

Interrogating theoretical models of neural computation with deep inference  
Sean R. Bittner<sup>1</sup>, Agostina Palmigiano<sup>1</sup>, Alex T. Piet<sup>2,3,4</sup>, Chunyu A. Duan<sup>5</sup>, Carlos D. Brody<sup>2,3,6</sup>,  
Kenneth D. Miller<sup>1</sup>, and John P. Cunningham<sup>7</sup>.

<sup>1</sup>Department of Neuroscience, Columbia University,

<sup>2</sup>Princeton Neuroscience Institute,

<sup>3</sup>Princeton University,

<sup>4</sup>Allen Institute for Brain Science,

<sup>5</sup>Institute of Neuroscience, Chinese Academy of Sciences,

<sup>6</sup>Howard Hughes Medical Institute,

<sup>7</sup>Department of Statistics, Columbia University

## <sup>1</sup> 1 Abstract

<sup>2</sup> A cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures  
<sup>3</sup> a hypothesized neural mechanism. Such models are valuable when they give rise to an experi-  
<sup>4</sup> mentally observed phenomenon – whether behavioral or a pattern of neural activity – and thus  
<sup>5</sup> can offer insights into neural computation. The operation of these mechanistic circuits, like all  
<sup>6</sup> models, critically depends on the choices of model parameters. A key process in circuit modeling  
<sup>7</sup> is then to identify the model parameters consistent with observed phenomena: to solve the inverse  
<sup>8</sup> problem. To solve challenging inverse problems modeling neural datasets, neuroscientists have used  
<sup>9</sup> statistical inference techniques to much success. However, most research in theoretical neuroscience  
<sup>10</sup> focuses on how computation emerges in biologically interpretable circuit models, and how the model  
<sup>11</sup> parameters govern computation; it is not focused on the latent structure of empirical models of  
<sup>12</sup> noisy experimental datasets. In this work, we present a novel technique that brings the power  
<sup>13</sup> and versatility of the probabilistic modeling toolkit to theoretical inverse problems. Our method  
<sup>14</sup> uses deep neural networks to learn parameter distributions with rich structure that have specific  
<sup>15</sup> computational properties in biologically relevant models. This methodology is explained through  
<sup>16</sup> a motivational example inferring conductance parameters in an STG subcircuit model. Then, with  
<sup>17</sup> RNNs of increasing size, we show that only EPI allows precise control over the behavior of inferred  
<sup>18</sup> parameters, and that EPI scales better in parameter dimension than alternative techniques. In the  
<sup>19</sup> remainder of this work, we explain novel theoretical insights through the examination of intricate  
<sup>20</sup> parametric structure in complex circuit models. In a model of primary visual cortex with multiple

21 neuron-types, where analysis becomes untenable with each additional neuron-type, we discovered  
22 how noise distributed across neuron-types governs the excitatory population. Finally, in a model  
23 of superior colliculus, we identified and characterized two distinct regimes of connectivity that  
24 facilitate switching between opposite tasks amidst interleaved trials. We also found that all task-  
25 switching connectivities in this model reproduce behaviors from inactivation experiments, further  
26 establishing this hypothesized circuit model. Beyond its scientific contribution, this work illustrates  
27 the variety of analyses possible once deep learning is harnessed towards solving theoretical inverse  
28 problems.

## 29 2 Introduction

30 The fundamental practice of theoretical neuroscience is to use a mathematical model to understand  
31 neural computation, whether that computation enables perception, action, or some intermediate  
32 processing. A neural circuit is systematized with a set of equations – the mechanistic model – and  
33 these equations are motivated by biophysics, neurophysiology, and other conceptual considerations  
34 [1–4]. The function of this system is governed by the choice of model *parameters*, which when  
35 configured in a particular way, give rise to a measurable signature of a computation. The work  
36 of analyzing a model then requires solving the inverse problem: given a computation of interest,  
37 how can we reason about particular parameter configurations? The inverse problem is crucial for  
38 reasoning about likely parameter values, uniquenesses and degeneracies, and predictions made by  
39 the model [5, 6].

40 Consider the idealized practice: one carefully designs a model and analytically derives how compu-  
41 tational properties determine model parameters. Seminal examples of this gold standard include  
42 our field’s understanding of memory capacity in associative neural networks [7], chaos and au-  
43 tocorrelation timescales in random neural networks [8], the paradoxical effect [9], and decision  
44 making [10]. Unfortunately, as circuit models include more biological realism, theory via analytical  
45 derivation becomes intractable. Still, we can gain insight into these complex models by identifying  
46 the distribution of parameters that produce computations. By solving the inverse problem in this  
47 way, scientific analysis of biologically realistic models is made possible [6, 11–14].

48 While theoretical neuroscience is concerned with how model parameters govern computational  
49 properties, existing methodology for statistical inference in neuroscience [15–36] (see review, [37])  
50 requires that parameters be conditioned on an explicit dataset. The scientific insight for a model

51 of computation is then limited by the quantity and quality of available neural data. Even with a  
52 vast amount of high-quality recordings, neural data often reflect uninstructed behaviors [38–40],  
53 and thus may only reflect the computation of interest amidst a sea of task-irrelevant factors. A  
54 common alternative is to synthesize an explicit dataset that is exemplary of that computation, so  
55 that the framework of statistical inference can be applied for parameter identification. In this case,  
56 well-defined computational properties are being shoehorned into artificial datasets for the purpose  
57 of methodological compatibility.

58 Another key challenge is that as models of computation become more complex, statistical inference  
59 becomes intractable. Such mechanistic models in theoretical neuroscience are noisy systems of  
60 differential equations that can only be sampled or realized through forward simulation [41, 42];  
61 they lack a tractable likelihood function, which is necessary for statistical inference. Therefore, the  
62 most popular approaches to parameter inference in mechanistic models have been simulation-based  
63 inference methods [43, 44], in which reasonable parameters are obtained via simulation and rejection.  
64 A new class of techniques [45–47] use deep learning to improve upon traditional simulation-based  
65 inference approaches. However, to use these methods in theoretical neuroscience, we must represent  
66 computation with an explicit dataset in some way. Theorists are therefore barred from using the  
67 probabilistic modeling toolkit for science with circuit models, unless they reformulate their inverse  
68 problem into a framework for observational datasets.

69 To address the methodological incongruity between explicit datasets and emergent properties, we  
70 present a statistical inference method for conditioning parameters of neural circuit models directly  
71 on computation. In this work, we define computation by an emergent property, which is a statistical  
72 description of the phenomena to be produced by the neural circuit model. In emergent property  
73 inference (EPI), we infer the distribution of model parameters that produce this emergent property.  
74 With EPI, parameters are conditioned directly on an implicit dataset defined by the computation  
75 of interest. By using recent optimization techniques [48], EPI uses deep learning to make rich,  
76 flexible approximations to the parameter distributions [49], the structure of which reveals scientific  
77 insight about how parameters govern the emergent property.

78 Equipped with this method, we prove out the potential of EPI by demonstrating its capabilities and  
79 presenting novel theoretical findings borne from its analysis. First, we show EPI’s ability to handle  
80 mechanistic models using a classic model of parametric degeneracy in biology: the stomatogastric  
81 ganglion [50, 51]. Then, we show EPI’s scalability to high dimensional parameter distributions by  
82 inferring connectivities of recurrent neural networks (RNNs) that exhibit stable, yet amplified re-

sponses – a hallmark of neural responses throughout the brain [52–54]. In a model of primary visual cortex (V1) [55, 56] with different neuron-types, we show that the equation for excitatory variability become analytically intractable as more populations are added. Strikingly, the way in which noisy inputs across neuron-types governs excitatory variability is salient in the visualized structure of the EPI inferred parameter distribution. Finally, we investigated the possible connectivities of superior colliculus (SC) that allow execution of different tasks on interleaved trials [57]. EPI discovered a rich distribution containing two connectivity regimes with different solution classes. We queried the deep probability distribution learned by EPI to produce a mechanistic understanding of cortical responses in each regime. Intriguingly, all inferred connectivities reproduced results from optogenetic inactivation experiments in this behavioral paradigm – emergent phenomena that EPI was not conditioned upon. These theoretical insights afforded by EPI illustrate the value of deep inference for the interrogation of neural circuit models.

### 3 Results

#### 3.1 Motivating emergent property inference of theoretical models

Consideration of the typical workflow of theoretical modeling clarifies the need for emergent property inference. First, one designs or chooses an existing 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 [58]. Despite full knowledge of STG connectivity and a precise characterization of its rhythmic pattern generation, biophysical models of the STG have complicated relationships between circuit parameters and computation [12, 50]. A subcircuit model of the STG [51] is shown schematically in Figure 1A. The jagged connections indicate electrical coupling having electrical conductance  $g_{el}$ , smooth connections in the diagram are inhibitory synaptic projections having strength  $g_{synA}$  onto the hub neuron, and  $g_{synB} = 5nS$  for mutual inhibitory connections. Note that the behavior of this model will be critically dependent on its parameterization – the choices of conductance parameters  $\mathbf{z} = [g_{el}, g_{synA}]$ . 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 must specify what the model should produce. In this STG

model, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model 1B. An emergent property studied by Guttierrez et al. of this stochastic model is the hub neuron firing at an intermediate frequency between the intrinsic spiking rates of the fast and slow populations. This emergent property is shown in Figure 1C at an average frequency of 0.55Hz. Our notion of intermediate hub frequency is not strictly 0.55Hz, but also moderate deviations of this frequency between the fast (.35Hz) and slow (.68Hz) frequencies, which are quantified in the emergent property with variance  $0.025^2\text{Hz}^2$ .

Third, the model parameters producing these outputs are inferred. To infer the STG parameters of intermediate hub frequency with existing methodology, we need an explicit dataset: experimentally recorded or synthesized. By precisely quantifying the emergent property of interest as a statistical feature of the model, we use EPI to condition directly on this emergent property. EPI learns a probability distribution of model parameters constrained to produce the emergent property. In this last step lies the opportunity for a shift away from a dataset-oriented representation of model output towards that of an implicit dataset, where the only structure is the emergent property of interest.

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 a parameter distribution that produces the emergent property.

### 3.2 A deep generative modeling approach to emergent property inference

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

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

where  $C_m=1\text{nF}$ , and  $\mathbf{h}$  is a sum of the leak, calcium, potassium, hyperpolarization, electrical, and synaptic currents, all of which have their own complicated dependence on activity  $\mathbf{x}$  and parameters  $\mathbf{z} = [g_{el}, g_{synA}]$ , and  $d\mathbf{B}$  is white gaussian noise [51, 59] (see Section 5.2.1 for more detail).



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

142 Second, we stipulate that our model should produce the emergent property of “intermediate hub  
 143 frequency” (Figure 1C). We stipulate that the hub neuron’s spiking frequency – denoted  $\omega_{\text{hub}}(\mathbf{x})$   
 144 is close to a frequency of 0.55Hz, between that of the slow and fast frequencies. Mathematically,  
 145 we define this emergent property with two statistical constraints: that the mean hub frequency is  
 146 0.55Hz,

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

147 and that the variance of the hub frequency is moderate

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

148 The hub neuron frequency is constrained over the distribution of parameters  $\mathbf{z}$  and the distribution  
 149 of the data  $\mathbf{x}$  that those parameters produce. Formally, the emergent property is the collection of  
 150 these two constraints

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [\omega_{\text{hub}}(\mathbf{x}; \mathbf{z})] = 0.55, \quad \text{Var}_{\mathbf{z}, \mathbf{x}} [\omega_{\text{hub}}(\mathbf{x}; \mathbf{z})] = 0.025^2. \quad (4)$$

151 In general, an emergent property is a collection of first-, second- and higher moments of statistics  
 152 that together define the phenomena.

153 Third, we perform emergent property inference: we find a distribution over parameter configura-  
 154 tions  $\mathbf{z}$  that produces the emergent property; in other words, they obey the constraints intro-  
 155 duced in Equation 4. This distribution will be chosen from a family of probability distributions  
 156  $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$ , defined by a deep neural network [49, 60, 61] (Figure 1D, EPI box). Deep  
 157 probability distributions map a simple random variable  $\mathbf{z}_0$  through a deep neural network with  
 158 weights and biases  $\boldsymbol{\theta}$  to parameters  $\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0)$  to a suitably complicated distribution (see Section  
 159 5.1.2 for more details). Many distributions in  $\mathcal{Q}$  will respect the emergent property constraints,  
 160 so we select the most random (highest entropy) distribution, which is the same choice made in  
 161 Bayesian posterior inference (see Section 5.1.6). In EPI optimization, stochastic gradient steps in  
 162  $\boldsymbol{\theta}$  are taken such that entropy is maximized, and the emergent property  $\mathcal{X}$  is produced (see Section  
 163 5.1) The inferred EPI distribution is denoted  $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ , since it is conditioned upon emergent  
 164 property  $\mathcal{X}$ . This is meant to share the same notation as a posterior distribution  $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$  that is  
 165 conditioned upon an explicit dataset.

166 EPI produces parameter distributions that can be queried for scientific insight. The modes of  
 167  $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$  indicate parameter choices exemplary of the emergent property (Fig. 1E yellow star). As  
 168 probability in the EPI inferred distribution decreases, the emergent property deteriorates. Perturb-  
 169 ing  $\mathbf{z}$  along a dimension in which  $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$  does not change will not disturb the emergent property,

making this parameter combination *degenerate* with respect to the emergent property. In contrast, if  $\mathbf{z}$  is perturbed along a dimension that strongly decreases  $q_{\theta}(\mathbf{z} \mid \mathcal{X})$ , we call that parameter combination *sensitive*. By querying the second order derivative (Hessian) of  $\log q_{\theta}(\mathbf{z} \mid \mathcal{X})$  at a mode, we can quantitatively identify how sensitive (or robust) each eigenvector is by its eigenvalue; the more negative the eigenvalue, the more sensitive. Indeed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity ( $v_1$ , smaller eigenvalue) and robustness ( $v_2$ , greater eigenvalue) (Fig. 1E, arrows) agree with error contours (Fig. 1E contours) and have diminished or preserved hub frequency, respectively (Fig. 1F activity traces). Once an EPI distribution has been inferred, this Hessian calculation requires trivial computation (when the correct architecture class is chosen, see Section 5.1.2).

In the following sections, we demonstrate EPI on three neural circuit models across ranges of biological realism, neural system function, and network scale. First, we demonstrate the superior scalability of EPI compared to alternative techniques by inferring high-dimensional distributions of recurrent neural network (RNN) connectivities that exhibit amplified, yet stable responses. Also in this RNN example, we emphasize that EPI is the only technique that controls the predictions made by the inferred parameter distribution. Next, in a model of primary visual cortex [55,56], we show how EPI captures a curved manifold of parametric degeneracy, revealing how input variability across neuron types affects the excitatory population. Finally, in a model of superior colliculus [57], we used EPI to capture multiple parametric regimes of task switching, and queried the dimensions of sensitivity ( $\mathbf{v}_1(\mathbf{z})$ ) to mechanistically characterize each regime.

### 3.3 Scaling inference of RNN connectivity with EPI

Transient amplification is a hallmark of neural activity throughout cortex, and is often thought to be intrinsically generated by recurrent connectivity in the responding cortical area [52–54]. It has been shown that to generate such amplified, yet stabilized responses, the connectivity of RNNs must be non-normal [52,62], and satisfy additional constraints [63]. In theoretical neuroscience, RNNs are optimized and then examined to show how dynamical systems could execute a given computation [64,65], but such biologically realistic constraints on connectivity are ignored during optimization for practical reasons. In general, access to distributions of connectivity adhering to theoretical criteria like stable amplification, chaotic fluctuations [8], or low tangling [66] would add scientific value and context to existing research with RNNs. Here, we use EPI to learn RNN connectivities producing stable amplification, and demonstrate the superior scalability and efficiency of EPI to

201 alternative approaches.

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

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}, \quad (5)$$

203 where  $U = [\mathbf{u}_1 \ \mathbf{u}_2] + g\chi^{(U)}$ ,  $V = [\mathbf{v}_1 \ \mathbf{v}_2] + g\chi^{(V)}$ ,  $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2 \in [-1, 1]^N$ , and  $\chi_{i,j}^{(U)}, \chi_{i,j}^{(V)} \sim$   
204  $\mathcal{N}(0, 1)$ . We infer connectivity parameterizations  $\mathbf{z} = [\mathbf{u}_1^\top, \mathbf{u}_2^\top, \mathbf{v}_1^\top, \mathbf{v}_2^\top]^\top$  that produce stable amplification.  
205 Two conditions are necessary and sufficient for RNNs to exhibit stable amplification [63]:  
206  $\text{real}(\lambda_1) < 1$  and  $\lambda_1^s > 1$ , where  $\lambda_1$  is the eigenvalue of  $W$  with greatest real part and  $\lambda^s$  is the maximum eigenvalue of  $W^s = \frac{W+W^\top}{2}$ . RNNs with  $\text{real}(\lambda_1) = 0.5 \pm 0.5$  and  $\lambda_1^s = 1.5 \pm 0.5$  will be stable  
207 with modest decay rate ( $\text{real}(\lambda_1)$  close to its upper bound of 1) and exhibit modest amplification  
208 ( $\lambda_1^s$  close to its lower bound of 1). EPI can naturally condition on this emergent property  
209

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}, \end{aligned} \quad (6)$$

210 under the notion that variance constraints with standard deviation 0.25 predicate that the vast  
211 majority of samples (those within two standard deviations) are within the specified ranges.

212 For comparison, we infer the parameters  $\mathbf{z}$  likely to produce stable amplification using two al-  
213 ternative simulation-based inference approaches. We ran sequential Monte Carlo approximate  
214 Bayesian computation (SMC-ABC) [43] and sequential neural posterior estimation (SNPE) [45]  
215 with observation  $\mathbf{x}_0 = \boldsymbol{\mu}$ . SMC-ABC is a rejection sampling approach that SMC techniques to  
216 improve efficiency, and SNPE approximates posteriors with deep probability distributions using  
217 a two-network architecture (see Section 5.1.1). Unlike EPI, these statistical inference techniques  
218 do not constrain the statistics of the predictive distribution, and these predictions of the inferred  
219 posteriors are typically affected by model characteristics (e.g.  $N$  and  $g$ , Fig. 11). To compare the  
220 efficiency of these different techniques, we measured the time and number of simulations necessary  
221 for the distance of the predictive mean to be less than 0.5 from  $\boldsymbol{\mu} = \mathbf{x}_0$  (see Section 5.2.2).

