaque characterization of the way model parameters govern computation. In this work, we replaced this opaque procedure of parameter identification

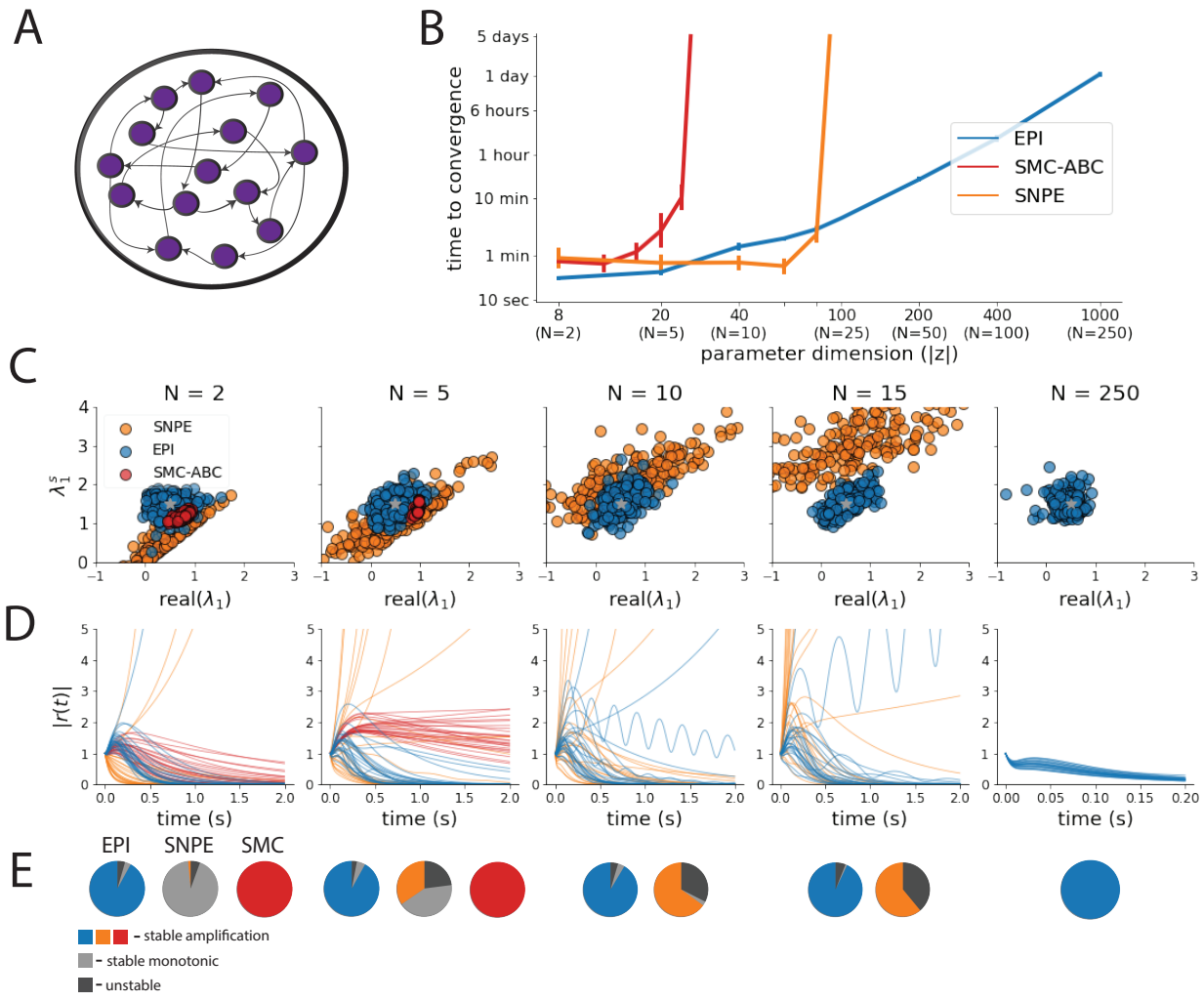

Figure 4: **A**. Recurrent neural network. **B**. EPI scales with z to high dimensions. Convergence definitions: EPI (blue) - satisfies all moment constraints, SNPE (orange)- produces at least $2/n_{\text{train}}$ parameter samples are in the bounds of emergent property (mean +- 0.5), and SMC-ABC (red) - 100 particles with $\epsilon <$ 0.5 are produced. **C**. Posterior predictive distributions of EPI (blue), SNPE (orange), and SMC-ABC (red). Gray star indicates emergent property mean, and gray dashed lines indicate two standard deviations corresponding to the variance constraint. For $N <= 6$ where SMC-ABC converges, samples are not diverse (path degeneracies). For $N >= 25$, SNPE does not produce a posterior approximation yielding parameters with simulations near x_0. **D**. Simulations of network parameters resulting from each method ($\tau = 100ms$). Each trace corresponds to simulation of one z. **E**. Ratio of obtained samples producing stable amplification.

in theoretical neuroscience with emergent property inference, opening the door to careful inference in careful models of neural computation.

Biologically realistic models of neural circuits often prove formidable to analyze. Two main factors contribute to the difficulty of this endeavor. First, in most neural circuit models, the number of parameters scales quadratically with the number of neurons, limiting analysis of its parameter space. Second, even in low dimensional circuits, the structure of the parametric regimes governing emergent properties is intricate. For example, these circuit models can support more than one steady state [61] and non-trivial dynamics on strange attractors [62].

In Section 3.3, we advanced the tractability of low-dimensional neural circuit models by showing that EPI offers insights about cell-type specific input-responsivity that cannot be afforded through the available linear analytical methods [26, 42, 43]. By flexibly conditioning this V1 model on different emergent properties, we performed an exploratory analysis of a *model* rather than a dataset, generating a set of testable hypotheses, which were proved out. Furthermore, exploratory analyses can be directed towards formulating hypotheses of a specific form. For example, model parameter dependencies on behavioral performance can be assessed by using EPI to condition on various levels of task accuracy (See Section 3.4). This analysis identified experimentally testable predictions (proved out *in-silico*) of patterns of effective connectivity in SC that should be correlated with increased performance.

In our final analysis, we presented a novel procedure for doing statistical inference on interpretable parameterizations of RNNs executing simple tasks. Specifically, we analyzed RNNs solving a posterior conditioning problem in the spirit of [63, 64]. This methodology relies on recently extended theory of responses in random neural networks with low-rank structure [49]. While we focused on rank-1 RNNs, which were sufficient for solving this task, this inference procedure generalizes to RNNs of greater rank necessary for more complex tasks. The ability to apply the probabilistic model selection toolkit to RNNs should prove invaluable as their use in neuroscience increases.

EPI leverages deep learning technology for neuroscientific inquiry in a categorically different way than approaches focused on training neural networks to execute behavioral tasks [65]. These works focus on examining optimized deep neural networks while considering the objective function, learning rule, and architecture used. This endeavor efficiently obtains sets of parameters that can be reasoned about with respect to such considerations, but lacks the careful probabilistic treatment of parameter inference in EPI. These approaches can be used complementarily to enhance the practice of theoretical neuroscience.

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 Postdoctoral 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 analysed during the current study are available from the corresponding author upon reasonable request.

Code availability statement:

The software written for the current study is available from the corresponding author upon reasonable request.

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[8] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. *International Conference on Machine Learning*, 2014.

[9] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical neural population models through nonlinear embeddings. In *Advances in neural information processing systems*, pages 163–171, 2016.

[10] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.

[11] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders. *bioRxiv*, page 399246, 2018.

[12] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, page 1, 2018.

[13] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

[14] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

[15] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall, Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet: nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural Information Processing Systems*, 2019.

[16] 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.

[17] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

[18] David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. 2014.

[19] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.

[20] 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.

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

[22] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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

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

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

[31] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

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

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

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing Wang. Paradoxical response reversal of top- down modulation in cortical circuits with three interneuron types. *Elife*, 6:1–15, 2017.

[43] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying the response of mouse cortical networks to optogenetic manipulation. 2019.

[44] 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.

[45] 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.

[46] 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.

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

[48] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.

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

[50] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*, 13(8):1713–1720, 2001.

[51] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–7425, 1998.

[52] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.

[53] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

[54] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan Segev. A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.

[55] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.

[56] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In *Advances in neural information processing systems*, pages 1692–1700, 2011.

[57] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

[58] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in visual path integration. *Neuron*, 99(1):194–206, 2018.

[59] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[60] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic. Inferring and validating mechanistic models of neural microcircuits based on spike-train data. *Nature Communications*, 10(4933), 2019.

[61] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of Sciences*, 115(13):3464–3469, 2018.

[62] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emergent dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint arXiv:1605.04463*, 2016.

[63] Xaq Pitkow and Dora E Angelaki. Inference in the brain: statistics flowing in redundant population codes. *Neuron*, 94(5):943–953, 2017.

[64] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*, page 696088, 2019.

[65] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 2019.

[66] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[67] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

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

[75] 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.

[76] Yazan N Billeh, Binghuang Cai, Sergey L Gratiy, Kael Dai, Ramakrishnan Iyer, Nathan W Gouwens, Reza Abbasi-Asl, Xiaoxuan Jia, Joshua H Siegle, Shawn R Olsen, et al. Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *bioRxiv*, page 662189, 2019.

[77] Nicolas Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of computational neuroscience*, 8(3):183–208, 2000.

[78] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science*, 304(5667):78–80, 2004.

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

5 Methods

5.1 Emergent property inference (EPI)

Consider model parameterization $\mathbf{z}$ and data $\mathbf{x}$ which has an intractable likelihood $p(\mathbf{x} \mid \mathbf{z})$ defined by a model simulator of which samples are available $\mathbf{x} \sim p(\mathbf{x} \mid \mathbf{z})$. EPI optimizes a distribution $q_{\boldsymbol{\theta}}(\mathbf{z})$ (itself parameterized by $\boldsymbol{\theta}$) of model parameters $\mathbf{z}$ to produce an emergent property of interest $\mathcal{X}$ defined by the means and variances of emergent property statistics $f(\mathbf{x}; \mathbf{z})$

$$\mathcal{X} : \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x}; \mathbf{z})\right] = \boldsymbol{\mu}, \mathrm{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x}; \mathbf{z})\right] = \boldsymbol{\sigma}^2. \tag{15}$$

Precisely, the emergent property statistics $f(\mathbf{x})$ must have means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ over the EPI distribution of parameters $q_{\boldsymbol{\theta}}(\mathbf{z})$ and stochasticity of the data given the parameters defined by the model $p(\mathbf{x} \mid \mathbf{z})$. This is a viable way to represent emergent properties in theoretical models, as we have demonstrated in the main text, and enables the EPI optimization.

With EPI, we use deep probability distributions to learn flexible approximations to model parameter distributions $q_{\boldsymbol{\theta}}(\mathbf{z})$. In deep probability distributions, a simple random variable $\mathbf{z}_0 \sim q_0(\mathbf{z}_0)$ 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:

$$\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0) = g_l(..g_1(\mathbf{z}_0)) \sim q_{\boldsymbol{\theta}}(\mathbf{z}). \tag{16}$$

Given a simulator defined by a theoretical model $\mathbf{x} \sim p(\mathbf{x} \mid \mathbf{z})$ and some emergent property of interest $\mathcal{X}$, $q_{\boldsymbol{\theta}}(\mathbf{z})$ 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:

$$\begin{aligned} q^*_{\boldsymbol{\theta}}(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in Q}{\mathrm{argmax}}\, H(q_{\boldsymbol{\theta}}(\mathbf{z})) \\ \text{s.t. } \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x}; \mathbf{z})\right] = \boldsymbol{\mu}, \mathrm{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x}; \mathbf{z})\right] = \boldsymbol{\sigma}^2. \end{aligned} \tag{17}$$

Since we are optimizing parameters $\boldsymbol{\theta}$ of our deep probability distribution with respect to the entropy $H(q_{\boldsymbol{\theta}}(\mathbf{z}))$, we must take gradients with respect to the log probability density of samples from the deep probability distribution. Entropy of $q_{\boldsymbol{\theta}}(\mathbf{z})$ can be expressed as an expectation of the negative log density of parameter samples $\mathbf{z}$ over the randomness in the parameterless initial distribution q_0:

$$H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \int -q_{\boldsymbol{\theta}}(\mathbf{z}) \log(q_{\boldsymbol{\theta}}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}}\left[-\log(q_{\boldsymbol{\theta}}(\mathbf{z}))\right] = \mathbb{E}_{\mathbf{z}_0 \sim q_0}\left[-\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))\right]. \tag{18}$$

Thus, the gradient of the entropy of the deep probability distribution can be estimated as an average of gradients of the log density of samples $\mathbf{z}$:

$$\nabla_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0 \sim q_0} \left[-\nabla_{\boldsymbol{\theta}} \log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0))) \right]. \tag{19}$$

In EPI, MEFNs are purposed towards variational learning of model parameter distributions.

5.1.1 Related work

TODO: rewrite this whole section.

A closely related methodology, variational inference, uses optimization to approximate posterior distributions [66]. Standard methods like stochastic gradient variational Bayes [7] or black box variational inference [67] simply do not work for inference in theoretical models of neural circuits, since they require tractable likelihoods $p(\mathbf{x} \mid \mathbf{z})$. Work on likelihood-free variational inference (LFVI) [20], which like EPI seeks to do inference in models with intractable likelihoods, employs an addtional deep neural network as a ratio estimator, enabling an estimation of the optimization objective for variational inference. Like LFVI, EPI can be framed as variational inference (see Section 5.1.4). But, unlike LFVI, EPI uses a single deep network to learn a distribution and is optimized to produce an emergent property, rather than condition on data points. Optimizing the EPI objective is a technological challenge, the details of which we elaborate in Section 5.1.3. Before going through those details, we ground this optimization in a toy example. We note that, during our preparation and early presentation of this work [68, 69], another work has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural circuits ([70, 71, 72], preprint posted simultaneously with this preprint). 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 [24], theirs similarly [73]). These distinct methodologies and scientific investigations emphasize the increased importance and timeliness of both works.