222 As the number of neurons  $N$  in the RNN is scaled, and thus the dimension of the parameter  
223 space  $\mathbf{z} \in [-1, 1]^{4N}$ , we see that EPI converges at greater speed and at greater dimension than  
224 SMC-ABC and SNPE (Fig. 2A). It also becomes most efficient to use EPI in terms of simulation  
225 count at  $N = 50$  (Fig. 2B). It is well known that ABC techniques struggle in dimensions greater  
226 than about 30 [67], yet we were careful to assess the scalability of the more comparable approach



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

227 SNPE. Between EPI and SNPE, we closely controlled the number of parameters in deep probability  
228 distributions by dimensionality (Fig. 10), and tested more aggressive SNPE hyperparameterizations  
229 when SNPE failed to converge (Fig. 12). From this analysis, we see that deep inference techniques  
230 EPI and SNPE are far more amenable to inference of high dimensional parameter distributions than  
231 rejection sampling techniques like SMC-ABC, and that EPI outperforms SNPE in both criteria in  
232 high dimensions.

233 No matter the number of neurons, EPI always produces connectivity distributions with mean and  
234 variance of  $\text{real}(\lambda_1)$  and  $\lambda_1^s$  according to  $\mathcal{X}$  (Fig. 2C, blue). For the dimensionalities in which  
235 SMC-ABC is tractable, the inferred parameters are concentrated and offset from  $\mathbf{x}_0$  (Fig. 2C,  
236 green). When using SNPE the predictions of the inferred parameters are highly concentrated at  
237 some RNN sizes and widely varied in others (Fig. 2C, orange). We see these properties reflected in  
238 simulations from the inferred distributions: EPI produces a consistent variety of stable, amplified  
239 activity norms  $|r(t)|$ , SMC-ABC produces a limited variety of responses, and the changing variety  
240 of responses from SNPE emphasizes the control of EPI on parameter predictions.

241 EPI outperforms SNPE in high dimensions by using gradient information (from  $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z}) =$   
242  $\nabla_{\mathbf{z}} [\text{real}(\lambda_1), \lambda_1^s]^{\top}$ ) on each iteration of optimization. This agrees with recent speculation that such  
243 gradient information could improve the efficiency of simulation-based inference techniques [68].  
244 Since gradients of the emergent property statistics are necessary in EPI optimization, gradient  
245 tractability is a key criteria when determining the suitability of a simulation-based inference tech-  
246 nique. Evidenced by this analysis, EPI is a clear choice for inferring high dimensional parameter  
247 distributions when the emergent property gradient is efficiently calculated. This can be invaluable  
248 for understanding how RNNs produce complex emergent phenomena. Even with a high degree  
249 of biophysical realism and expensive emergent property gradients, EPI was run successfully on  
250 intermediate hub frequency in a 5-neuron subcircuit model of the STG (Section 3.1). However,  
251 conditioning on the pyloric rhythm [69] in a model of the pyloric subnetwork model [12] proved to  
252 be prohibitive with EPI. The pyloric subnetwork requires many time steps for simulation and many  
253 key emergent property statistics (e.g. burst duration and phase gap) are not calculable or easily  
254 approximated with differentiable functions. In such cases, gradient-free approaches like SNPE have  
255 proved to be a powerful option [45]. In the next two sections, we use EPI for novel scientific insight  
256 by examining the structure of inferred distributions.

257 **3.4 EPI reveals how noisy input across neuron-types governs excitatory vari-  
258 ability in a V1 model**

259 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output  
260 function have succeeded in explaining a host of experimentally documented phenomena. In a regime  
261 characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to  
262 paradoxical responses [9], selective amplification [52, 62], surround suppression [70] and normaliza-  
263 tion [71]. Despite their strong predictive power, E-I circuit models rely on the assumption that  
264 inhibition can be studied as an indivisible unit. However, experimental evidence shows that inhibi-  
265 tion is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) – composing  
266 80% of GABAergic interneurons in V1 [72–74], and that these inhibitory cell types follow specific  
267 connectivity patterns (Fig. 3A) [75]. While research has shown that V1 only shares specific dimen-  
268 sions of neuronal variability with downstream areas [76], the role played by recurrent dynamics and  
269 the connectivity across neuron-type populations is not understood. Here, in a model of V1 with  
270 biologically realistic connectivity, we use EPI to show how the structure of input across neuron  
271 types affects the variability of the excitatory population – the population largely responsible for  
272 projecting to other brain areas [77].

273 We considered response variability of a nonlinear dynamical V1 circuit model (Fig. 3A) with a  
274 state comprised of each neuron-type population’s rate  $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$ . Each population  
275 receives recurrent input  $W\mathbf{x}$ , where  $W$  is the effective connectivity estimated from post-synaptic  
276 potential and connectivity rate measurements (see Section 5.2.3). Each population also experiences  
277 an external input  $\mathbf{h}$ , which determines population rate via supralinear nonlinearity  $\phi(\cdot) = [\cdot]_+^2$ .  
278 There is also an additive noisy input  $\epsilon$  parameterized by variances for each neuron type population  
279  $\sigma^2 = [\sigma_E^2, \sigma_P^2, \sigma_S^2, \sigma_V^2]$ . This noise has a slower dynamical timescale  $\tau_{\text{noise}} > \tau$  than the population  
280 rate, allowing fluctuations around a stimulus-dependent steady-state (Fig. 3B). This model is the  
281 stochastic stabilized supralinear network (SSSN) [78]

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \epsilon). \quad (7)$$

282 generalized to have multiple inhibitory neuron types, and introduces stochasticity to previous four  
283 neuron-type models of V1 [55]. Stochasticity and inhibitory multiplicity introduce substantial  
284 complexity to mathematical derivations (see Section 5.2.4) motivating the treatment of this model  
285 with EPI. Here, we consider fixed weights  $W$  and input  $\mathbf{h}$  [56], and study the effect of input  
286 variability  $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$  on excitatory variability.

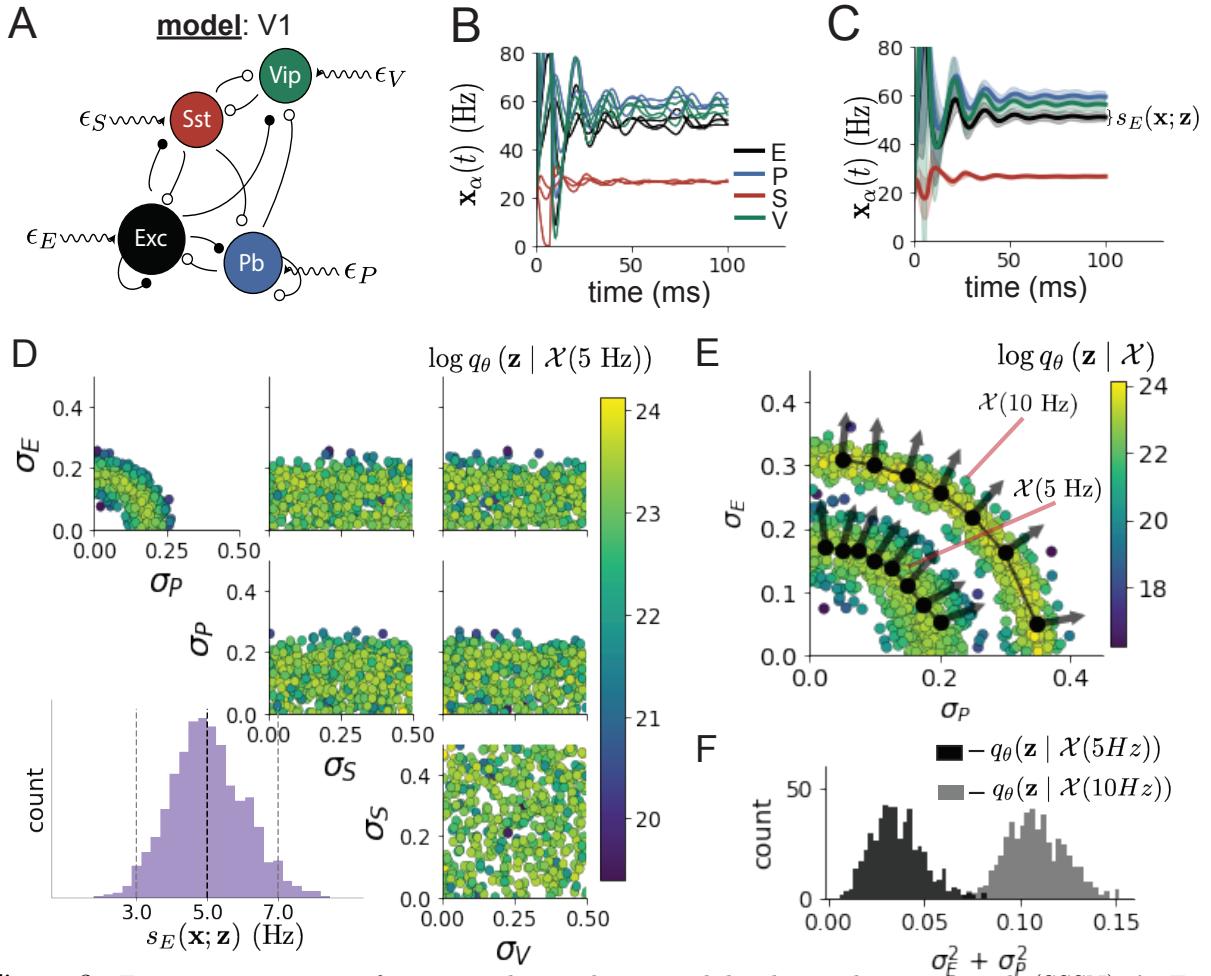


Figure 3: Emergent property inference in the stochastic stabilized supralinear network (SSSN). **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively).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

287 We quantify levels  $y$  of E-population variability with the emergent property

$$\begin{aligned}\mathcal{X}(y) &: \mathbb{E}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] = y \\ \text{Var}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= 1\text{Hz}^2,\end{aligned}\tag{8}$$

288 where  $s_E(\mathbf{x}; \mathbf{z})$  is the standard deviation of the stochastic  $E$ -population response about its steady  
289 state (Fig. 3C). In the following analyses, we compare levels of 5Hz and 10Hz, and select 1 Hz<sup>2</sup>  
290 variance such that the two emergent properties do not overlap in  $s_E(\mathbf{z}; \mathbf{x})$ .

291 First, we ran EPI to obtain parameter distribution  $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$  producing E-population vari-  
292 ability around 5 Hz (Fig. 3D). From the marginal distribution of  $\sigma_E$  and  $\sigma_P$  (Fig. 3D, top-left), we  
293 can see that  $s_E(\mathbf{x}; \mathbf{z})$  is sensitive to various combinations of  $\sigma_E$  and  $\sigma_P$ . Alternatively, both  $\sigma_S$  and  
294  $\sigma_V$  are degenerate with respect to  $s_E(\mathbf{x}; \mathbf{z})$  evidenced by the high variability in those dimensions  
295 (Fig. 3D, bottom-right). Together, these observations imply a curved path with respect to  $s_E(\mathbf{x}; \mathbf{z})$   
296 of 5 Hz, which is indicated by the modes along  $\sigma_P$  (Fig. 3E).

297 Figure 3E suggests a quadratic relationship in E-population fluctuations and the standard deviation  
298 of E- and P-population input; as the square of either  $\sigma_E$  or  $\sigma_P$  increases, the other compensatorily  
299 decreases to preserve the level of  $s_E(\mathbf{x}; \mathbf{z})$ . This quadratic relationship is preserved at greater level  
300 of E-population variability  $\mathcal{X}(10 \text{ Hz})$  (Fig. 3E). Indeed, the sum of squares of  $\sigma_E$  and  $\sigma_P$  is larger  
301 in  $q_{\theta}(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$  than  $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$  (Fig 3F,  $p < 1 \times 10^{-10}$ ), while the sum of squares of  
302  $\sigma_S$  and  $\sigma_V$  are not significantly different in the two EPI distributions (Fig. 15,  $p = .40$ ), in which  
303 parameters were bounded from 0 to 0.5. The strong interactive influence of E- and P-population  
304 input variability on excitatory variability is intriguing, since this circuit exhibits a paradoxical effect  
305 in the P-population (and no other inhibitory types) (Fig. 15), meaning that the E-population is  
306 P-stabilized. Future research may uncover a link between the population of network stabilization  
307 and compensatory interactions governing excitatory variability.

308 EPI revealed the quadratic dependence of excitatory variability on input variability to the E- and  
309 P-populations, as well as its independence to input from the other two inhibitory populations. We  
310 show that with each neuron-type population added to this E-I model, calculations of excitatory  
311 variability with respect to noise parameters become unruly and challenging to work with (see  
312 Section 5.2.4). This emphasizes the value of streamlined methods for gaining understanding about  
313 theoretical models when mathematical analysis becomes onerous or impractical. While EPI can  
314 be used to investigate fundamental aspects of sensory processing, in the next section, we use the  
315 probabilistic tools of EPI to identify and characterize two distinct parametric regimes of a neural  
316 circuit executing a computation, and then relate these insights to behavioral experiments.

317 **3.5 EPI identifies two regimes of rapid task switching**

318 It has been shown that rats can learn to switch from one behavioral task to the next on randomly  
 319 interleaved trials [79], and an important question is what types of neural connectivity allow this  
 320 ability. In this experimental setup, rats were explicitly cued on each trial to either orient towards  
 321 a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A) task  
 322 (Fig. 4A). Neural recordings in superior colliculus (SC) exhibited two populations of neurons that  
 323 represented task context (Pro or Anti). Furthermore, Pro/Anti neurons in each hemisphere were  
 324 strongly correlated with the animal’s decision [57]. These results motivated a model of SC that  
 325 is a four-population dynamical system with functionally-defined neuron-types. Here, our goal is  
 326 to understand how connectivity in this circuit model governs the ability to perform rapid task  
 327 switching: to respond with satisfactory accuracy in both tasks on randomly interleaved trials.

328 In this SC model, there are Pro- and Anti-populations in each hemisphere (left (L) and right  
 329 (R)) with activity variables  $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$ . The connectivity of these populations is  
 330 parameterized by self  $sW$ , vertical  $vW$ , diagonal  $dW$  and horizontal  $hW$  connections (Fig. 4B). The  
 331 input  $\mathbf{h}$  is comprised of a positive cue-dependent signal to the Pro or Anti populations, a positive  
 332 stimulus-dependent input to either the Left or Right populations, and a choice-period input to the  
 333 entire network (see Section 5.2.5). Model responses are bounded from 0 to 1 as a function  $\phi$  of an  
 334 internal variable  $\mathbf{u}$

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

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

338 Rapid task switching is formalized mathematically as an emergent property with two statistics:  
 339 accuracy in the Pro task  $p_P(\mathbf{x}; \mathbf{z})$  and Anti task  $p_A(\mathbf{x}; \mathbf{z})$ . We stipulate that accuracy be on average  
 340 .75 in each task with variance  $.075^2$

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .75 \\ .75 \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .075^2 \\ .075^2 \end{bmatrix}. \end{aligned} \tag{10}$$



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

341 75% accuracy is a realistic level of performance in each task, and with the chosen variance, inferred  
342 models will not exhibit fully random responses (50%), nor perfect performance (100%).

343 The EPI inferred distribution (Fig. 4C) produces Pro and Anti task accuracies (Fig. 4C, middle-  
344 left) consistent with rapid task switching (Equation 10). This parameter distribution has intricate  
345 structure, that is not captured well by simple linear correlations (Fig. 17). Specifically, the shape  
346 of the EPI distribution changes dramatically on different sides of parameter space. This is most  
347 saliently pointed out in the marginal distribution of  $sW-hW$  (Fig. 4C top-right), where anticorrela-  
348 tion between  $sW$  and  $hW$  switches to correlation with decreasing  $sW$ . Not only has EPI captured  
349 this complicated distribution of connectivities producing rapid task switching, we can query the  
350 EPI distribution  $q_{\theta}(\mathbf{z} | \mathcal{X})$  to understand these two parametric regimes of SC connectivity.

351 To distinguish these two regimes, we use the EPI distribution to identify two sets of modes. By  
352 fixing  $hW$  to different values and doing gradient ascent on  $\log q_{\theta}(\mathbf{z} | \mathcal{X})$ , we arrive at two solutions  
353  $\mathbf{z}^*(hW_{\text{fixed}}, r)$  where regime  $r \in [1, 2]$ , and regime 1 is that of greater  $sW$  (see Section 5.2.5). As  
354  $hW_{\text{fixed}}$  increases, the modes coalesce to intermediate parameters reflecting a transition between  
355 the two sets of modes (Fig. 20 top). By using EPI to connect these two regimes through this  
356 transitional region of parameter space, we can explore what distinguishes the two regimes by  
357 stepping from the prototypical connectivity of regime 1 to that of regime 2.

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

364 When perturbing connectivity along the sensitivity dimension, Pro accuracy monotonically in-  
365 creases in both regimes (Fig. 4D, top-left). However, there is a stark difference between regimes in  
366 Anti accuracy. Anti accuracy falls in either direction of  $\mathbf{v}_1$  in regime 1, yet monotonically increases  
367 along with Pro accuracy in regime 2 (Fig. 4D, top-right). These distinct pathologies of rapid task  
368 switching are caused by distinct connectivity changes ( $\mathbf{v}_1(\mathbf{z}^*(r = 1))$  vs  $\mathbf{v}_1(\mathbf{z}^*(r = 2))$ ) and explain  
369 the sharp change in local structure of the EPI distribution. With further perturbation analyses  
370 along dimensions of connectivity having established effects on processing, we are able to distinguish  
371 these two regimes with mechanistic insights (Fig. 4D, middle, bottom) (Section 5.2.5).

372 **3.6 EPI inferred SC connectivities reproduce results from optogenetic inacti-**  
373 **vation experiments**

374 During the delay period of this task, the circuit must prepare to execute the correct task based on  
375 the cue input. Experimental results from Duan et al. found that optogenetic inactivation of SC  
376 during the delay period consistently decreased performance in the Anti task, but had no effect on  
377 the Pro task (Fig. 5A). All network connectivities inferred by EPI exhibited this same effect, when  
378 network activities were silenced during the delay period (see Section 5.2.5). Notably, EPI inferred  
379 connectivities were only conditioned upon the emergent property of rapid task switching, not on  
380 Anti task failure during delay period silencing.