5.1.2 Normalizing flows

Deep probability distributions are comprised of multiple layers of fully connected neural networks. 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 $\mathbf{z}_i = g_i(\mathbf{z_{i-1}})$,

$$p(\mathbf{z}_i) = p(g_i^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial g_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} \right| = p(\mathbf{z}_{i-1}) \left| \det \frac{\partial g_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right|^{-1}. \tag{20}$$

However, this computation has cubic complexity in dimensionality for fully connected layers. By restricting our layers to normalizing flows [21] – bijective functions with fast log determinant Jacobian computations, we can tractably optimize deep generative models with objectives that are a function of sample density, like entropy. TODO: (clean up) We use Real NVP because it's a coupling architecture, which is fast to run either forwards (probability with samples) and backwards (prrobability or hessian). Normalizing flow architectures for deep probability distributions used in EPI are specified by the number of masks, neural network layers per mask, units per layer, and batch normalization momentum parameter.

5.1.3 Augmented Lagrangian optimization

To optimize $q_{\boldsymbol{\theta}}(\mathbf{z})$ in Equation 17, the constrained optimization is executed using the augmented Lagrangian method. The following objective is minimized:

$$L(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{opt}}, c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^{\top} R(\boldsymbol{\theta}) + \frac{c}{2} ||R(\boldsymbol{\theta})||^2 \tag{21}$$

where $R(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}} \left[\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\mathbf{z})} \left[T(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_{\text{opt}} \right] \right]$, $\boldsymbol{\eta}_{\text{opt}} \in \mathbb{R}^m$ are the Lagrange multipliers where $m = |\boldsymbol{\mu}_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})|$, and c is the penalty coefficient. These Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of exponential families (see Section 5.1.4). Deep neural network weights and biases $\boldsymbol{\theta}$ of the deep probability distribution are optimized according to Equation 21 using the Adam optimizer with its standard parameterization [74]. $\boldsymbol{\eta}_{\text{opt}}$ is initialized to the zero vector 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. For example, the initial value of c was $c_0 = 10^{-3}$ during EPI with the oscillating 2D LDS (Fig. S1C). The penalty coefficient is updated based on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of $\mathbb{E}[||R(\boldsymbol{\theta}_{k+1})||] > \gamma \mathbb{E}[||R(\boldsymbol{\theta}_k)||]$ is computed, and c_{k+1} is updated to βc_k with probability $1-p$. The other update rule is $\boldsymbol{\eta}_{\text{opt},k+1} = \boldsymbol{\eta}_{\text{opt},k} + c_k \frac{1}{n} \sum_{i=1}^{n} (T(\mathbf{x}^{(i)}) - \boldsymbol{\mu})$ given a batch size n. Throughout the study, $\beta = 4.0$, $\gamma = 0.25$, and the batch size was a hyperparameter, which varied according to the application of EPI.

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. In the oscillating 2D LDS example, each augmented Lagrangian epoch ran for 2,000 iterations (Fig. S1C-D). Notice the initial entropic growth, and subsequent reduction upon each update of $\boldsymbol{\eta}_{\text{opt}}$ and c. The momentum parameters of the Adam optimizer were 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 Gaussain 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 5.2.2).

To assess whether EPI distribution $q_{\boldsymbol{\theta}}(\mathbf{z})$ produces the emergent property, we defined a hypothesis testing convergence criteria. The algorithm has 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 from k resamplings of $\mathbf{z}$ from a finite sample of size n taken at the end of the augmented Lagrangian epoch. k is determined by a fraction of the batch size ν, which varies according to the application. In the linear two-dimensional system example, we used a batch size of $n = 1000$ and set $\nu = 0.1$ resulting in convergence after the ninth epoch of optimization. (Fig. S1C-D black dotted line).

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 5.1.6) 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 5.2.1). In such cases, optimization will be expensive.

5.1.4 Maximum entropy distributions and exponential families

Maximum entropy distributions have a fundamental link to exponential family distributions. A maximum entropy distribution of form:

$$\begin{aligned} p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}}\, H(p(\mathbf{z})) \\ \text{s.t. } \mathbb{E}_{\mathbf{z} \sim p}\left[T(\mathbf{z})\right] = \boldsymbol{\mu}_{\text{opt}}. \end{aligned} \tag{22}$$

will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \tag{23}$$

The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are formally hard to identify [75].

In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equation 2

$$\mathbb{E}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \boldsymbol{\sigma}^2. \tag{24}$$

The variance constraint is a second moment constraint on $f(\mathbf{x};\mathbf{z})$

$$\text{Var}_{\mathbf{z},\mathbf{x}}\left[f(\mathbf{x};\mathbf{z})\right] = \mathbb{E}_{\mathbf{z},\mathbf{x}}\left[\left(f(\mathbf{x};\mathbf{z}) - \boldsymbol{\mu}\right)^2\right] \tag{25}$$

As a general maximum entropy distribution (Equation 22), the sufficient statistics vector contains both first and second order moments of $f(\mathbf{x};\mathbf{z})$

$$T(\mathbf{x};\mathbf{z}) = \begin{bmatrix} f(\mathbf{x};\mathbf{z}) \\ \left(f(\mathbf{x};\mathbf{z}) - \boldsymbol{\mu}\right)^2 \end{bmatrix}, \tag{26}$$

which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma}^2 \end{bmatrix}. \tag{27}$$

5.1.5 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 $p(z \mid \mathbf{x})$, 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 [66], are utilized.

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

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

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

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

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

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)

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

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 26). Thus, EPI is implicitly executing variational inference with a uniform prior and a mean-field gaussian likelihood on the emergent property statistics. The data $\mathbf{x}$ used by this implicit variational inference program would be that generated by the adapting variational approximation $\mathbf{x} \sim p(\mathbf{x} \mid \mathbf{z}) q_{\boldsymbol{\theta}}(\mathbf{z})$, and the likelihood parameters $\boldsymbol{\eta}_f$ of EPI optimization epoch k are predicated by $\boldsymbol{\eta}_{\text{opt},k}$. However, in EPI we have not specified a prior distribution, or collected data, which can inform us about model parameters. Instead we have a mathematical specification of an emergent property, which the model must produce, and a maximum entropy selection principle. 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).