381 Following delay period inactivation, there are strong similarities in the responses of Pro and Anti  
382 trials during the choice period (Fig. 21A). We interpreted these similarities to suggest that delay  
383 period inactivation flips the internal representation of task in the model. Connectivity patterns  
384 inducing greater Pro task accuracy would antagonistically reduce accuracy in Anti trials subject to  
385 delay period inactivation in which this flip in task representation occurs. In fact, across connectiv-  
386 ities in the EPI inferred distribution, there was strong anti-correlation between Pro task accuracy  
387  $p_P$  and Anti task accuracy during delay period silencing  $p_{A,opto}$  (Fig. 5B).

388 We explored whether this antagonism between  $p_P$  and  $p_{A,opto}$  was present in each connectivity  
389 regime. As in the previous section, we tested perturbations in the dimension of sensitivity  $\mathbf{v}_1(\mathbf{z})$   
390 from regime 1 to regime 2. In both regimes, we recall that perturbations along  $\mathbf{v}_1(\mathbf{z})$  yield increasing  
391  $p_P$ . We found that only in regime 2, does  $p_{A,opto}$  decrease as  $p_P$  increases (Fig. 5C). This context  
392 further distinguishes the two regimes discovered with EPI.

393 In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task  
394 switching. This EPI distribution revealed two regimes of rapid task switching, which we char-  
395 acterized using the probabilistic toolkit afforded by EPI. We found that both of these parametric  
396 regimes identified by EPI reproduce results from optogenetic inactivation experiments: when activ-  
397 ity is silenced during the delay period, only Anti accuracy suffers. We then identified connectivity  
398 mechanisms governing Anti accuracy during delay period silencing, and showed the parametric  
399 regime they describe.

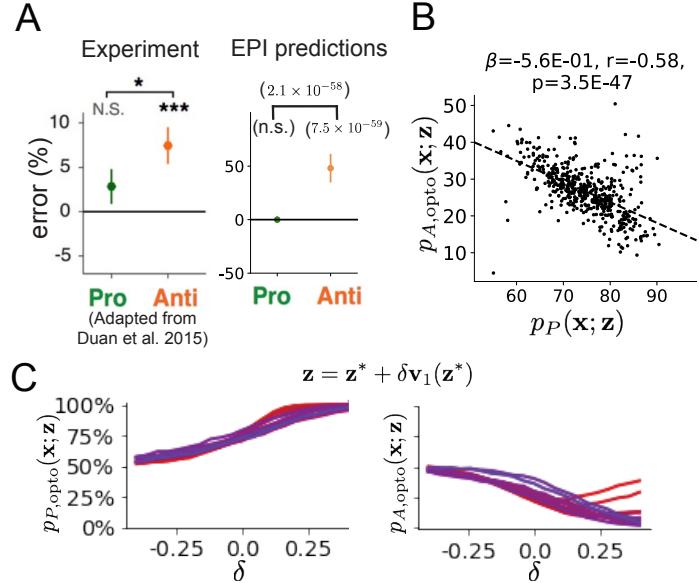


Figure 5: **A.** The EPI distribution predicts experimental results (left) showing no change in the Pro task, but larger error in the Anti task (right). **B.** Accuracy in the Anti task during delay period optogenetic inactivation  $p_{A,\text{opto}}$  is strongly anticorrelated with accuracy in the Pro task. **C.** Mean and standard error ( $N_{\text{test}} = 25$ ) of accuracy with respect to perturbation along the sensitivity dimension of each mode  $\mathbf{z}^*$ .

## 4 Discussion

In neuroscience, machine learning has primarily been used to reveal structure in neural datasets [37]. Careful inference procedures are developed for these statistical models allowing precise, quantitative reasoning, which clarifies the way data informs beliefs about the model parameters. However, these statistical models often lack resemblance to the underlying biology, making it unclear how to go from the structure revealed by these methods, to the neural mechanisms giving rise to it. In contrast, theoretical neuroscience has focused on careful mechanistic modeling and the production of emergent properties of computation, rather than measuring structure in some noisy observed dataset. In this work, we improve upon parameter inference techniques in theoretical neuroscience with emergent property inference, harnessing deep learning towards parameter inference with respect to emergent phenomena in interpretable models of neural computation (see Section 5.1.1).

Methodology for statistical inference in mechanistic models of neural circuits has evolved considerably in recent years. Early work used rejection sampling techniques [43, 80, 81], but more recently developed methodology employs deep learning to improve efficiency or provide flexible distribution approximations. SNPE [45] and other sequential techniques for inference in mechanistic models developed along with EPI (see Section 5.1.1) have been used for posterior inference with noisy

416 experimental datasets. On the other hand, EPI is a deep inference technique designed to condition  
417 directly on emergent properties, such that the parameter distribution only produces the computa-  
418 tion of interest. EPI is thus ideally suited for questions in theoretical neuroscience, and we show  
419 that it has superior scaling properties to these other inference techniques (see Section 3.3).

420 **Acknowledgements:**

421 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-  
422 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-  
423 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-  
424 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program  
425 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James  
426 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

427 **Data availability statement:**

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

430 **Code availability statement:**

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

432 **References**

- 433 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern  
434 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 435 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,  
436 21(1):25–45, 1998.
- 437 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 438 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in  
439 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 440 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R  
441 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology  
442 models. *PLoS Comput Biol*, 3(10):e189, 2007.

- 443 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network  
444 homeostasis, and pathological compensation from a biologically plausible ion channel expres-  
445 sion model. *Neuron*, 82(4):809–821, 2014.
- 446 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-  
447 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 448 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural  
449 networks. *Physical review letters*, 61(3):259, 1988.
- 450 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-  
451 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,  
452 17(11):4382–4388, 1997.
- 453 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration  
454 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 455 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley  
456 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 457 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit  
458 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 459 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron  
460 model. *PLoS computational biology*, 2(7):e94, 2006.
- 461 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar  
462 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 463 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,  
464 13(8):1713–1720, 2001.
- 465 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.  
466 A statistical paradigm for neural spike train decoding applied to position prediction from  
467 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–  
468 7425, 1998.
- 469 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding  
470 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.

- 471 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.  
472 A point process framework for relating neural spiking activity to spiking history, neural  
473 ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 474 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correla-  
475 tions imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–  
476 1012, 2006.
- 477 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan  
478 Segev. A novel multiple objective optimization framework for constraining conductance-based  
479 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 480 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature  
481 analysis. *Neural computation*, 19(4):1022–1038, 2007.
- 482 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and  
483 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of  
484 neural population activity. In *Advances in neural information processing systems*, pages  
485 1881–1888, 2009.
- 486 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and  
487 Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural  
488 information processing systems*, 24:1350–1358, 2011.
- 489 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In  
490 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 491 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-  
492 dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,  
493 9(3):e1002922, 2013.
- 494 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W  
495 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.  
496 *Science*, 349(6244):184–187, 2015.
- 497 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq  
498 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in  
499 visual path integration. *Neuron*, 99(1):194–206, 2018.

- 500 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable  
501 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- 502
- 503 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.  
504 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.  
505 *Nature Communications*, 10(4933), 2019.
- 506 [30] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical  
507 neural population models through nonlinear embeddings. In *Advances in neural information  
508 processing systems*, pages 163–171, 2016.
- 509 [31] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-  
510 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 511 [32] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-  
512 encoders. *bioRxiv*, page 399246, 2018.
- 513 [33] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,  
514 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R  
515 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-  
516 encoders. *Nature methods*, page 1, 2018.
- 517 [34] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M  
518 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.  
519 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 520 [35] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R  
521 Datta. Composing graphical models with neural networks for structured representations and  
522 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 523 [36] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,  
524 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:  
525 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural  
526 Information Processing Systems*, 2019.
- 527 [37] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-  
528 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,  
529 2018.

- 530 [38] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral  
531 state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.
- 532 [39] Aman B Saleem, Aslı Ayaz, Kathryn J Jeffery, Kenneth D Harris, and Matteo Carandini.  
533 Integration of visual motion and locomotion in mouse visual cortex. *Nature neuroscience*,  
534 16(12):1864–1869, 2013.
- 535 [40] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Church-  
536 land. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuro-  
537 science*, 22(10):1677–1686, 2019.
- 538 [41] Peter Dayan, Laurence F Abbott, et al. Theoretical neuroscience: computational and mathe-  
539 matical modeling of neural systems. *Journal of Cognitive Neuroscience*, 15(1):154–155, 2003.
- 540 [42] Eugene M Izhikevich. *Dynamical systems in neuroscience*. MIT press, 2007.
- 541 [43] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.  
542 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 543 [44] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.  
544 A framework for parameter estimation and model selection from experimental data in systems  
545 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 546 [45] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan  
547 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P  
548 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural  
549 dynamics. *bioRxiv*, page 838383, 2019.
- 550 [46] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast  
551 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on  
552 Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 553 [47] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized  
554 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–  
555 4248. PMLR, 2020.
- 556 [48] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow  
557 networks. *International Conference on Learning Representations*, 2017.

- 558 [49] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.  
559 *International Conference on Machine Learning*, 2015.
- 560 [50] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-  
561 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,  
562 2001.
- 563 [51] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an  
564 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.  
565 *Neuron*, 77(5):845–858, 2013.
- 566 [52] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of  
567 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 568 [53] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dy-  
569 namics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–  
570 1406, 2014.
- 571 [54] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding  
572 and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.
- 573 [55] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and  
574 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysi-  
575 ology*, 115(3):1399–1409, 2016.
- 576 [56] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel  
577 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the  
578 operating regime of cortex. *bioRxiv*, 2020.
- 579 [57] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J  
580 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor  
581 routing. *bioRxiv*, page 245613, 2018.
- 582 [58] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-  
583 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 584 [59] Catherine Morris and Harold Lecar. Voltage oscillations in the barnacle giant muscle fiber.  
585 *Biophysical journal*, 35(1):193–213, 1981.

- 586 [60] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
587 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 588 [61] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for  
589 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,  
590 2017.
- 591 [62] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,  
592 2009.
- 593 [63] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural  
594 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 595 [64] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in  
596 neurobiology*, 25:156–163, 2014.
- 597 [65] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current  
598 opinion in neurobiology*, 46:1–6, 2017.
- 599 [66] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H  
600 Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex  
601 embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966,  
602 2018.
- 603 [67] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian com-  
604 putation*. CRC Press, 2018.
- 605 [68] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based infer-  
606 ence. *Proceedings of the National Academy of Sciences*, 2020.
- 607 [69] Eve Marder and Allen I Selverston. *Dynamic biological networks: the stomatogastric nervous  
608 system*. MIT press, 1992.
- 609 [70] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory  
610 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–  
611 592, 2009.
- 612 [71] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralin-  
613 ear network: a unifying circuit motif underlying multi-input integration in sensory cortex.  
614 *Neuron*, 85(2):402–417, 2015.

- 615 [72] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and  
616 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,  
617 5(10):793, 2004.
- 618 [73] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of  
619 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-*  
620 *robiology*, 71(1):45–61, 2011.
- 621 [74] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-  
622 tex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 623 [75] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-  
624 bition of inhibition in visual cortex: the logic of connections between molecularly distinct  
625 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 626 [76] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn.  
627 Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.
- 628 [77] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate  
629 cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
- 630 [78] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D  
631 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-  
632 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 633 [79] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and  
634 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,  
635 2015.
- 636 [80] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-  
637 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 638 [81] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo  
639 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,  
640 2003.
- 641 [82] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural  
642 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.

- 643 [83] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and  
644 Edward Teller. Equation of state calculations by fast computing machines. *The journal of*  
645 *chemical physics*, 21(6):1087–1092, 1953.
- 646 [84] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.  
647 1970.
- 648 [85] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte  
649 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,  
650 73(2):123–214, 2011.
- 651 [86] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-  
652 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–  
653 820, 2011.
- 654 [87] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-  
655 free variational inference. In *Advances in Neural Information Processing Systems*, pages  
656 5523–5533, 2017.
- 657 [88] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-  
658 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience*  
659 *Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 660 [89] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,  
661 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience  
662 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 663 [90] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and  
664 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In  
665 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 666 [91] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-  
667 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,*  
668 *Germany*, 2019.
- 669 [92] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-  
670 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural  
671 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.

- 672 [93] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and  
673 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 674 [94] Sean R Bittner and John P Cunningham. Approximating exponential family models (not  
675 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 676 [95] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural  
677 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,  
678 2012.
- 679 [96] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary  
680 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,  
681 2018.
- 682 [97] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable  
683 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 684 [98] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula  
685 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially  
686 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–  
687 1929, 2009.
- 688 [99] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter  
689 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 690 [100] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of  
691 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 692 [101] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-  
693 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv  
694 preprint arXiv:1912.02762*, 2019.
- 695 [102] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-  
696 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 697 [103] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.  
698 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-  
699 tion processing systems*, 29:4743–4751, 2016.

- 700 [104] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.
- 701
- 702 [105] Emmanuel Klinger, Dennis Rickert, and Jan Hasenauer. pyabc: distributed, likelihood-free  
703 inference. *Bioinformatics*, 34(20):3591–3593, 2018.
- 704 [106] David S Greenberg, Marcel Nonnenmacher, and Jakob H Macke. Automatic posterior trans-  
705 formation for likelihood-free inference. *International Conference on Machine Learning*, 2019.

706 **5 Methods**

707 **5.1 Emergent property inference (EPI)**

708 Determining the combinations of model parameters that can produce observed data or a desired  
709 output is a key part of scientific practice. Solving inverse problems is especially important in  
710 neuroscience, since we require complex models to describe the complex phenomena of neural com-  
711 putations. While much machine learning research has focused on how to find latent structure  
712 in large-scale neural datasets, less has focused on inverting theoretical circuit models conditioned  
713 upon the emergent phenomena they produce. Here, we introduce a novel method for statistical  
714 inference, which finds distributions of parameter solutions that only produce the desired emer-  
715 gent property. This method seamlessly handles neural circuit models with stochastic nonlinear  
716 dynamical generative processes, which are predominant in theoretical neuroscience.

717 Consider model parameterization  $\mathbf{z}$ , which is a collection of scientifically interesting variables that  
718 govern the complex simulation of data  $\mathbf{x}$ . For example (see Section 3.1),  $\mathbf{z}$  may be the electrical  
719 conductance parameters of an STG subcircuit, and  $\mathbf{x}$  the evolving membrane potentials of the five  
720 neurons. In terms of statistical modeling, this circuit model has an intractable likelihood  $p(\mathbf{x} | \mathbf{z})$ ,  
721 which is predicated by the stochastic differential equations that define the model. Even so, we do  
722 not scientifically reason about how  $\mathbf{z}$  governs all of  $\mathbf{x}$ , but rather specific phenomena that are a  
723 function of the data  $f(\mathbf{x}; \mathbf{z})$ . In the STG example,  $f(\mathbf{x}; \mathbf{z})$  measures hub neuron frequency from the  
724 evolution of  $\mathbf{x}$  governed by  $\mathbf{z}$ . With EPI, we learn distributions of  $\mathbf{z}$  that results in an average and  
725 variance of  $f(\mathbf{x}; \mathbf{z})$ , denoted  $\boldsymbol{\mu}$  and  $\sigma^2$ . We refer to the collection of these statistical moments as an  
726 emergent property. Such emergent properties  $\mathcal{X}$  are defined through choice of  $f(\mathbf{x}; \mathbf{z})$  (which may  
727 be one or multiple statistics),  $\boldsymbol{\mu}$ , and  $\sigma^2$

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \sigma^2. \quad (11)$$

728 Precisely, the emergent property statistics  $f(\mathbf{x}; \mathbf{z})$  must have means  $\boldsymbol{\mu}$  and variances  $\sigma^2$  over the  
729 EPI distribution of parameters and stochasticity of the data given the parameters. By defining  
730 these means and variances over both levels of stochasticity – the inferred distribution and that of  
731 the model – there is a fine degree of control over predictions made by the inferred parameters.  
732 In EPI, deep probability distributions are optimized to learn the inferred distribution. In deep  
733 probability distributions, a simple random variable  $\mathbf{z}_0 \sim q_0(\mathbf{z}_0)$  is mapped deterministically via a  
734 sequence of deep neural network layers ( $g_1, \dots, g_l$ ) parameterized by weights and biases  $\boldsymbol{\theta}$  to the

735 support of the distribution of interest:

$$\mathbf{z} = g_{\theta}(\mathbf{z}_0) = g_l(\dots g_1(\mathbf{z}_0)) \sim q_{\theta}(\mathbf{z}). \quad (12)$$

736 Such deep probability distributions embed the inferred distribution in a deep network. Once opti-  
737 mized, this deep network representation has remarkably useful properties: fast sampling, probability  
738 evaluations, and also first- and second-order probability gradient evaluations.

739 By choosing a neural circuit model, often represented as a system of differential equations, we  
740 implicitly define a model likelihood  $p(\mathbf{x} | \mathbf{z})$ , which may be unknown or intractable for our purposes.  
741 Given this model choice and that of an emergent property  $\mathcal{X}$ ,  $q_{\theta}(\mathbf{z})$  is optimized via the neural  
742 network parameters  $\theta$  to find a maximally entropic distribution  $q_{\theta}^*$  within the deep variational  
743 family  $\mathcal{Q}$  producing the emergent property  $\mathcal{X}$ :

$$q_{\theta}(\mathbf{z} | \mathcal{X}) = q_{\theta}^*(\mathbf{z}) = \operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}(\mathbf{z})) \quad (13)$$
$$\text{s.t. } \mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2.$$

744 Entropy is chosen as the normative selection principle to match that of Bayesian inference (see  
745 Section 5.1.5). However, a key difference is that variational inference and other Bayesian methods  
746 do not constrain the predictions of their inferred parameter distribution. This optimization is  
747 executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [48].

748 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.  
749 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-  
750 ond, we describe the special class of deep probability distributions used in EPI called normalizing  
751 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-  
752 tion 13 (Section 5.1.3). Then, we demonstrate the details of this optimization in a toy example  
753 (Section 5.1.4). Finally, we establish the known relationship between maximum entropy distribu-  
754 tions and exponential families (Section 5.1.5), which is used to explain how EPI can be viewed as  
755 a form of variational inference (Section 5.1.6).