5.1.6 Example: 2D LDS

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

$$\tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \tag{39}$$

with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \tag{40}$$

To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fixing $\tau = 1$), the emergent property statistics $T(\mathbf{x})$ were chosen to contain the first and second moments of the oscillatory frequency, $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, real(λ_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 real(λ_1) having mean zero with variance 0.25^2, and the oscillation frequency 2πimag(λ_1) having mean $\omega = 1$ Hz with variance $(0.1\text{Hz})^2$:

$$\mathbb{E}\left[T(\mathbf{x})\right] \triangleq \mathbb{E}\begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}. \tag{41}$$

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:

$$\lambda = \frac{(\frac{a_1+a_4}{\tau}) \pm \sqrt{(\frac{a_1+a_4}{\tau})^2 + 4(\frac{a_2a_3-a_1a_4}{\tau})}}{2}. \tag{42}$$

Importantly, even though $\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\mathbf{z})}\left[T(\mathbf{x})\right]$ 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 [75]. 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.

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4-a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases}. \tag{43}$$

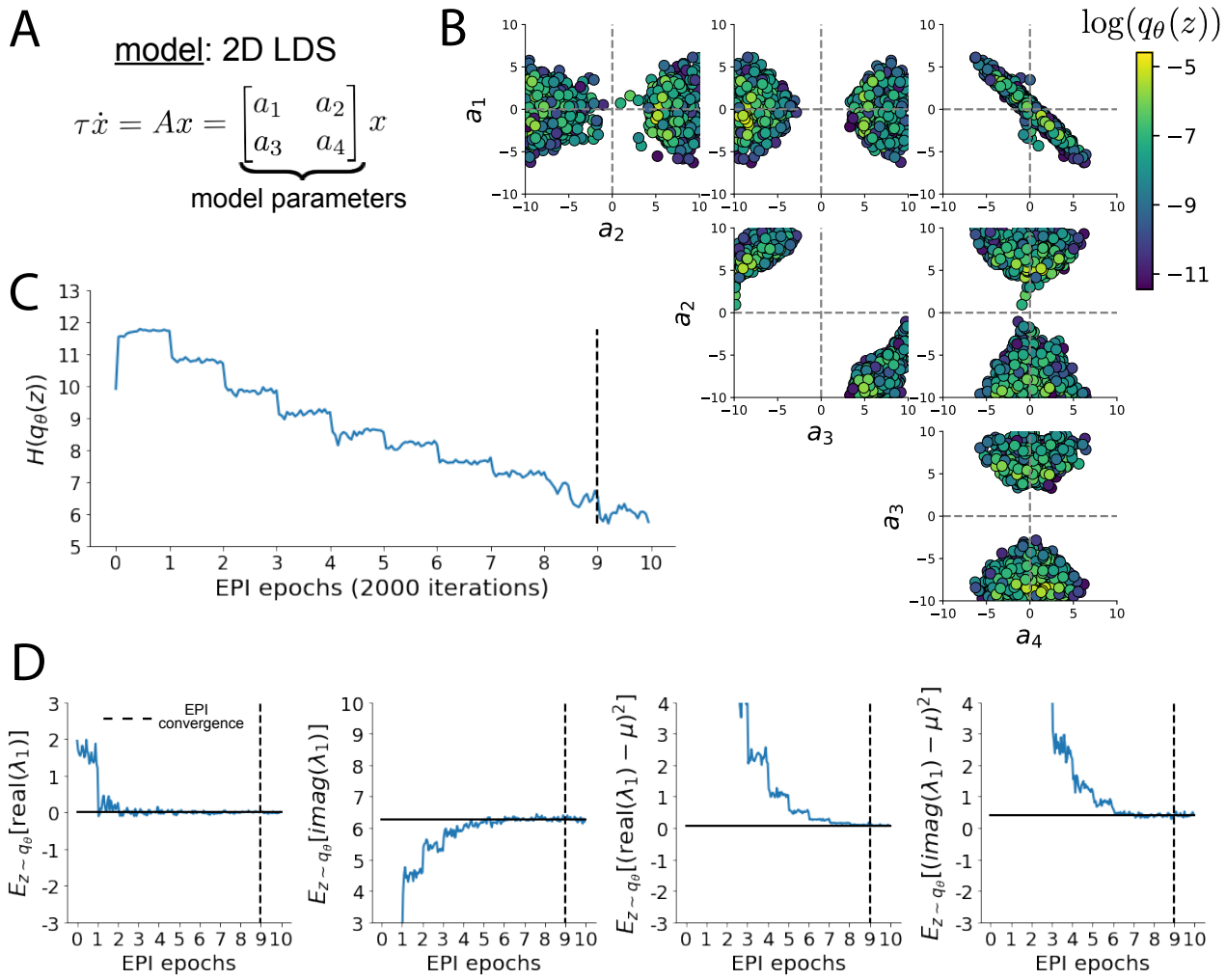

Figure 5: (LDS1): A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed 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.

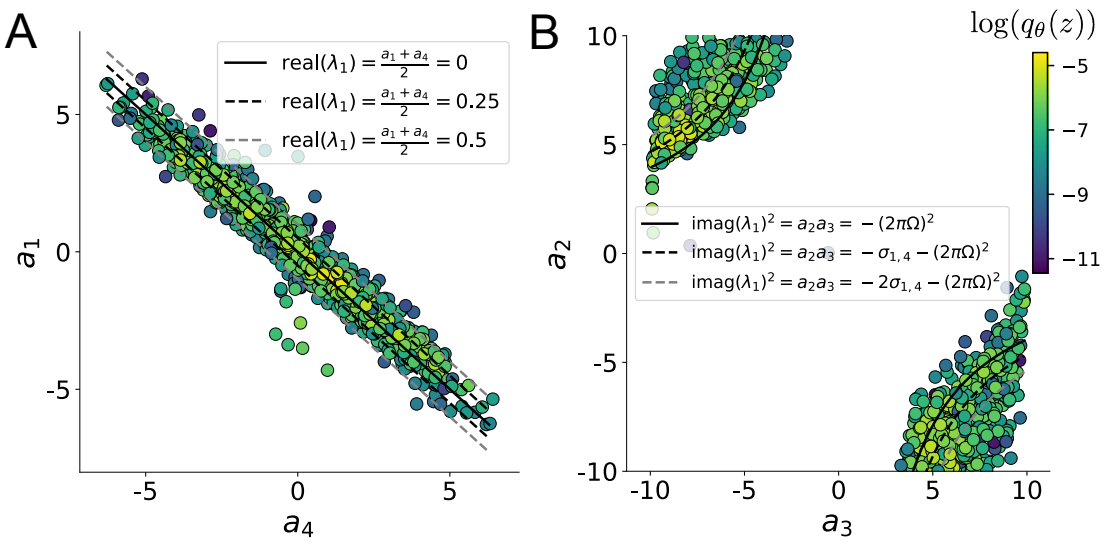

Figure 6: (LDS2): A. Probability contours in the a_1-a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor real(λ_1). B. Probability contours in the a_2-a_3 plane were derived from the emergent property statistic of oscillation frequency 2πimag(λ_1).

When $\tau = 1$ and $a_1a_4 > a_2a_3$ (center of distribution above), we have the following equation for the other two dimensions:

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \tag{44}$$

Since we constrained $\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}[\text{imag}(\lambda)] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\theta}}}(|a_1a_4 - E_{q_{\boldsymbol{\theta}}}[a_1a_4]|)$, we show the contours as $a_1a_4 = 0$ (black), $a_1a_4 = -\sigma_{1,4}$ (black dotted), and $a_1a_4 = -2\sigma_{1,4}$ (grey dotted). This validates the curved structure of the inferred distribution learned through EPI. We took steps in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive values a_1a_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).

5.2 Theoretical models

In this study, we used emergent property inference to examine several models relevant to theoretical neuroscience. Here, we provide the details of each model and the related analyses.

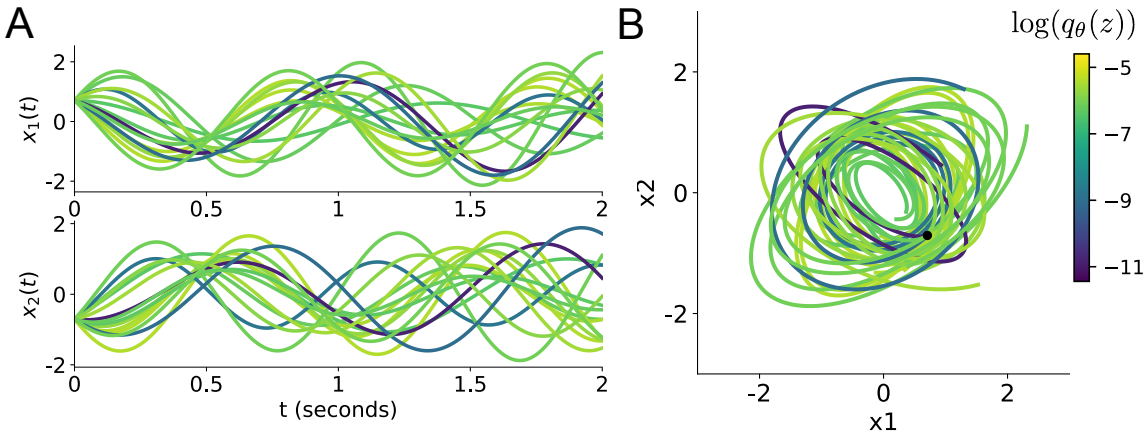

Figure 7: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\boldsymbol{\theta}}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. A. Each dimension of the simulated trajectories throughout time. B The simulated trajectories in phase space.

5.2.1 Stomatogastric ganglion

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) [30] 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:

$$C_m \frac{dx_\alpha}{dt} = -\left[h_{leak}(\mathbf{x};\mathbf{z}) + h_{Ca}(\mathbf{x};\mathbf{z}) + h_K(\mathbf{x};\mathbf{z}) + h_{hyp}(\mathbf{x};\mathbf{z}) + h_{elec}(\mathbf{x};\mathbf{z}) + h_{syn}(\mathbf{x};\mathbf{z})\right] + dB. \quad (45)$$

The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization, electrical and synaptic currents as well as gaussian noise dB. Each current component is a function of all membrane potentials and the conductance parameters $\mathbf{z}$.

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:

$$h_{leak}(\mathbf{x};\mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (46)$$

$$h_{elec}(\mathbf{x};\mathbf{z}) = g_{\text{el}}(x_\alpha^{post} - x_\alpha^{pre}) \quad (47)$$

$$h_{syn}(\mathbf{x};\mathbf{z}) = g_{syn} S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (48)$$

$$h_{Ca}(\mathbf{x};\mathbf{z}) = g_{Ca} M_\infty (x_\alpha - V_{Ca}) \tag{49}$$

$$h_K(\mathbf{x};\mathbf{z}) = g_K N (x_\alpha - V_K) \tag{50}$$

$$h_{hyp}(\mathbf{x};\mathbf{z}) = g_h H (x_\alpha - V_{hyp}). \tag{51}$$

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:

$$M_\infty = 0.5\left(1 + \tanh\left(\frac{x_\alpha - v_1}{v_2}\right)\right) \tag{52}$$

$$\frac{dN}{dt} = \lambda_N (N_\infty - N) \tag{53}$$

$$N_\infty = 0.5\left(1 + \tanh\left(\frac{x_\alpha - v_3}{v_4}\right)\right) \tag{54}$$

$$\lambda_N = \phi_N \cosh\left(\frac{x_\alpha - v_3}{2v_4}\right) \tag{55}$$

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \tag{56}$$

$$H_\infty = \frac{1}{1 + \exp\left(\frac{x_\alpha + v_5}{v_6}\right)} \tag{57}$$

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp\left(\frac{-x_\alpha + v_7}{v_8}\right)}\right). \tag{58}$$

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:

$$S_\infty = \frac{1}{1 + \exp\left(\frac{v_{th} - x_\alpha}{v_9}\right)}. \tag{59}$$

When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear dynamical system. Gaussian noise of variance $(1 \times 10^{-12})^2$ amps makes the model stochastic, and introduces variability in frequency at each parameterization $\mathbf{z}$.

In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for $T = 300$ time steps of $dt = 25ms$. 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

$$\Phi = [0.0, 0.01, ..., 1.0]^\top .. \tag{60}$$

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 $\alpha \in \{\text{f1}, \text{f2}, \text{hub}, \text{s1}, \text{s2}\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta |X_\alpha| \odot i), \tag{61}$$

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

$$\omega_\alpha = 0.01 \psi_\alpha \text{Hz}. \tag{62}$$

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, and variance was chosen to be $(0.025\text{Hz})^2$ to capture variation in frequency between 0.5Hz and 0.6Hz (Equation 2). 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 26 and 27)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \tag{63}$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \tag{64}$$

Throughout optimization, the augmented Lagrangian parameters η and c, were updated after each epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

For EPI in Fig 1E, we used a real NVP architecture with three coupling layers of affine transformations parameterized by two-layer neural networks of 25 units per layer. The initial distribution was