### 756 5.1.1 Related approaches

757 When Bayesian inference problems lack conjugacy, scientists use approximate inference methods like  
758 variational inference (VI) [82] and Markov chain Monte Carlo (MCMC) [83, 84]. After optimization,  
759 variational methods return a parameterized posterior distribution, which we can analyze. Also, the  
760 variational approximating distribution class is often chosen such that it permits fast sampling. In

761 contrast MCMC methods only produce samples from the approximated posterior distribution. No  
762 parameterized distribution is estimated, and additional samples are always generated with the same  
763 sampling complexity. Inference in models defined by systems of differential has been demonstrated  
764 with MCMC [85], although this approach requires tractable likelihoods. Advancements have lever-  
765 aged structure in stochastic differential equation models to improve likelihood approximations, thus  
766 expanding the domain of applicable models [86].

767 Simulation-based inference [68] is model parameter inference in the absence of a tractable likelihood  
768 function. The most prevalent approach to simulation-based inference is approximate Bayesian  
769 computation [80], in which satisfactory parameter samples are kept from random prior sampling  
770 according to a rejection heuristic. The obtained set of parameters do not have a probabilities,  
771 and further insight about the model must be gained from examination of the parameter set and  
772 their generated activity. Methodological advances to ABC methods have come through the use  
773 of Markov chain Monte Carlo (MCMC-ABC) [81] and sequential Monte Carlo (SMC-ABC) [43]  
774 sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still struggles  
775 to scale in dimensionality (cf. Fig. 2). Furthermore, once a parameter set has been obtained by  
776 SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again from scratch  
777 with a new population of initialized particles to obtain additional samples.

778 For scientific model analysis, we seek a parameter distribution exhibiting the properties of a well-  
779 chosen variational approximation: a parametric form conferring analytic calculations, and trivial  
780 sampling time. For this reason, ABC and MCMC techniques are unattractive, since they only  
781 produce a set of parameter samples and have unchanging sampling rate. EPI infers parameters  
782 in mechanistic models using the MEFN [48] algorithm using a deep variational approximation.  
783 The deep neural network of EPI defines the parametric form of the distribution approximation.  
784 Furthermore, the EPI distribution is constrained to produce an emergent property. In other words,  
785 the summary statistics of the posterior predictive distribution are fixed to have certain first and  
786 second moments. EPI optimization is enabled using stochastic gradient techniques in the spirit  
787 of likelihood-free variational inference [87]. The analytic relationship between EPI and variational  
788 inference is explained in Secton 5.1.6.

789 We note that, during our preparation and early presentation of this work [88, 89], another work  
790 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural  
791 circuits ([45, 90, 91]). We are encouraged by this general problem being recognized by others in the  
792 community, and we emphasize that these works offer complementary neuroscientific contributions

793 (different theoretical models of focus) and use different technical methodologies (ours is built on  
794 our prior work [48], theirs similarly [92]).

795 The method EPI differs from SNPE in some key ways. SNPE belongs to a “sequential” class  
796 of recently developed simulation-based inference methods in which two neural networks are used  
797 for posterior inference. This first neural network is a deep probability distribution (normalizing  
798 flow) used to estimate the posterior  $p(\mathbf{z} | \mathbf{x})$  (SNPE) or the likelihood  $p(\mathbf{x} | \mathbf{z})$  (sequential neural  
799 likelihood (SNL [46])). A recent advance uses an unconstrained neural network to estimate the  
800 likelihood ratio (sequential neural ratio estimation (SNRE [47])). In SNL and SNRE, MCMC  
801 sampling techniques are used to obtain samples from the approximated posterior. This contrasts  
802 with EPI and SNPE, which use deep probability distributions to model parameters, which facilitates  
803 immediate measurements of sample probability, gradient, or Hessian for system analysis. The  
804 second neural network in this sequential class of methods is the amortizer. This unconstrained  
805 deep network maps data  $\mathbf{x}$  (or statistics  $f(\mathbf{x}; \mathbf{z})$ ) or model parameters  $\mathbf{z}$  to the weights and biases of  
806 the first neural network. These methods are optimized on a conditional density (or ratio) estimation  
807 objective. The data used to optimize this objective are generated via an adaptive procedure, in  
808 which training data pairs  $(\mathbf{x}_i, \mathbf{z}_i)$  become sequentially closer to the true data and posterior.

809 The approximating fidelity of the deep probability distribution in sequential approaches is opti-  
810 mized to generalize across the training distribution of the conditioning variable. This generalization  
811 property of the sequential methods can reduce the accuracy at the singular posterior of interest.

812 Whereas in EPI, the entire expressivity of the deep probability distribution is dedicated to learning  
813 a single distribution as well as possible. Amortization is not possible in EPI, since EPI learns  
814 an exponential family distribution parameterized by its mean (see Section 5.1.5). Since EPI dis-  
815 tributions are defined by the mean  $\mu$  of their statistics, there is the well-known inverse mapping  
816 problem of exponential families [93] that prohibits an amortization based approach. However, we  
817 have shown that the same two-network architecture of the sequential simulation-based inference  
818 methods can be used for amortized inference in intractable exponential family posteriors using their  
819 natural parameterization [94].

820 Finally, one important differentiating factor between EPI and sequential simulation-based infer-  
821 ence methods is that EPI leverages gradients  $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$  during optimization. These gradients can  
822 improve convergence time and scalability, as we have shown on an example conditioning low-rank  
823 RNN connectivity on the property of stable amplification (see Section 3.3). With EPI, we prove  
824 out the suggestion that a deep inference technique can improve efficiency by leveraging these model

825 gradients when they are tractable. Sequential simulation-based inference techniques may be better  
 826 suited for scientific problems where  $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$  is intractable or unavailable: when there is a non-  
 827 differentiable model or it requires lengthy simulations. However, the sequential simulation-based  
 828 inference techniques cannot constrain the predictions of the inferred distribution in the manner of  
 829 EPI.

830 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in  
 831 natural models. Around a point, one can measure the Jacobian. One approach that scales well is  
 832 EAR [95]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [96] and  
 833 its stochastic adaptation [97]. Casting identifiability as a statistical estimation problem, the profile  
 834 likelihood can assess via iterated optimization while holding parameters fixed [98]. An exciting  
 835 recent method is capable of recovering the functional form of such unidentifiabilities away from a  
 836 point by following degenerate dimensions of the fisher information matrix [99]. Global structural  
 837 non-identifiabilities can be found for models with polynomial or rational dynamics equations using  
 838 DAISY [100]. With EPI, we have all the benefits given by a statistical inference method plus the  
 839 ability to query the first- or second-order gradient of the probability of the inferred distribution at  
 840 any chosen parameter value. The second-order gradient of the log probability (the Hessian), which  
 841 is directly afforded by EPI distributions, produces salient information about parametric sensitivity  
 842 of the emergent property. For example, the eigenvector with most negative eigenvalue of the Hessian  
 843 shows parametric combinations away from a parameter choice that decrease the in EPI distribution  
 844 probability the fastest. We refer to this eigenvector as the sensitivity dimension, and it is used to  
 845 generate scientific insight about a model of superior colliculus connectivity (see Section 3.5).

846 **5.1.2 Deep probability distributions and normalizing flows**

847 Deep probability distributions are comprised of multiple layers of fully connected neural networks  
 848 (Equation 12). When each neural network layer is restricted to be a bijective function, the sample  
 849 density can be calculated using the change of variables formula at each layer of the network. For  
 850  $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$ ,

$$p(\mathbf{z}_i) = p(g_i^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial g_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} \right| = p(\mathbf{z}_{i-1}) \left| \det \frac{\partial g_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right|^{-1}. \quad (14)$$

851 However, this computation has cubic complexity in dimensionality for fully connected layers. By  
 852 restricting our layers to normalizing flows [49, 101] – bijective functions with fast log determinant  
 853 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log

854 probability calculation confers efficient optimization of the maximum entropy objective (see Section  
 855 5.1.3). We use the Real NVP [60] normalizing flow class, because its coupling architecture confers  
 856 both fast sampling (forward) and fast log probability evaluation (backward). Fast probability  
 857 evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout  
 858 parameter space. Glow permutations were used in between coupling stages [102]. This is in contrast  
 859 to autoregressive architectures [61, 103], in which only one of the forward or backward passes can  
 860 be efficient. In this work, normalizing flows are used as flexible posterior approximations  $q_{\theta}(\mathbf{z})$   
 861 having weights and biases  $\theta$ . We specify the architecture used in each application by the number  
 862 of Real-NVP affine coupling stages, and the number of neural network layers and units per layer  
 863 of the conditioning functions.

### 864 5.1.3 Augmented Lagrangian optimization

865 To optimize  $q_{\theta}(\mathbf{z})$  in Equation 13, the constrained maximum entropy optimization is executed using  
 866 the augmented Lagrangian method. The following objective is minimized:

$$L(\theta; \eta_{\text{opt}}, c) = -H(q_{\theta}) + \eta_{\text{opt}}^{\top} R(\theta) + \frac{c}{2} \|R(\theta)\|^2 \quad (15)$$

867 where average constraint violations  $R(\theta) = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [T(\mathbf{x}; \mathbf{z}) - \mu_{\text{opt}}]]$ ,  $\eta_{\text{opt}} \in \mathbb{R}^m$  are the  
 868 Lagrange multipliers where  $m = |\mu_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})| = 2|f(\mathbf{x}; \mathbf{z})|$ , and  $c$  is the penalty coefficient. The  
 869 sufficient statistics  $T(\mathbf{x}; \mathbf{z})$  and mean parameter  $\mu_{\text{opt}}$  are determined by the means  $\mu$  and variances  
 870  $\sigma^2$  of emergent property statistics  $f(\mathbf{x}; \mathbf{z})$  defined in Equation 13 (see Section 5.1.6). Specifically,  
 871  $T(\mathbf{x}; \mathbf{z})$  is a concatenation of the first and second moments,  $\mu_{\text{opt}}$  is a concatenation of  $\mu$  and  $\sigma^2$   
 872 (see section 5.1.5), and the Lagrange multipliers are closely related to the natural parameters  $\eta$  of  
 873 exponential families (see Section 5.1.5). Weights and biases  $\theta$  of the deep probability distribution  
 874 are optimized according to Equation 15 using the Adam optimizer with learning rate  $10^{-3}$  [104].

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

$$H(q_{\theta}(\mathbf{z})) = \int -q_{\theta}(\mathbf{z}) \log(q_{\theta}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [-\log(q_{\theta}(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\log(q_{\theta}(g_{\theta}(\mathbf{z}_0)))]. \quad (16)$$

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

$$\nabla_{\theta} H(q_{\theta}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\nabla_{\theta} \log(q_{\theta}(g_{\theta}(\mathbf{z}_0)))]. \quad (17)$$

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

889 The intention is that  $c$  and  $\eta_{\text{opt}}$  start at values encouraging entropic growth early in optimization.  
890 With each training epoch in which the update rule for  $c$  is invoked by unsatisfactory constraint  
891 error reduction, the constraint satisfaction terms are increasingly weighted, resulting in a decreased  
892 entropy. This encourages the discovery of suitable regions of parameter space, and the subsequent  
893 refinement of the distribution to produce the emergent property (see example in Section 5.1.4). The  
894 momentum parameters of the Adam optimizer are reset at the end of each augmented Lagrangian  
895 epoch.

896 Rather than starting optimization from some  $\theta$  drawn from a randomized distribution, we found  
897 that initializing  $q_{\theta}(\mathbf{z})$  to approximate an isotropic Gaussian distribution conferred more stable, con-  
898 sistent optimization. The parameters of the Gaussian initialization were chosen on an application-  
899 specific basis. Throughout the study, we chose isotropic Gaussian initializations with mean  $\mu_{\text{init}}$   
900 at the center of the distribution support and some standard deviation  $\sigma_{\text{init}}$ , except for one case,  
901 where an initialization informed by random search was used (see Section 5.2.1).

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

910 When assessing the suitability of EPI for a particular modeling question, there are some important  
911 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 5.2.2) 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, backpropagating gradients through the SDE evolution will be expensive.

#### 5.1.4 Example: 2D LDS

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

$$\tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (18)$$

with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (19)$$

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,  $\text{real}(\lambda_1)$ , of the oscillating system.  $\lambda_1$  is the eigenvalue of greatest real part when the imaginary component is zero, and alternatively of positive imaginary component when the eigenvalues are complex conjugate pairs. To learn the distribution of real entries of  $A$  that produce a band of oscillating systems around 1Hz, we formalized this emergent property as  $\text{real}(\lambda_1)$  having mean zero with variance  $0.25^2$ , and the oscillation frequency  $2\pi\text{imag}(\lambda_1)$  having mean  $\omega = 1$  Hz with variance  $(0.1\text{Hz})^2$ :

$$\mathbb{E}[T(\mathbf{x})] \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}. \quad (20)$$

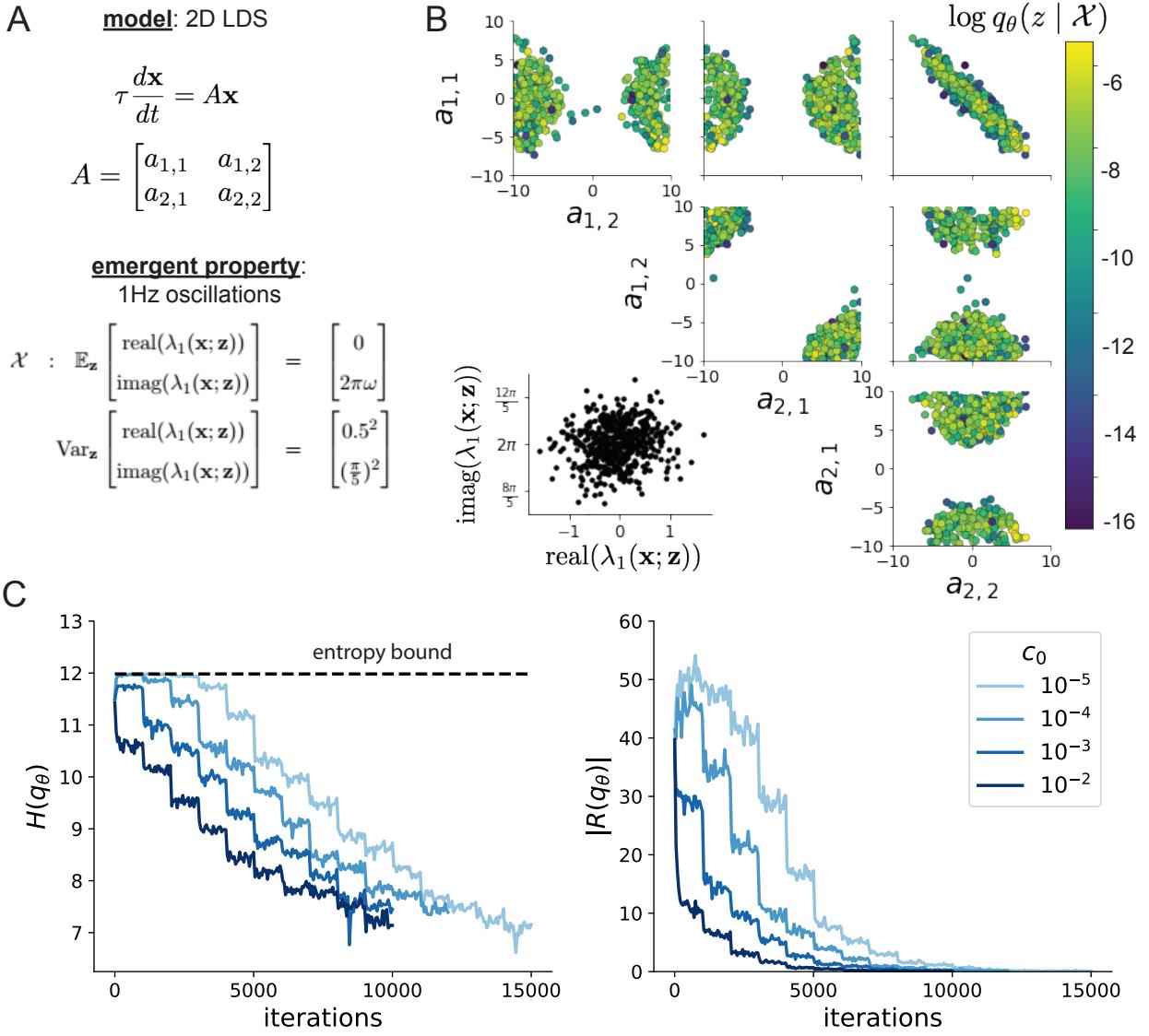


Figure 6: (LDS1): **A.** Two-dimensional linear dynamical system model, where real entries of the dynamics matrix  $A$  are the parameters. **B.** The EPI distribution for a two-dimensional linear dynamical system with  $\tau = 1$  that produces an average of 1Hz oscillations with some small amount of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

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:

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (21)$$

Importantly, even though  $\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})}[T(\mathbf{x})]$  is calculable directly via a closed form function and does not require simulation, we cannot derive the distribution  $q_{\theta}^*$  directly. This fact is due to the formally hard problem of the backward mapping: finding the natural parameters  $\eta$  from the mean parameters  $\mu$  of an exponential family distribution [93]. 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  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ , dotted black line at the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$ , and the dotted gray line at twice the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.5$  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  $\text{real}(\lambda_1)$  have similar log densities.

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4-a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (22)$$

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

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \quad (23)$$

Since we constrained  $\mathbb{E}_{\mathbf{z} \sim q_{\theta}}[\text{imag}(\lambda)] = 2\pi$  (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  $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_{\theta}}(|a_1a_4 - E_{q_{\theta}}[a_1a_4]|)$ , 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

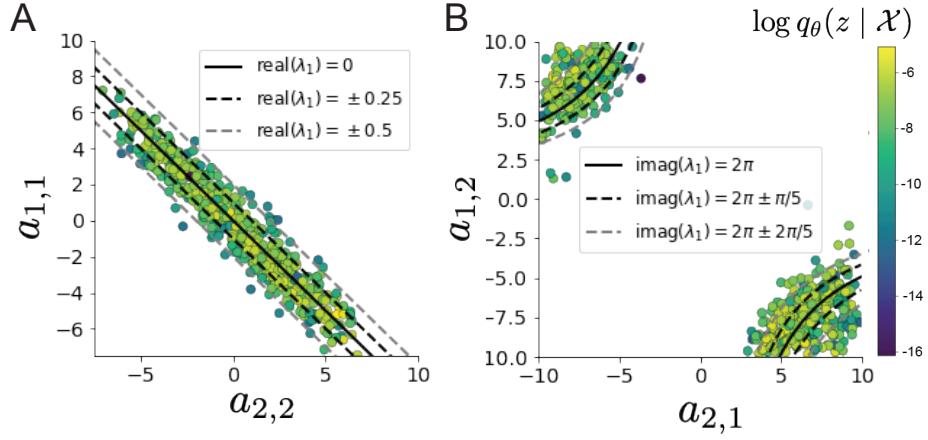


Figure 7: (LDS2): **A.** Probability contours in the  $a_1$ - $a_4$  plane were derived from the relationship to emergent property statistic of growth/decay factor  $\text{real}(\lambda_1)$ . **B.** Probability contours in the  $a_2$ - $a_3$  plane were derived from the emergent property statistic of oscillation frequency  $2\pi\text{imag}(\lambda_1)$ .

961 took steps in negative standard deviation of  $a_1a_4$  (dotted and gray lines), since there are few positive  
 962 values  $a_1a_4$  in the learned distribution. Subtler combinations of model and emergent property will  
 963 have more complexity, further motivating the use of EPI for understanding these systems. As we  
 964 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz  
 965 (Fig. S3).

### 966 5.1.5 Maximum entropy distributions and exponential families

967 EPI is a maximum entropy distribution, which have fundamental links to exponential family dis-  
 968 tributions. A maximum entropy distribution of form:

$$p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \quad (24)$$

s.t.  $\mathbb{E}_{\mathbf{z} \sim p} [T(\mathbf{z})] = \boldsymbol{\mu}_{\text{opt}}$ .

969 will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \quad (25)$$

970 The mappings between the mean parameterization  $\boldsymbol{\mu}_{\text{opt}}$  and the natural parameterization  $\boldsymbol{\eta}$  are  
 971 formally hard to identify except in special cases [93].

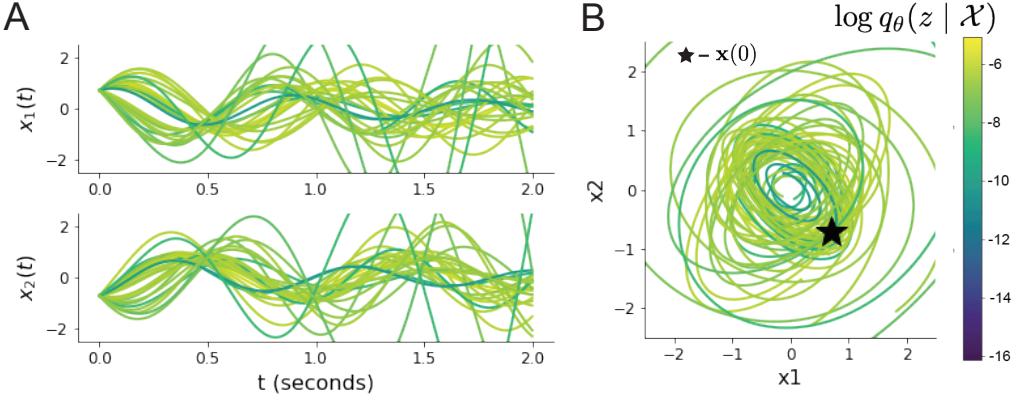


Figure 8: (LDS3): Sampled dynamical systems  $\mathbf{z} \sim q_\theta(\mathbf{z})$  and their simulated activity from  $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$  colored by log probability. **A.** Each dimension of the simulated trajectories throughout time. **B.** The simulated trajectories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

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

$$\text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \mathbb{E}_{\mathbf{z}, \mathbf{x}} \left[ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \right] \quad (26)$$

974 As a general maximum entropy distribution (Equation 24), the sufficient statistics vector contains  
975 both first and second order moments of  $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (27)$$

976 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (28)$$

### 977 5.1.6 EPI as variational inference

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

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \quad (29)$$

982 The posterior distribution is analytically available when the prior is conjugate with the likelihood.  
 983 However, conjugacy is rare in practice, and alternative methods, such as variational inference [82],  
 984 are utilized.

985 In variational inference, a posterior approximation  $q_{\theta}^*$  is chosen from within some variational family  
 986  $\mathcal{Q}$

$$q_{\theta}^*(\mathbf{z}) = \operatorname{argmin}_{q_{\theta} \in \mathcal{Q}} KL(q_{\theta}(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})). \quad (30)$$

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

$$KL(q_{\theta}(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(q_{\theta}(\mathbf{z}))] - \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(p(\mathbf{z} \mid \mathbf{x}))] \quad (31)$$

988

$$= -H(q_{\theta}) - \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(p(\mathbf{x} \mid \mathbf{z})) + \log(p(\mathbf{z})) - \log(p(\mathbf{x}))] \quad (32)$$

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

$$\operatorname{argmin}_{q_{\theta} \in \mathcal{Q}} KL(q_{\theta} \parallel p(\mathbf{z} \mid \mathbf{x})) = \operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}) + \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(p(\mathbf{x} \mid \mathbf{z})) + \log(p(\mathbf{z}))]. \quad (33)$$

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

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) \mid \boldsymbol{\mu}_f, \Sigma_f), \quad (34)$$

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

$$\log p(\mathbf{x} \mid \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \quad (35)$$

996 where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\boldsymbol{\mu}_f}{\sigma_f^2} \\ \frac{-1}{2\sigma_f^2} \end{bmatrix}, \text{ and} \quad (36)$$

997

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \quad (37)$$

998 The variational objective is then

$$\operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}) + \boldsymbol{\eta}_f^\top \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [T(\mathbf{x}; \mathbf{z})] \quad (38)$$

999 Comparing this to the Lagrangian objective (without augmentation) of EPI, we see they are the  
1000 same

$$\begin{aligned} q_{\theta}^*(\mathbf{z}) &= \underset{q_{\theta} \in Q}{\operatorname{argmin}} -H(q_{\theta}) + \boldsymbol{\eta}_{\text{opt}}^\top (\mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}_{\text{opt}}) \\ &= \underset{q_{\theta} \in Q}{\operatorname{argmin}} -H(q_{\theta}) + \boldsymbol{\eta}_{\text{opt}}^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})]. \end{aligned} \quad (39)$$

1001 where  $T(\mathbf{x}; \mathbf{z})$  consists of the first and second moments of the emergent property statistic  $f(\mathbf{x}; \mathbf{z})$   
1002 (Equation 27). Thus, EPI is implicitly executing variational inference with a uniform prior and a  
1003 mean-field gaussian likelihood on the emergent property statistics. The mean and variances of the  
1004 mean-field gaussian likelihood are predicated by  $\boldsymbol{\eta}_{\text{opt}}$  (Equations 36 and 38), which is adapted after  
1005 each EPI optimization epoch based on  $\mathcal{X}$  (see Section 5.1.3). In EPI, the inferred distribution is  
1006 not conditioned on a finite dataset as in variational inference, but rather the emergent property  
1007  $\mathcal{X}$  dictates the likelihood parameterization such that the inferred distribution will produce the  
1008 emergent property. As a note, we could not simply choose  $\boldsymbol{\mu}_f$  and  $\boldsymbol{\sigma}_f$  directly from the outset, since  
1009 we do not know which of these choices will produce the emergent property  $\mathcal{X}$ , which necessitates  
1010 the EPI optimization routine that adapts  $\boldsymbol{\eta}_{\text{opt}}$ . Accordingly, we replace the notation of  $p(\mathbf{z} | \mathbf{x})$   
1011 with  $p(\mathbf{z} | \mathcal{X})$  conceptualizing an inferred distribution that obeys emergent property  $\mathcal{X}$  (see Section  
1012 5.1).

## 1013 5.2 Theoretical models

1014 In this study, we used emergent property inference to examine several models relevant to theoretical  
1015 neuroscience. Here, we provide the details of each model and the related analyses.

### 1016 5.2.1 Stomatogastric ganglion

1017 We analyze how the parameters  $\mathbf{z} = [g_{el}, g_{synA}]$  govern the emergent phenomena of intermediate  
1018 hub frequency in a model of the stomatogastric ganglion (STG) [51] shown in Figure 1A with  
1019 activity  $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$ , using the same hyperparameter choices as Gutierrez et al.  
1020 Each neuron's membrane potential  $x_{\alpha}(t)$  for  $\alpha \in \{f1, f2, hub, s1, s2\}$  is the solution of the following  
1021 stochastic differential equation:

$$C_m \frac{dx_{\alpha}}{dt} = -[h_{leak}(\mathbf{x}; \mathbf{z}) + h_{Ca}(\mathbf{x}; \mathbf{z}) + h_K(\mathbf{x}; \mathbf{z}) + h_{hyp}(\mathbf{x}; \mathbf{z}) + h_{elec}(\mathbf{x}; \mathbf{z}) + h_{syn}(\mathbf{x}; \mathbf{z})] + dB. \quad (40)$$

1022 The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization,  
1023 electrical and synaptic currents as well as gaussian noise  $dB$ . Each current component is a function

1024 of all membrane potentials and the conductance parameters  $\mathbf{z}$ .

1025 The capacitance of the cell membrane was set to  $C_m = 1nF$ . Specifically, the currents are the  
 1026 difference in the neuron's membrane potential and that current type's reversal potential multiplied  
 1027 by a conductance:

$$1028 \quad h_{leak}(\mathbf{x}; \mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (41)$$

$$1029 \quad h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{post} - x_\alpha^{pre}) \quad (42)$$

$$1030 \quad h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (43)$$

$$1031 \quad h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (44)$$

$$1032 \quad h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (45)$$

$$1032 \quad h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (46)$$

1033 The reversal potentials were set to  $V_{leak} = -40mV$ ,  $V_{Ca} = 100mV$ ,  $V_K = -80mV$ ,  $V_{hyp} = -20mV$ ,  
 1034 and  $V_{syn} = -75mV$ . The other conductance parameters were fixed to  $g_{leak} = 1 \times 10^{-4}\mu S$ ,  $g_{Ca}$ ,  
 1035  $g_K$ , and  $g_{hyp}$  had different values based on fast, intermediate (hub) or slow neuron. The fast  
 1036 conductances had values  $g_{Ca} = 1.9 \times 10^{-2}$ ,  $g_K = 3.9 \times 10^{-2}$ , and  $g_{hyp} = 2.5 \times 10^{-2}$ . The intermediate  
 1037 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  
 1038 slow conductances had values  $g_{Ca} = 8.5 \times 10^{-3}$ ,  $g_K = 1.5 \times 10^{-2}$ , and  $g_{hyp} = 1.0 \times 10^{-2}$ .

1039 Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating  
 1040 dynamics dependent on steady-state gating variables  $M_\infty$ ,  $N_\infty$  and  $H_\infty$ , respectively:

$$1041 \quad M_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_\alpha - v_1}{v_2} \right) \right) \quad (47)$$

$$1042 \quad \frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (48)$$

$$1043 \quad N_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_\alpha - v_3}{v_4} \right) \right) \quad (49)$$

$$1044 \quad \lambda_N = \phi_N \cosh \left( \frac{x_\alpha - v_3}{2v_4} \right) \quad (50)$$

$$1045 \quad \frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (51)$$

$$1046 \quad H_\infty = \frac{1}{1 + \exp \left( \frac{x_\alpha + v_5}{v_6} \right)} \quad (52)$$

$$1045 \quad \tau_h = 272 - \left( \frac{-1499}{1 + \exp \left( \frac{-x_\alpha + v_7}{v_8} \right)} \right). \quad (53)$$

1047 where we set  $v_1 = 0mV$ ,  $v_2 = 20mV$ ,  $v_3 = 0mV$ ,  $v_4 = 15mV$ ,  $v_5 = 78.3mV$ ,  $v_6 = 10.5mV$ ,  
 1048  $v_7 = -42.2mV$ ,  $v_8 = 87.3mV$ ,  $v_9 = 5mV$ , and  $v_{th} = -25mV$ .

1049 Finally, there is a synaptic gating variable as well:

$$S_\infty = \frac{1}{1 + \exp\left(\frac{v_{th}-x_\alpha}{v_9}\right)}. \quad (54)$$

1050 When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear  
 1051 dynamical system. Gaussian noise  $d\mathbf{B}$  of variance  $(1 \times 10^{-12})^2 \text{ A}^2$  makes the model stochastic, and  
 1052 introduces variability in frequency at each parameterization  $\mathbf{z}$ .

1053 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for  
 1054  $T = 300$  time steps of  $dt = 25\text{ms}$ . The chosen  $dt$  and  $T$  were the most computationally convenient  
 1055 choices yielding accurate frequency measurement. We used a basis of complex exponentials with  
 1056 frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

$$\Phi = [0.0, 0.01, \dots, 1.0]^\top \dots \quad (55)$$

1057 To measure spiking frequency, we processed simulated membrane potentials with a relu (spike  
 1058 extraction) and low-pass filter with averaging window of size 20, then took the frequency with the  
 1059 maximum absolute value of the complex exponential basis coefficients of the processed time-series.  
 1060 The first 20 temporal samples of the simulation are ignored to account for initial transients.

1061 To differentiate through the maximum frequency identification, we used a soft-argmax Let  $X_\alpha \in$   
 1062  $\mathcal{C}^{|\Phi|}$  be the complex exponential filter bank dot products with the signal  $x_\alpha \in \mathbb{R}^N$ , where  $\alpha \in$   
 1063  $\{\text{f1, f2, hub, s1, s2}\}$ . The soft-argmax is then calculated using temperature parameter  $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta|X_\alpha| \odot i), \quad (56)$$

1064 where  $i = [0, 1, \dots, 100]$ . The frequency is then calculated as

$$\omega_\alpha = 0.01\psi_\alpha \text{Hz}. \quad (57)$$

1065 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean  
 1066 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron  
 1067 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be  $(0.025\text{Hz})^2$  to  
 1068 capture variation in frequency between 0.5Hz and 0.6Hz (Equation 4). As a maximum entropy dis-  
 1069 tribution,  $T(\mathbf{x}, \mathbf{z})$  is comprised of both these first and second moments of the hub neuron frequency

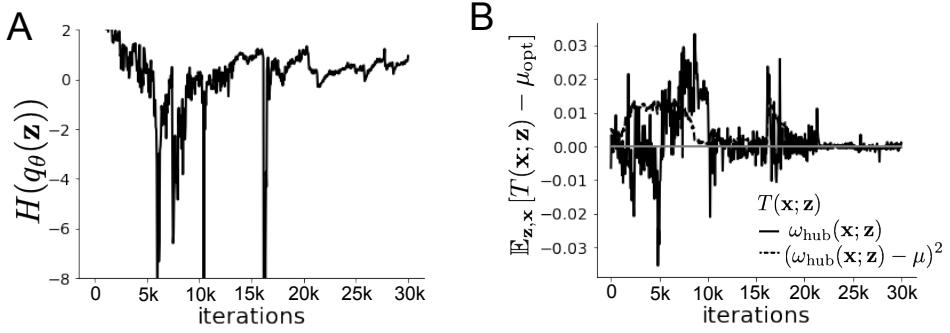


Figure 9: (STG1): EPI optimization of the STG model producing network syncing. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

1070 (as in Equations 27 and 28)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \quad (58)$$

1071

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \quad (59)$$

1072 Throughout optimization, the augmented Lagrangian parameters  $\eta$  and  $c$ , were updated after each  
1073 epoch of 5,000 iterations (see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

1074 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-  
1075 layer neural networks of 25 units per layer. The normalizing flow architecture mapped  $z_0 \sim \mathcal{N}(\mathbf{0}, I)$   
1076 to a support of  $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}] \in [4, 8] \times [0.01, 4]$ , initialized to a gaussian approximation of samples  
1077 returned by a preliminary ABC search. We did not include  $g_{\text{synA}} < 0.01$ , for numerical stability.  
1078 EPI optimization was run using 5 different random seeds for architecture initialization  $\boldsymbol{\theta}$  with an  
1079 augmented Lagrangian coefficient of  $c_0 = 10^5$ , a batch size  $n = 400$ , and  $\beta = 2$ . The distribution  
1080 shown is that of the architecture converging with criteria  $N_{\text{test}} = 100$  at greatest entropy across  
1081 random seeds.

1082 We calculated the Hessian at the mode of the inferred EPI distribution. The Hessian of a probability  
1083 model is the second order gradient of the log probability density  $\log q_{\boldsymbol{\theta}}(\mathbf{z})$  with respect to the  
1084 parameters  $\mathbf{z}$ :  $\frac{\partial^2 \log q_{\boldsymbol{\theta}}(\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^T}$ . With EPI, we can examine the Hessian, which is analytically available  
1085 throughout distribution, to indicate the dimensions of parameter space that are sensitive (strongly  
1086 negative eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the

1087 emergent property produced. In Figure 1D, the eigenvectors of the Hessian  $v_1$  (solid) and  $v_2$   
 1088 (dashed) are shown evaluated at the mode of the distribution. The length of the arrows is inversely  
 1089 proportional to the square root of absolute value of their eigenvalues  $\lambda_1 = -10.7$  and  $\lambda_2 = -3.22$ .  
 1090 Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D parameter  
 1091 space are chosen arbitrarily.

## 1092 5.2.2 Scaling EPI for stable amplification in RNNs

1093 We examined the scaling properties of EPI by learning connectivities of RNNs of increasing size  
 1094 that exhibit stable amplification. Rank-2 RNN connectivity was modeled as  $W = UV^\top$ , where  
 1095  $U = [\mathbf{u}_1 \ \mathbf{u}_2] + g\chi^{(W)}$ ,  $V = [\mathbf{v}_1 \ \mathbf{v}_2] + g\chi^{(V)}$ , and  $\chi_{i,j}^{(W)}, \chi_{i,j}^{(V)} \sim \mathcal{N}(0, 1)$ . This RNN model has  
 1096 dynamics

$$1097 \tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}. \quad (60)$$

1097 In this analysis, we inferred connectivity parameterizations  $\mathbf{z} = [\mathbf{u}_1^\top, \mathbf{u}_2^\top, \mathbf{v}_1^\top, \mathbf{v}_2^\top]^\top \in [-1, 1]^{(4N)}$   
 1098 that produced stable amplification using EPI, SMC-ABC [43], and SNPE [45] (see Section Related  
 1099 Methods).