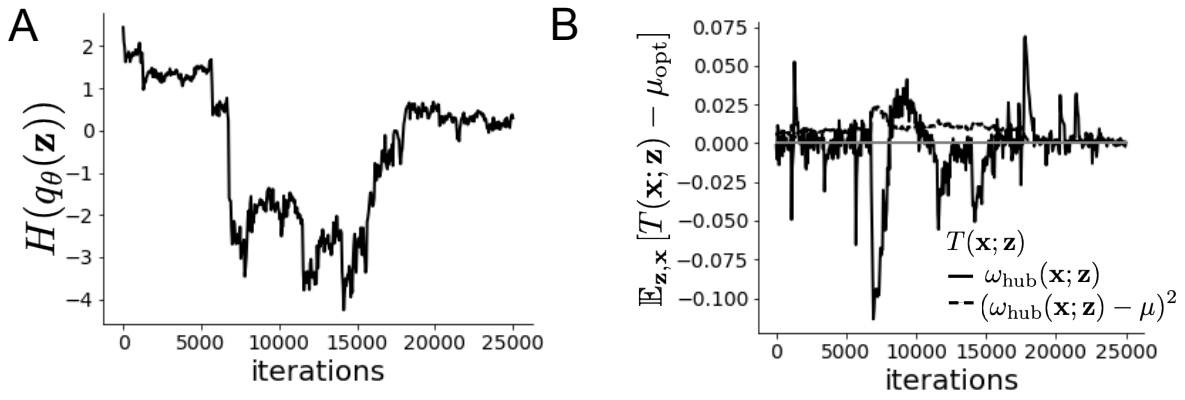

Figure 8: (STG1): EPI optimization of the STG model producing network syncing. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

a standard isotropic gaussian $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to a support of $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}] \in [4, 8] \times [0.01, 4]$. We did not include $g_{\text{synA}} < 0.01$, since conductances that low make the circuit simulations numerically unstable. We used an augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, set $\nu = 0.25$, and initialized $q_{\boldsymbol{\theta}}(\mathbf{z})$ to produce a gaussian approximation to samples returned from an initial ABC search. This initialization had much greater entropy and a different emergent property than the the returned EPI posterior.

TODO write about specifics of the Hessian analysis.

5.2.2 Primary visual cortex

Connectivity (W_{fit}) and input ($\mathbf{h}_{b,\text{fit}}$ and $\mathbf{h}_{c,\text{fit}}$) parameters were fit using the deterministic V1 circuit model [45]

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \tag{65}$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \tag{66}$$

and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \tag{67}$$

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

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \tag{68}$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \tag{69}$$

and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \tag{70}$$

Since the E-population of this network increases exponentially in the abscense of recurrent inhibitory feedback, we may also observe a paradoxical effect in the inhibitory populations (which is present in E-I networks). At 50% contrast (Fig. 2B, dots), this network exhibits a paradoxical effect in the P-population (Fig. 2C), but no others (Fig. 9). That is, for a small increase in h_P, $\mathbb{E}_t[x_P]$ decreases.

Fano factor is calculated as the temporal variance divided by the temporal mean following some time t_{ss} following dynamical evolution from the initial state at $\mathbf{x}(t = 0)$.

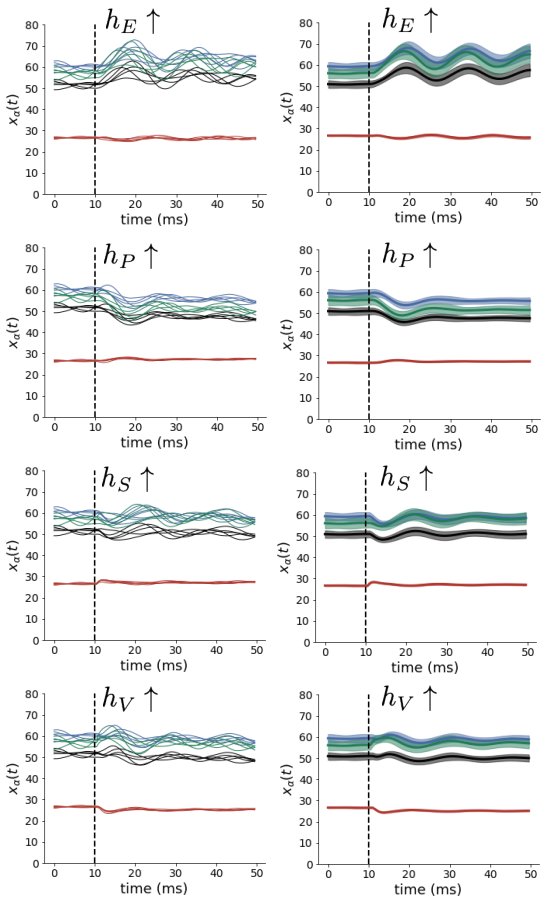

Figure 9: Supplemental Figure: (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average and standard deviation of stochastic fluctuations of responses.

5.2.3 Superior colliculus

In the model of Duan et al [27], there are four total units: two in each hemisphere corresponding to the Pro/Contra and Anti/Ipsi populations. They 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

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2}\tanh\left(\frac{u_\alpha - a}{b}\right) + \frac{1}{2}\right) \tag{71}$$

where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlinearity, respectively. During periods of optogenetic inactivation, activity was decreased proportional

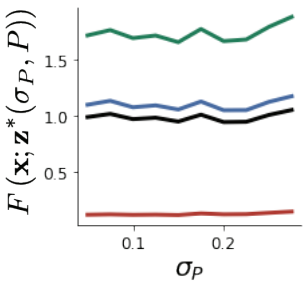

Figure 10: Supplemental Figure: Fano factors along the ridge of the posterior in Fig. 2E.

to the optogenetic strength γ

$$x_\alpha = (1-\gamma)\phi(u_\alpha). \tag{72}$$

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

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \tag{73}$$

which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \tag{74}$$

with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2. The weight matrix has 4 parameters sW, vW, hW, and dW:

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \tag{75}$$

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

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \tag{76}$$

There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1,1,1,1]^\top, \tag{77}$$

a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]\top, \tag{78}$$

rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \tag{79}$$

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \tag{80}$$

a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \tag{81}$$

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

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases}. \tag{82}$$

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 $I_{\text{light}} = 0.5$.

The accuracies of p_P and p_A are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} \left[\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]\right] \tag{83}$$

and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} \left[\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]\right] \tag{84}$$

given that the stimulus is on the left side, where Θ is the Heaviside step function.

The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \tag{85}$$

where $\beta = 100$.

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 26 and 27)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p(\mathbf{x}; \mathbf{z})_P \\ p(\mathbf{x}; \mathbf{z})_A \\ (p(\mathbf{x}; \mathbf{z})_P - 75\%)^2 \\ (p(\mathbf{x}; \mathbf{z})_A - 75\%)^2 \end{bmatrix}, \tag{86}$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 5\%^2 \\ 5\%^2 \end{bmatrix}. \tag{87}$$

Throughout optimization, the augmented Lagrangian parameters η and c, were updated after each epoch of 2,000 iterations(see Section 5.1.3). The optimization converged after six epochs (Fig. 15).

For EPI in Fig. 3C, 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 $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ 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 $n = 100$, set $\nu = 0.5$, and initialized $q_{\boldsymbol{\theta}}(\mathbf{z})$ to produce an isotropic gaussian with mean 0 and variance 2.5^2. Accuracies were estimated over 200 trials of random gaussian noise, which was sampled independently for each drawn parameter $\mathbf{z}$ and each iteration of the EPI optimization.