1100 For this RNN model to be stable, all real eigenvalues of  $W$  must be less than 1:  $\text{real}(\lambda_1) < 1$ ,  
 1101 where  $\lambda_1$  denotes the greatest real eigenvalue of  $W$ . For a stable RNN to amplify at least one input  
 1102 pattern, the symmetric connectivity  $W^s = \frac{W+W^\top}{2}$  must have an eigenvalue greater than 1:  $\lambda_1^s > 1$ ,  
 1103 where  $\lambda^s$  is the maximum eigenvalue of  $W^s$ . These two conditions are necessary and sufficient for  
 1104 stable amplification in RNNs [63]. We defined the emergent property of stable amplification with  
 1105 means of these eigenvalues (0.5 and 1.5, respectively) that satisfy these conditions. To complete  
 1106 the emergent property definition, we chose variances ( $0.25^2$ ) about those means such that samples  
 1107 rarely violate the eigenvalue constraints. In terms of the EPI optimization variables, this is written  
 1108 as

$$1109 T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \lambda_1^s(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0.5)^2 \\ (\lambda_1^s(\mathbf{x}; \mathbf{z}) - 1.5)^2 \end{bmatrix}, \quad (61)$$

$$1109 \boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.25^2 \\ 0.25^2 \end{bmatrix}. \quad (62)$$

1110 Gradients of maximum eigenvalues of Hermitian matrices like  $W^s$  are available with modern auto-  
 1111 automatic differentiation tools. To differentiate through the  $\text{real}(\lambda_1)$ , we solved the following equation  
 1112 for eigenvalues of rank-2 matrices using the rank reduced matrix  $W^r = V^\top U$

$$\lambda_{\pm} = \frac{\text{Tr}(W^r) \pm \sqrt{\text{Tr}(W^r)^2 - 4\text{Det}(W^r)}}{2}. \quad (63)$$

1113 For EPI in Fig. 2, we used a real NVP architecture with three coupling layers of affine transfor-  
 1114 mations parameterized by two-layer neural networks of 100 units per layer. The initial distribution  
 1115 was a standard isotropic gaussian  $z_0 \sim \mathcal{N}(\mathbf{0}, I)$  mapped to the support of  $\mathbf{z}_i \in [-1, 1]$ . We used  
 1116 an augmented Lagrangian coefficient of  $c_0 = 10^3$ , a batch size  $n = 200$ ,  $\beta = 4$ , and chose to use  
 1117 500 iterations per augmented Lagrangian epoch and emergent property constraint convergence was  
 1118 evaluated at  $N_{\text{test}} = 200$  (Fig. 2B blue line, and Fig. 2C-D blue).

1119 We compared EPI to two alternative simulation-based inference techniques, since the likelihood  
 1120 of these eigenvalues given  $\mathbf{z}$  is not available. Approximate Bayesian computation (ABC) [80] is a  
 1121 rejection sampling technique for obtaining sets of parameters  $\mathbf{z}$  that produce activity  $\mathbf{x}$  close to some  
 1122 observed data  $\mathbf{x}_0$ . Sequential Monte Carlo approximate Bayesian computation (SMC-ABC) is the  
 1123 state-of-the-art ABC method, which leverages SMC techniques to improve sampling speed. We ran  
 1124 SMC-ABC with the pyABC package [105] to infer RNNs with stable amplification: connectivities  
 1125 having eigenvalues within an  $\epsilon$ -defined  $l_2$  distance of

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}. \quad (64)$$

1126 SMC-ABC was run with a uniform prior over  $\mathbf{z} \in [-1, 1]^{(4N)}$ , a population size of 1,000 particles  
 1127 with simulations parallelized over 32 cores, and a multivariate normal transition model.

1128 SNPE, the next approach in our comparison, is far more similar to EPI. Like EPI, SNPE treats pa-  
 1129 rameters in mechanistic models with deep probability distributions, yet the two learning algorithms  
 1130 are categorically different. SNPE uses a two-network architecture to approximate the posterior dis-  
 1131 tribution of the model conditioned on observed data  $\mathbf{x}_0$ . The amortizing network maps observations  
 1132  $\mathbf{x}_i$  to the parameters of the deep probability distribution. The weights and biases of the parameter  
 1133 network are optimized by sequentially augmenting the training data with additional pairs  $(\mathbf{z}_i, \mathbf{x}_i)$   
 1134 based on the most recent posterior approximation. This sequential procedure is important to get  
 1135 training data  $\mathbf{z}_i$  to be closer to the true posterior, and  $\mathbf{x}_i$  to be closer to the observed data. For  
 1136 the deep probability distribution architecture, we chose a masked autoregressive flow with affine  
 1137 couplings (the default choice), three transforms, 50 hidden units, and a normalizing flow mapping

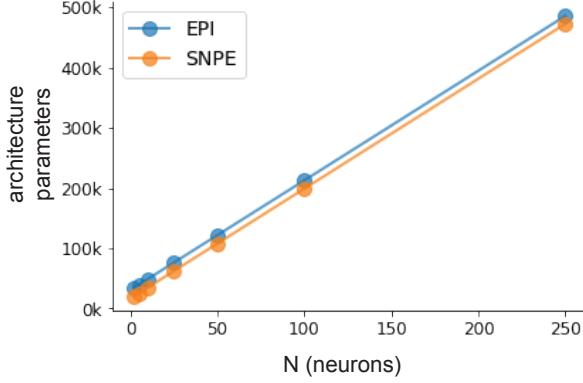


Figure 10: (RNN1): Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size ( $N$ ).

1138 to the support as in EPI. This architectural choice closely tracked the size of the architecture used  
 1139 by EPI (Fig. 10). As in SMC-ABC, we ran SNPE with  $\mathbf{x}_0 = \mu$ . All SNPE optimizations were  
 1140 run for a limit of 1.5 days on a Tesla V100 GPU, or until two consecutive rounds resulted in a  
 1141 validation log probability lower than the maximum observed for that random seed.

1142 To clarify the difference in objectives of EPI and SNPE, we show their results on RNN models  
 1143 with different numbers of neurons  $N$  and random strength  $g$ . The parameters inferred by EPI  
 1144 consistently produces the same mean and variance of  $\text{real}(\lambda_1)$  and  $\lambda_1^s$ , while those inferred by  
 1145 SNPE change according to the model definition (Fig. 11A). For  $N = 2$  and  $g = 0.01$ , the SNPE  
 1146 posterior has greater concentration in eigenvalues around  $\mathbf{x}_0$  than at  $g = 0.1$ , where the model has  
 1147 greater randomness (Fig. 11B top, orange). At both levels of  $g$  when  $N = 2$ , the posterior of SNPE  
 1148 has lower entropy than EPI at convergence (Fig. 11B top). However at  $N = 10$ , SNPE results in  
 1149 a predictive distribution of more widely dispersed eigenvalues (Fig. 11A bottom), and an inferred  
 1150 posterior with greater entropy than EPI (Fig. 11B bottom). We highlight these differences not  
 1151 to focus on an insightful trend, but to emphasize that these methods optimize different objectives  
 1152 with different implications.

1153 Note that SNPE converges when it's validation log probability has saturated after several rounds  
 1154 of optimization (Fig. 11C), and that EPI converges after several epochs of its own optimization  
 1155 to enforce the emergent property constraints (Fig. 11D blue). Importantly, as SNPE optimizes  
 1156 its posterior approximation, the predictive means change, and at convergence may be different  
 1157 than  $\mathbf{x}_0$  (Fig. 11D orange, left). It is sensible to assume that predictions of a well-approximated  
 1158 SNPE posterior should closely reflect the data on average (especially given a uniform prior and

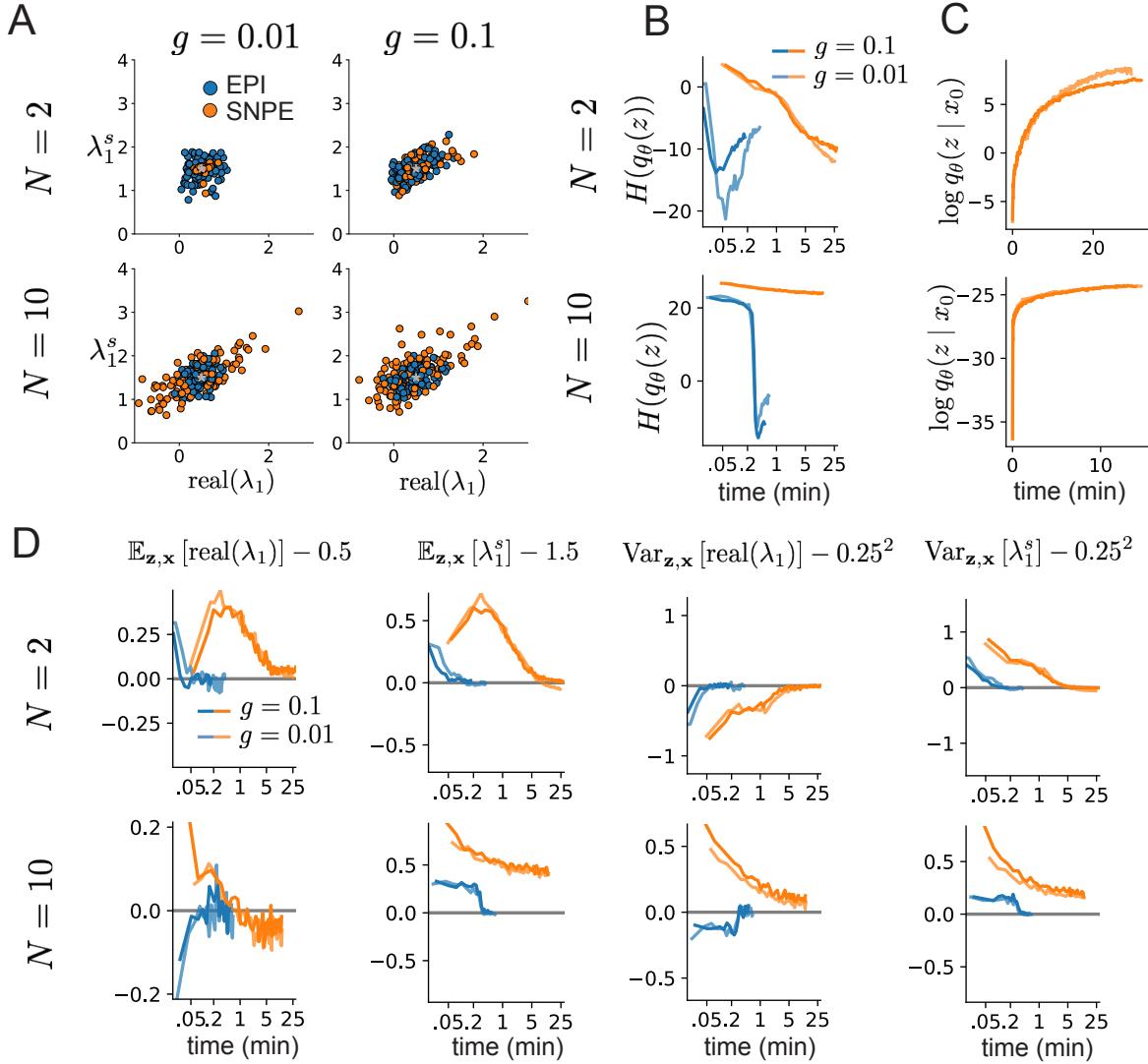


Figure 11: (RNN2): Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A.** Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with  $N = 2$  (top),  $N = 10$  (bottom),  $g = 0.01$  (left), and  $g = 0.1$  (right). **B.** Entropy of parameter distribution approximations throughout optimization with  $N = 2$  (top),  $N = 10$  (bottom),  $g = 0.1$  (dark shade), and  $g = 0.01$  (light shade). **C.** Validation log probabilities throughout SNPE optimization. Same conventions as B. **D.** Adherence to EPI constraints. Same conventions as B.

1159 a low degree of stochasticity), however this is not a given. Furthermore, no aspect of the SNPE  
1160 optimization controls the variance of the predictions (Fig. 11D orange, right).

1161 To compare the efficiency of these algorithms for inferring RNN connectivity distributions producing  
1162 stable amplification, we develop a convergence criteria that can be used across methods. While EPI  
1163 has its own hypothesis testing convergence criteria for the emergent property, it would not make  
1164 sense to use this criteria on SNPE and SMC-ABC which do not constrain the means and variances  
1165 of their predictions. Instead, we consider EPI and SNPE to have converged after completing its  
1166 most recent optimization epoch (EPI) or round (SNPE) in which the distance

$$d(q_\theta(z)) = |\mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}|_2 \quad (65)$$

1167 is less than 0.5. We consider SMC-ABC to have converged once the population produces samples  
1168 within the  $\epsilon = 0.5$  ball ensuring stable amplification.

1169 When assessing the scalability of SNPE, it is important to check that alternative hyperparameter-  
1170izations could not yield better performance. Key hyperparameters of the SNPE optimization are  
1171 the number of simulations per round  $n_{\text{round}}$ , the number of atoms used in the atomic proposals of  
1172 the SNPE-C algorithm [106], and the batch size  $n$ . To match EPI, we used a batch size of  $n = 200$   
1173 for  $N \leq 25$ , however we found  $n = 1,000$  to be helpful for SNPE in higher dimensions. While  
1174  $n_{\text{round}} = 1,000$  yielded SNPE convergence for  $N \leq 25$ , we found that a substantial increase to  
1175  $n_{\text{round}} = 25,000$  yielded more consistent convergence at  $N = 50$  (Fig. 12A). By increasing  $n_{\text{round}}$ ,  
1176 we also necessarily increase the duration of each round. At  $N = 100$ , we tried two hyperparameter  
1177 modifications. As suggested in [106], we increased  $n_{\text{atom}}$  by an order of magnitude to improve  
1178 gradient quality, but this had little effect on the optimization (much overlap between same random  
1179 seeds) (Fig. 12B). Finally, we increased  $n_{\text{round}}$  by an order of magnitude, which yielded convergence  
1180 in one case, but no others. We found no way to improve the convergence rate of SNPE without  
1181 making more aggressive hyperparameter choices requiring high numbers of simulations.

1182 In Figure 2C-D, we show samples from the random seed resulting in emergent property convergence  
1183 at greatest entropy (EPI), the random seed resulting in greatest validation log probability (SNPE),  
1184 and the result of all converged random seeds (SMC).

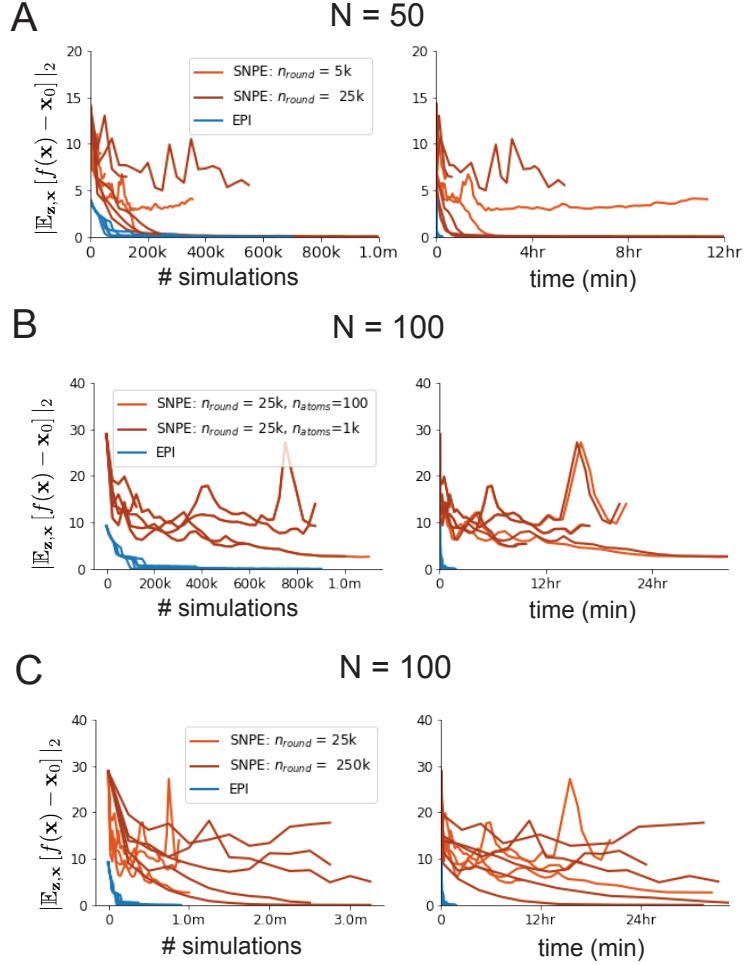


Figure 12: (RNN3): SNPE convergence was enabled by increasing  $n_{\text{round}}$ , not  $n_{\text{atom}}$ . **A.** Difference of mean predictions  $\mathbf{x}_0$  throughout optimization at  $N = 50$  with by simulation count (left) and wall time (right) of SNPE with  $n_{\text{round}} = 5,000$  (light orange), SNPE with  $n_{\text{round}} = 25,000$  (dark orange), and EPI (blue). Each line shows an individual random seed. **B.** Same conventions as A at  $N = 100$  of SNPE with  $n_{\text{atom}} = 100$  (light orange) and  $n_{\text{atom}} = 1,000$  (dark orange). **C.** Same conventions as A at  $N = 100$  of SNPE with  $n_{\text{round}} = 25,000$  (light orange) and  $n_{\text{round}} = 250,000$  (dark orange).

1185 **5.2.3 Primary visual cortex**

1186 In the stochastic stabilized supralinear network [78], population rate responses  $\mathbf{x}$  to input  $\mathbf{h}$ , recur-  
 1187 rent input  $W\mathbf{x}$  and slow noise  $\epsilon$  are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \epsilon), \quad (66)$$

1188 where the noise is an Ornstein-Uhlenbeck process  $\epsilon \sim OU(\tau_{\text{noise}}, \sigma)$

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (67)$$

1189 with  $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$ . The noisy process is parameterized as

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (68)$$

1190 so that  $\sigma$  parameterizes the variance of the noisy input in the absence of recurrent connectivity  
 1191 ( $W = \mathbf{0}$ ). As contrast  $c \in [0, 1]$  increases, input to the E- and P-populations increases relative to  
 1192 a baseline input  $\mathbf{h} = \mathbf{h}_b + c\mathbf{h}_c$ . Connectivity ( $W_{\text{fit}}$ ) and input ( $\mathbf{h}_{b,\text{fit}}$  and  $\mathbf{h}_{c,\text{fit}}$ ) parameters were fit  
 1193 using the deterministic V1 circuit model [56]

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (69)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (70)$$

1194 and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (71)$$

1195 To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equiv-  
 1196 alence class

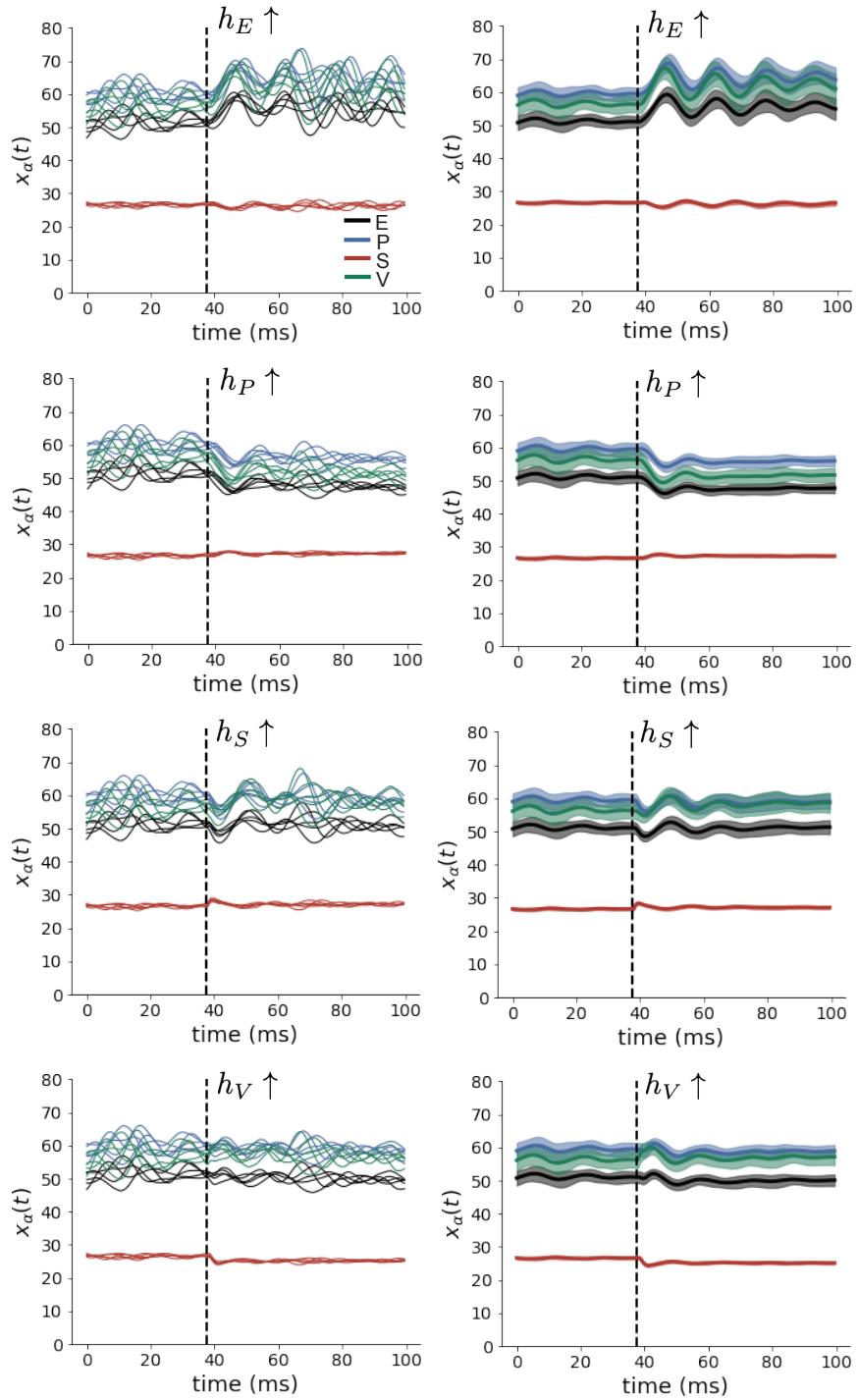


Figure 13: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

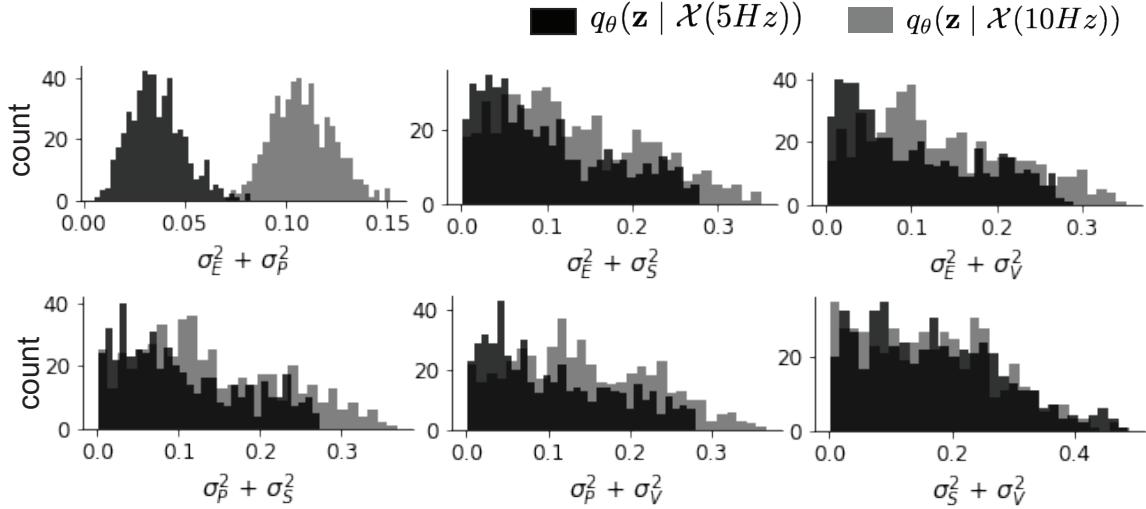


Figure 14: (V1 2) EPI predictive distributions of the sum of squares of each pair of noise parameters.

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (72)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (73)$$

1197 and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (74)$$

1198 Circuit responses are simulated using  $T = 200$  time steps at  $dt = 0.5\text{ms}$  from an initial condition  
 1199 drawn from  $\mathbf{x}(0) \sim U[10 \text{ Hz}, 25 \text{ Hz}]$ . Standard deviation of the E-population  $s_E(\mathbf{x}; \mathbf{z})$  is calculated  
 1200 as the square root of the temporal variance from  $t_{ss} = 75\text{ms}$  to  $Tdt = 100\text{ms}$  averaged over 100

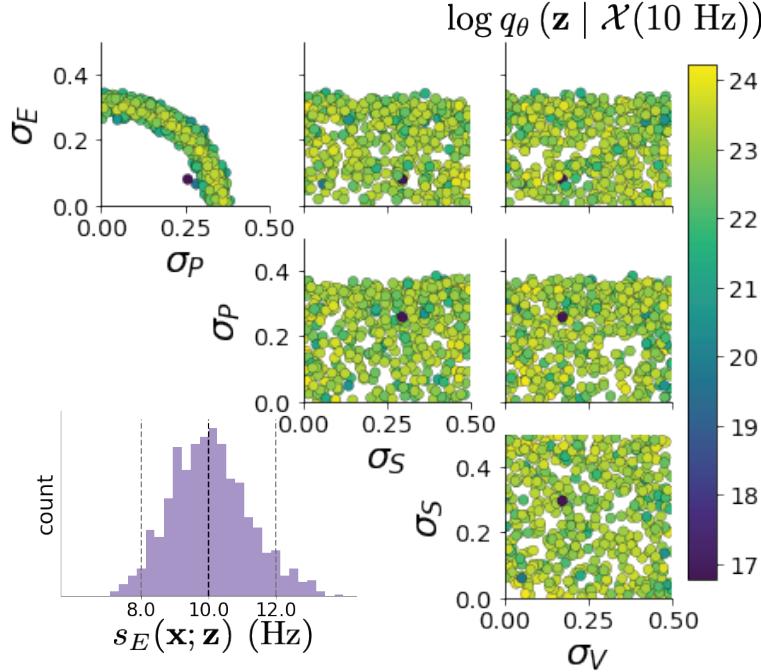


Figure 15: (V1 3) EPI inferred distribution for  $\mathcal{X}(10 \text{ Hz})$ .

1201 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[ \sqrt{\mathbb{E}_{t > t_{ss}} \left[ (x_E(t) - \mathbb{E}_{t > t_{ss}} [x_E(t)])^2 \right]} \right] \quad (75)$$

1202 For EPI in Fig 3D-E, we used a real NVP architecture with three Real NVP coupling layers  
 1203 and two-layer neural networks of 50 units per layer. The normalizing flow architecture mapped  
 1204  $z_0 \sim \mathcal{N}(\mathbf{0}, I)$  to a support of  $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V] \in [0.0, 0.5]^4$ . EPI optimization was run using three  
 1205 different random seeds for architecture initialization  $\theta$  with an augmented Lagrangian coefficient of  
 1206  $c_0 = 10^{-1}$ , a batch size  $n = 100$ , and  $\beta = 2$ . The distributions shown are those of the architectures  
 1207 converging with criteria  $N_{\text{test}} = 100$  at greatest entropy across random seeds.

1208 In Fig. 3E, we visualize the modes of  $q_\theta(\mathbf{z} | \mathcal{X})$  throughout the  $\sigma_E$ - $\sigma_P$  marginal. Specifically, we  
 1209 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} | \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (76)$$

1210 At each mode  $\mathbf{z}^*$ , we calculated the Hessian and visualized the sensitivity dimension in the direction  
 1211 of positive  $\sigma_E$ .

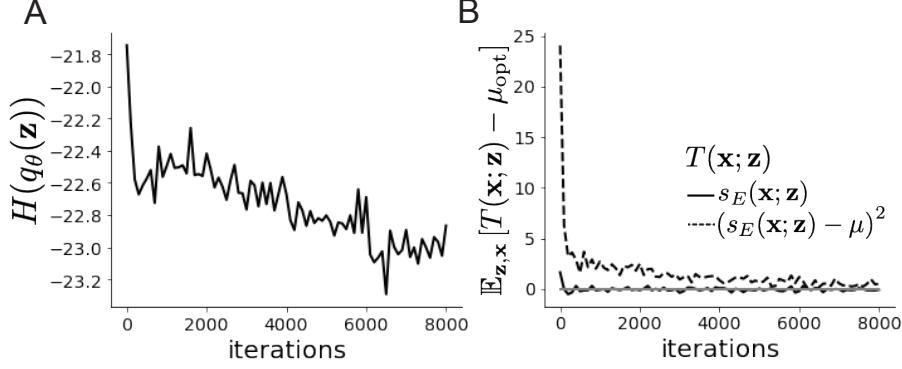


Figure 16: (V1 4) Optimization for V1

1212 **5.2.4 Primary visual cortex: challenges to analysis**

1213 TODO Agostina and I are putting this together now.

1214 **5.2.5 Superior colliculus**

1215 The ability to switch between two separate tasks throughout randomly interleaved trials, or “rapid  
1216 task switching,” has been studied in rats, and midbrain superior colliculus (SC) has been shown to  
1217 play an important role in this computation [79]. Neural recordings in SC exhibited two populations of  
1218 neurons that simultaneously represented both task context (Pro or Anti) and motor response (con-  
1219 tralateral or ipsilateral to the recorded side), which led to the distinction of two functional classes:  
1220 the Pro/Contra and Anti/Ipsi neurons [57]. Given this evidence, Duan et al. proposed a model  
1221 with four functionally-defined neuron-type populations: two in each hemisphere corresponding to  
1222 the Pro/Contra and Anti/Ipsi populations. We study how the connectivity of this neural circuit  
1223 governs rapid task switching ability.

1224 The four populations of this model are denoted as left Pro (LP), left Anti (LA), right Pro (RP)  
1225 and right Anti (RA). Each unit has an activity ( $x_\alpha$ ) and internal variable ( $u_\alpha$ ) related by

$$x_\alpha = \phi(u_\alpha) = \left( \frac{1}{2} \tanh \left( \frac{u_\alpha - a}{b} \right) + \frac{1}{2} \right), \quad (77)$$

1226 where  $\alpha \in \{LP, LA, RA, RP\}$ ,  $a = 0.05$  and  $b = 0.5$  control the position and shape of the nonlin-

1227 earity. We order the neural populations of  $x$  and  $u$  in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (78)$$

1228 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (79)$$

1229 with time constant  $\tau = 0.09s$ , step size 24ms and Gaussian noise  $d\mathbf{B}$  of variance  $0.2^2$ . These  
1230 hyperparameter values are motivated by modeling choices and results from [57].

1231 The weight matrix has 4 parameters for self  $sW$ , vertical  $vW$ , horizontal  $hW$ , and diagonal  $dW$   
1232 connections:

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (80)$$

1233 We study the role of parameters  $\mathbf{z} = [sW, vW, hW, dW]^\top$  in rapid task switching.

1234 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (81)$$

1235 There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (82)$$

1236 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (83)$$

1237 rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (84)$$

1238

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (85)$$

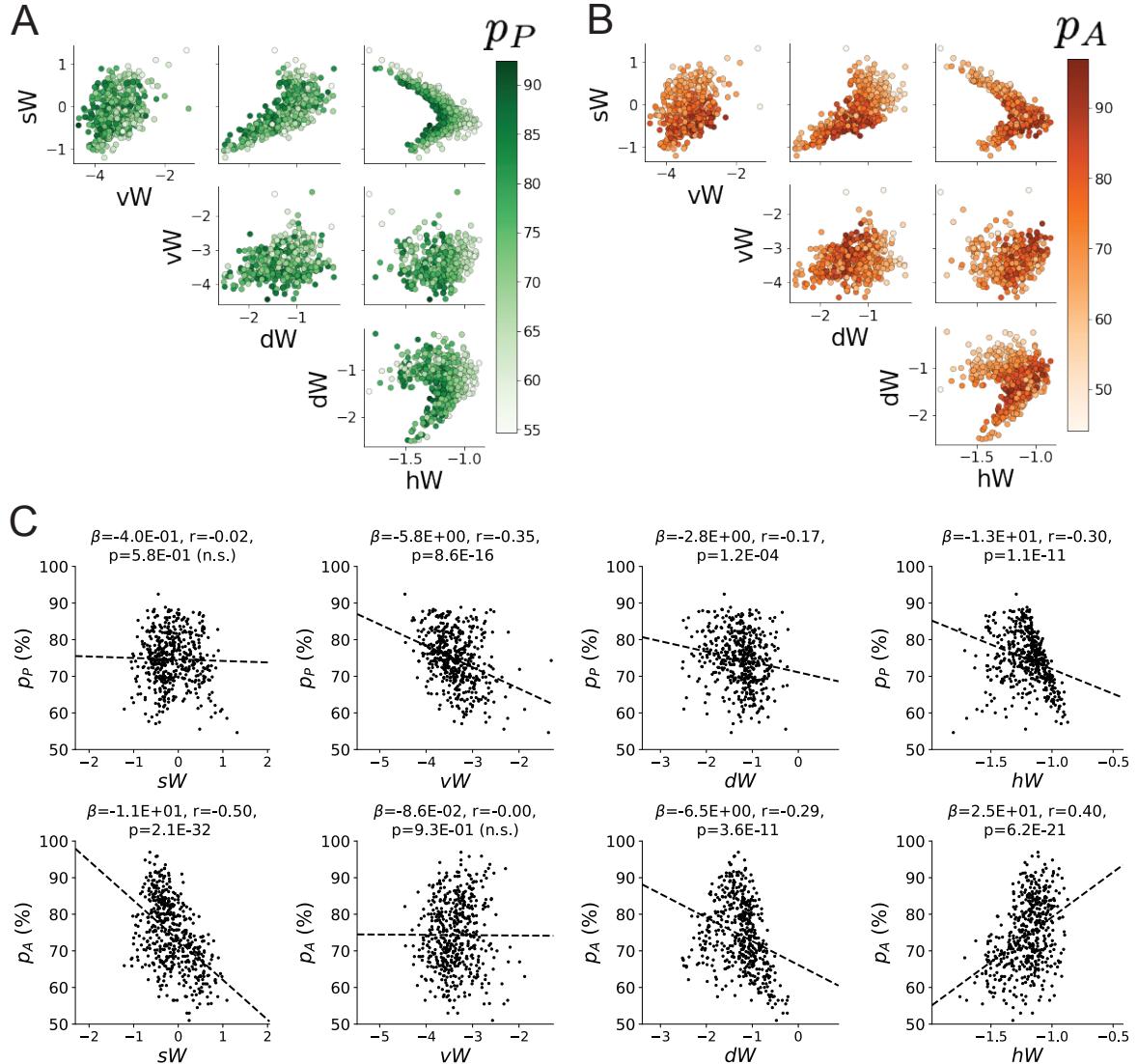


Figure 17: (SC1): **A.** Same pairplot as Fig. 4C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies.  $\beta$  is slope coefficient of linear regression,  $r$  is correlation, and  $p$  is the two-tailed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

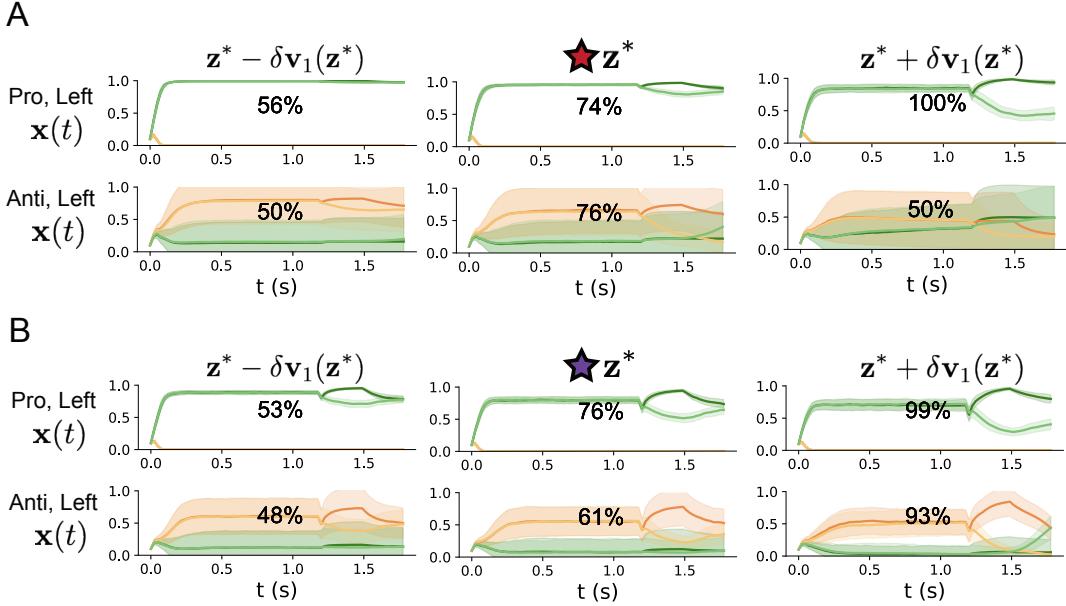


Figure 18: (SC2): **A.** Simulations in network regime 1 ( $hW_{\text{fixed}} = -1.2$ ) (center) with simulations given connectivity perturbations in the negative direction of the sensitivity vector  $\mathbf{v}_1$  (left) and positive direction (right). **B.** Same as A for network regime 2.