5.2.4 Rank-2 RNN

Traditional approaches to likelihood-free inference – approximate Bayesian computation (ABC) methods – randomly sample parameters $\mathbf{z}$ until a suitable set is obtained. State-of-the-art ABC methods leverage sequential monte-carlo (SMC) sampling techniques to obtain parameter sets more efficiently. To obtain more parameter samples, SMC-ABC must be run from scratch again. ABC methods do not confer log probabilities of samples. Like EPI, sequential neural posterior estimation (SNPE) uses deep learning to produce flexible posterior approximations. Like traditional Bayesian inference methods, SNPE conditions directly on the statistics of data. This differs from EPI, where posteriors are conditioned on emergent properties (moment constraints on the posterior predictive distribution). Peculiarities of SNPE (density estimation approach, two deep networks) make scaling in $\mathbf{z}$ prohibitive.

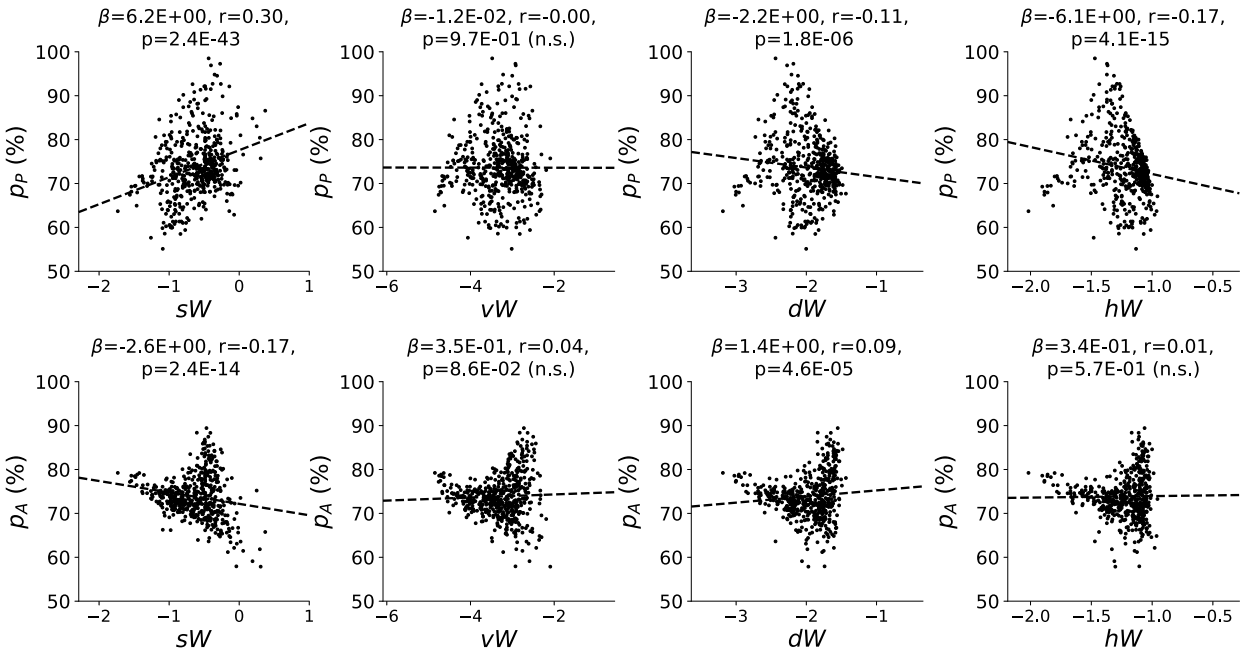

Figure 11: (SC1): Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p value.

SMC-ABC has many hyperparameters, of which pyABC selects automatically by running some initial diagnostics upon initialization. In concurrence with the literacture, SMC-ABC fails to converge around 25-30 dimensions, since it's proposal samples never get close enough to the target statistics. We searched over many SNPE hyperparameter choices: $n_{\text{train}} \in [2,000, 10,000, 100,000]$ is the number of simulations run per training epoch, and $n_{\text{mades}} \in [2, 3]$ is the number of masked autoregressive density estimators in the deep parameter distribution architecture. The greater n_{train}, the longer each epoch will take, but the more likely SNPE may converge during that epoch. Greater n_{mades} increases the flexibility of the deep parameter distribution of SNPE, but slows optimization. For the timing plot, we show the fastest among all of these choices, and for the convergence plot, we show the best convergence among all of these choices. During optimization, we used n_{atom}=100 atomic proposals as is recommended.