1239 a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^T, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (86)$$

1240 and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^T, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^T, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases}. \quad (87)$$

1241 The input parameterization was fixed to  $I_{\text{constant}} = 0.75$ ,  $I_{\text{P,bias}} = 0.5$ ,  $I_{\text{P,rule}} = 0.6$ ,  $I_{\text{A,rule}} = 0.6$ ,

1242  $I_{\text{choice}} = 0.25$ , and  $I_{\text{light}} = 0.5$ .

1243 The accuracies of each task  $p_P$  and  $p_A$  are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (88)$$

1244 and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (89)$$

1245 given that the stimulus is on the left side, where  $\Theta$  is the Heaviside step function, and the accuracy

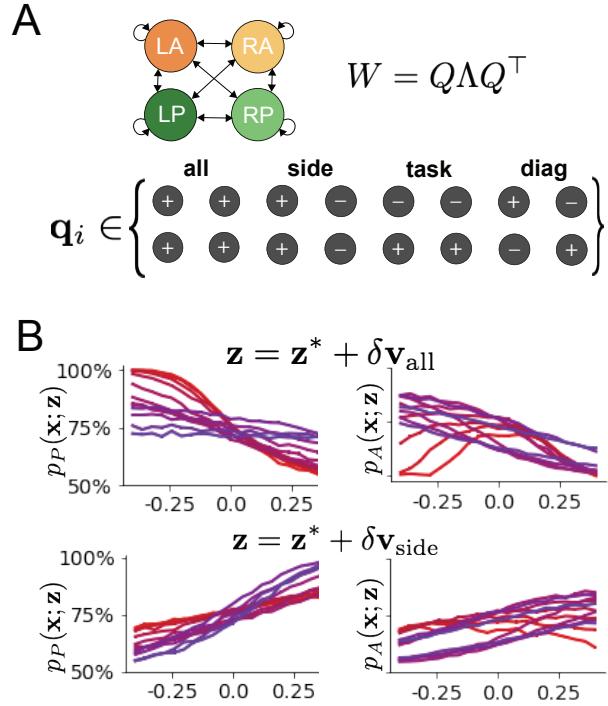


Figure 19: (SC3): **A.** Invariant eigenvectors of connectivity matrix  $W$ . **B.** Accuracies for connectivity perturbations for increasing  $\lambda_{\text{all}}$  and  $\lambda_{\text{side}}$  (rest shown in Fig. 4D).

<sub>1246</sub> is averaged over 200 independent trials. The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (90)$$

<sub>1247</sub> where  $\beta = 100$ .

<sub>1248</sub> Writing the EPI distribution as a maximum entropy distribution,  $T(\mathbf{x}, \mathbf{z})$  is comprised of both these  
<sub>1249</sub> first and second moments of the accuracy in each task (as in Equations 27 and 28)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \\ (p_P(\mathbf{x}; \mathbf{z}) - 75\%)^2 \\ (p_A(\mathbf{x}; \mathbf{z}) - 75\%)^2 \end{bmatrix}, \quad (91)$$

<sub>1250</sub>

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 7.5\%^2 \\ 7.5\%^2 \end{bmatrix}. \quad (92)$$

<sub>1251</sub> Throughout optimization, the augmented Lagrangian parameters  $\eta$  and  $c$ , were updated after each

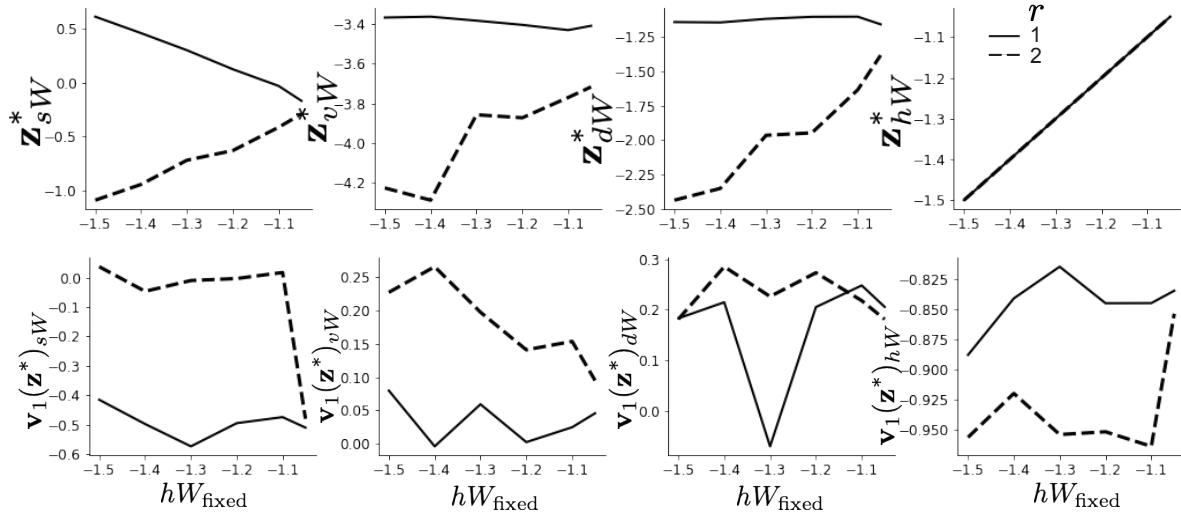


Figure 20: (SC4): **A.** The individual parameters of each mode throughout the two regimes. **B.** The individual sensitivities of parameters of each mode throughout the two regimes.

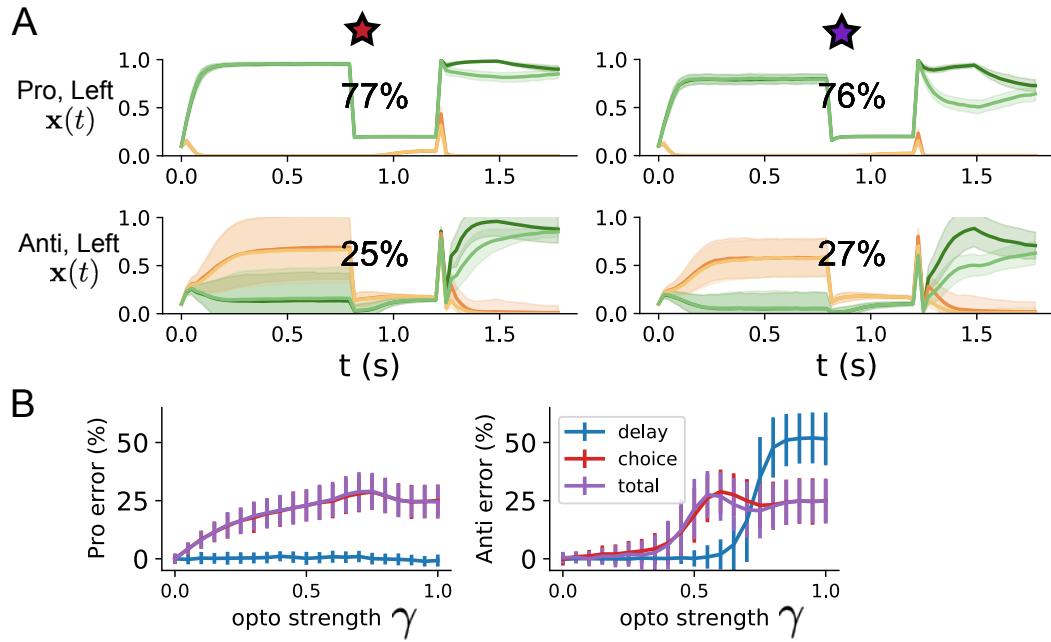


Figure 21: (SC5): **A.** Response of each parameter regime to optogenetic silencing during the delay period. **B.** Error induced by delay period inactivation with increasing optogenetic strength. Means and standard deviations are calculated across the entire EPI distribution.

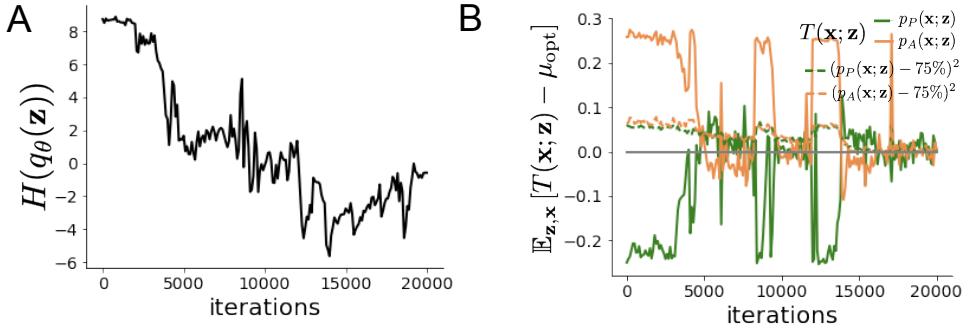


Figure 22: (SC6): **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

epoch of 2,000 iterations (see Section 5.1.3). The optimization converged after ten epochs (Fig. 22).

For EPI in Fig. 4C, we used a real NVP architecture with three coupling layers of affine transformations parameterized by two-layer neural networks of 50 units per layer. The initial distribution was a standard isotropic gaussian  $z_0 \sim \mathcal{N}(\mathbf{0}, I)$  mapped to a support of  $\mathbf{z}_i \in [-5, 5]$ . We used an augmented Lagrangian coefficient of  $c_0 = 10^2$ , a batch size  $n = 100$ , and  $\beta = 2$ . The distribution converged with criteria  $N_{\text{test}} = 25$ .

The EPI distribution of SC model connectivities producing rapid task switching has interesting structure. Throughout  $q_{\theta}(\mathbf{z} \mid \mathcal{X})$ , we see that the probability distribution is narrow in  $hW$  (Fig. 4C). This suggests that rapid task switching is sensitive to changes in  $hW$ , but this is only a single parameter. The local structure of the distribution varies across parameter space, and thus the nature in which parameter combinations affect rapid task switching. From visual inspection, we may hypothesize that there are two distinct regimes, most easily visualized in the  $sW$ - $hW$  marginal distribution: one where  $sW$  and  $hW$  are correlated for greater  $sW$  and one where  $sW$  and  $hW$  are anticorrelated for lesser  $sW$ .

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

1274 were always taken, and learning rates of  $2.5 \times 10^{-4}$  and  $5 \times 10^{-4}$  were used for regimes 1 and 2,  
 1275 respectively. Each of these modes is denoted  $\mathbf{z}^*(hW_{\text{fixed}}, r)$  for regime  $r \in \{1, 2\}$ .

1276 At each mode, we measure the sensitivity dimension (that of most negative eigenvalue in the Hessian  
 1277 of the EPI distribution)  $\mathbf{v}_1(\mathbf{z}^*)$ . To resolve sign degeneracy in eigenvectors, we chose  $\mathbf{v}_1(\mathbf{z}^*)$  to have  
 1278 negative element in  $hW$ . This tells us what parameter combination rapid task switching is most  
 1279 sensitive to at this parameter choice in the regime. We see that while the modes of each regime  
 1280 gradually converge to similar connectivities at  $hW_{\text{fixed}} = -1.05$  (Fig. 20 top), the sensitivity  
 1281 dimensions remain categorically different throughout the two regimes (Fig. 20 bottom). Only at  
 1282  $hW_{\text{fixed}} = -1.05$  is there a flip in sensitivity from regime 2 to regime 1 (in  $\mathbf{v}_1(\mathbf{z}^*)_{sW}$  and  $\mathbf{v}_1(\mathbf{z}^*)_{hW}$ ).  
 1283 There is thus some ambiguity regarding the “regime” of  $\mathbf{z}^*(-1.05, 2)$ , since the mode is derived  
 1284 from an initialization in regime 2, but has sensitivity like regime 1. We can consider this as an  
 1285 intermediate transitional region of parameter space between the two regimes. To emphasize this,  
 1286  $\mathbf{z}^*(-1.05, 1)$  and  $\mathbf{z}^*(-1.05, 2)$  have the same color.

1287 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-  
 1288 tion of the symmetric connectivity matrices  $W = Q\Lambda Q^{-1}$ , which results in the same basis vectors  
 1289  $\mathbf{q}_i$  for all  $W$  parameterized by  $\mathbf{z}$  (Fig. 19A). These basis vectors have intuitive roles in processing for  
 1290 this task, and are accordingly named the *all* eigenmode - all neurons co-fluctuate, *side* eigenmode  
 1291 - one side dominates the other, *task* eigenmode - the Pro or Anti populations dominate the other,  
 1292 and *diag* mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair.  
 1293 Due to the parametric structure of the connectivity matrix, the parameters  $\mathbf{z}$  are a linear function  
 1294 of the eigenvalues  $\boldsymbol{\lambda} = [\lambda_{\text{all}}, \lambda_{\text{side}}, \lambda_{\text{task}}, \lambda_{\text{diag}}]^\top$  associated with these eigenmodes.

$$\mathbf{z} = A\boldsymbol{\lambda} \quad (93)$$

1295

$$A = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}. \quad (94)$$

1296 We are interested in the effect of raising or lowering the amplification of each eigenmode in the  
 1297 connectivity matrix. To test this, we calculate the unit vector of changes in the connectivity  $\mathbf{z}$  that  
 1298 result from a change in the associated eigenvalues

$$\mathbf{v}_a = \frac{\frac{\partial \mathbf{z}}{\partial \lambda_a}}{\|\frac{\partial \mathbf{z}}{\partial \lambda_a}\|_2}, \quad (95)$$

1299 where

$$\frac{\partial \mathbf{z}}{\partial \lambda_a} = A \mathbf{e}_a, \quad (96)$$

1300 and e.g.  $\mathbf{e}_{\text{all}} = [1, 0, 0, 0]^\top$ . So  $\mathbf{v}_a$  is the normalized column of  $A$  corresponding to eigenmode  $a$ .

1301 While perturbations in the sensitivity dimension  $\mathbf{v}_1(\mathbf{z}^*)$  adapt with the mode  $\mathbf{z}^*$  chosen, perturba-  
1302 tions in  $\mathbf{v}_a$  for  $a \in \{\text{all}, \text{side}, \text{task}, \text{diag}\}$  are invariant to  $\mathbf{z}$  (Equation 96).

1303 To understand the connectivity mechanisms that distinguish these two regimes, we perturb connec-  
1304 tivity at each mode in dimensions that have well defined roles in processing for the Pro and Anti  
1305 tasks. A convenient property of this connectivity parameterization is that there are  $\mathbf{z}$ -invariant  
1306 eigenmodes of connectivity, whose eigenvalues (or degree of amplification) change with  $\mathbf{z}$ . These  
1307 eigenmodes have intuitive roles in processing in each task, and are accordingly named the *all*,  
1308 *side*, *task*, and *diag* eigenmodes (see Section 5.2.5). Furthermore, the parameter dimension  $\mathbf{v}_a$   
1309 ( $a \in \{\text{all}, \text{side}, \text{task}, \text{and diag}\}$ ) that increases the eigenvalue of connectivity  $\lambda_a$  is  $\mathbf{z}$ -invariant (un-  
1310 like the sensitivity dimension  $\mathbf{v}_1(\mathbf{z})$ ) and  $\mathbf{v}_a \perp \mathbf{v}_{b \neq a}$ . Thus, by changing the degree of amplification  
1311 of each processing mode by perturbing  $\mathbf{z}$  along  $\mathbf{v}_a$ , we can elicit the differentiating properties of  
1312 the two regimes.

1313 Through these connectivity perturbation analyses, we found that increasing  $\lambda_{\text{task}}$  strongly reduced  
1314 Pro accuracy in regime 1, yet strongly reduced Anti accuracy in regime 2. This suggests that  
1315 stronger task representations can inhibit both Pro and Anti task performance in different contexts.  
1316 Furthermore, changing  $\lambda_{\text{task}}$  in either direction decreases Anti performance in regime 1, showing  
1317 that Anti task performance in regime 1 is dependent on a specific level of task representation.  
1318 We also found that with increasing  $\lambda_{\text{diag}}$ , Pro accuracy increased in both regimes, but there were  
1319 opposite effects on Anti accuracy. In regime 1, stronger amplification of diagonal population pat-  
1320 terns decreased Anti accuracy, while in regime 2 accuracy increased. These findings give us an  
1321 understanding of the mechanistic differences in computation enabling rapid task switching in each  
1322 regime.

1323 We tested whether the inferred SC model connectivities could reproduce experimental effects of  
1324 optogenetic inactivation in rats [79]. During periods of simulated optogenetic inactivation, activity  
1325 was decreased proportional to the optogenetic strength  $\gamma$

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (97)$$

1326 Delay period inactivation was from  $0.8 < t < 1.2$ , choice period inactivation was for  $t > 1.2$  and  
1327 total inactivation was for the entire trial.