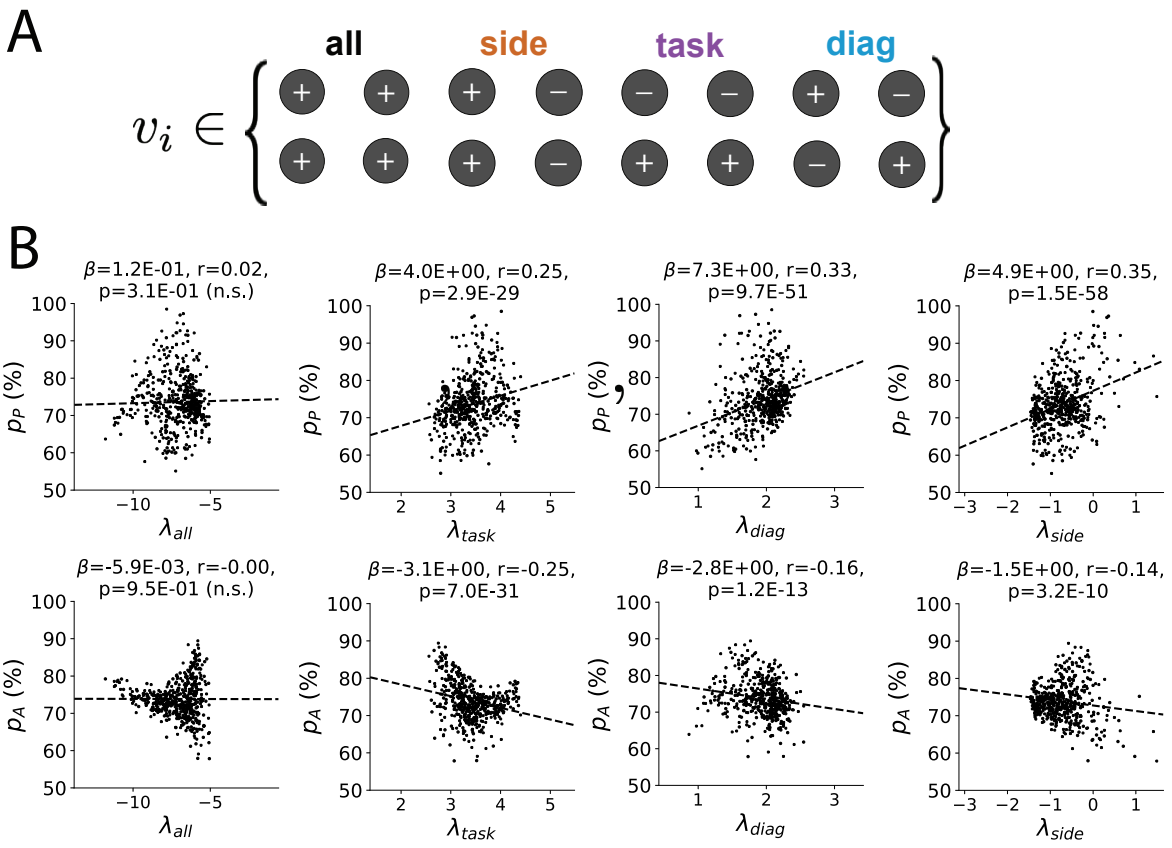

Figure 12: (SC2): A. Invariant eigenvectors of connectivity matrix W. B. Eigenvalues of connectivities of EPI distribution versus task accuracies.

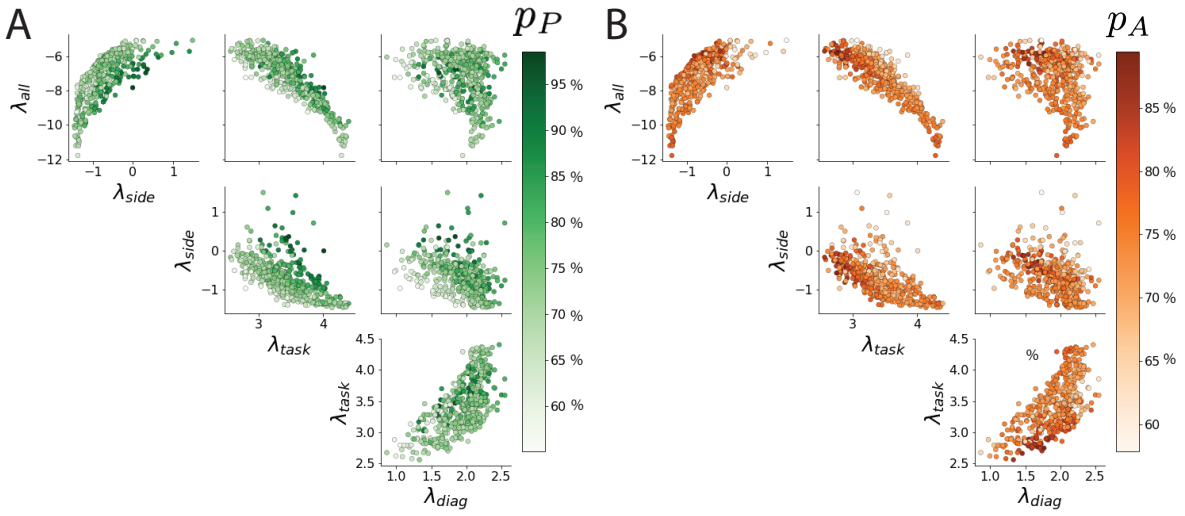

Figure 13: (SC3): A. Connectitivy eigenvalues of EPI parameter distribution colored by Pro task accuracy. B. Same for Anti task.

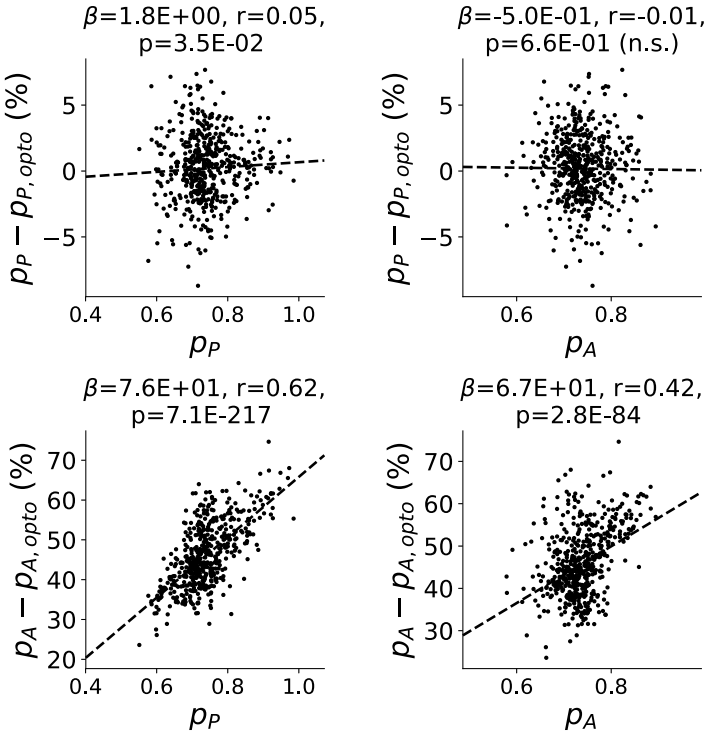

Figure 14: (SC4): Scatters of the effect of delay period inactivation in each task with task accuracy. Plots are shown at an opto strength of 0.8.

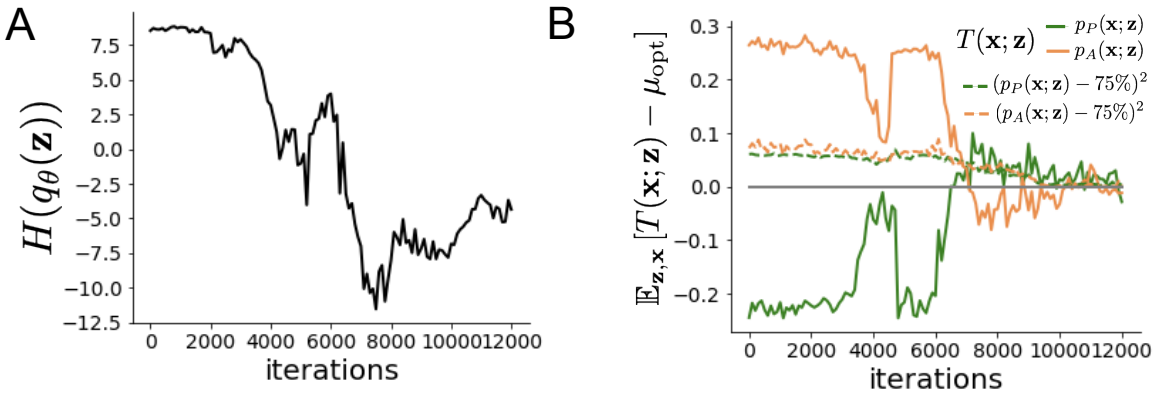

Figure 15: (SC5): EPI optimization of the SC model producing rapid task switching. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 12,000 iterations following the sixth augmented Lagrangian epoch.