

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 model – and these  
33 equations are motivated by biophysics, neurophysiology, and other conceptual considerations [1,  
34 2, 3, 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 com-  
41 putational properties determine model parameters. Seminal examples of this gold standard (which  
42 often adopt approaches from statistical physics) include our field’s understanding of memory ca-  
43 pacity in associative neural networks [7], chaos and autocorrelation timescales in random neural  
44 networks [8], the paradoxical effect [9], and decision making [10]. Unfortunately, as circuit models  
45 include more biological realism, theory via analytical derivation becomes intractable. Still, we can  
46 gain insight into these complex models by identifying the distribution of parameters that produce  
47 computations. By solving the inverse problem in this way, scientific analysis of complex biologi-  
48 cally realistic models is made possible [11, 12, 13, 6, 14]. However, this work clarifies an important  
49 incongruity between this methodology and theoretical approaches to neuroscience.

50 A pivotal detail in theoretical neuroscience is the manner in which the computation is specified. One

51 common approach is to use data that is exemplary of that computation, so that the framework of  
52 statistical inference can be applied for parameter identification. While a host of suitable algorithms  
53 have been developed for inferring parameters in models of neural data [15, 16, 17, 18, 19, 20, 21,  
54 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36] (see review, [37]), the level of insight gained  
55 strongly depends on the quantity and quality of available data. Ultimately theoretical neuroscience  
56 is concerned with the computational properties – the *emergent phenomena* – of our models [7, 8,  
57 9, 10], not noisy observed datasets. To use the aforementioned inference paradigm, scientists must  
58 shoehorn such mathematical criteria into an artificial dataset compatible with existing statistical  
59 approaches.

60 In addition, another crucial challenge in neuroscientific modeling has been the inversion of bio-  
61 logically realistic or “mechanistic” models. Most neural circuit models in theoretical neuroscience  
62 are noisy systems of differential equations that can only be sampled or realized through forward  
63 simulation; they lack the explicit likelihood necessary for statistical inference. Therefore, the most  
64 popular approaches to theoretical inverse problems have been likelihood-free inference (LFI) meth-  
65 ods [38, 39], in which reasonable parameters are obtained via simulation and rejection. A flourishing  
66 new class of techniques [40, 41, 42] use deep learning to improve upon traditional LFI approaches.  
67 However, to use these methods in theoretical neuroscience, we still must represent computation with  
68 data in some way. Theorists are therefore barred from using the probabilistic modeling toolkit for  
69 science with circuit models, unless they reformulate their inverse problem into an empirical frame-  
70 work.

71 These challenges motivate the development of a novel inference framework called emergent property  
72 inference (EPI). As an adaption of variational inference [43], EPI infers parameter distributions  
73 that produce an emergent property: not a singular dataset, but a collection of datasets exhibiting  
74 some mathematical criteria. EPI constrains the predictions of the inferred parameter distribution  
75 to produce the emergent property, which requires a variant of probabilistic inference methods [44].  
76 Importantly, EPI uses deep learning to make rich, flexible approximations to the parameter distri-  
77 bution [45] that produces an emergent property. The structure captured by these deep probability  
78 distributions are scientifically valuable, revealing the sensitivity and robustness of the emergent  
79 property to different parameter combinations. Perhaps most powerfully, EPI facilitates inference  
80 in mechanistic models, allowing theorists to capture rich parametric structure in biologically real-  
81 istic models that is conditioned upon the emergent phenomena of interest.

82 Equipped with this method, we prove out the potential of EPI by demonstrating its capabilities and

83 presenting novel theoretical findings borne from its analysis. First, we show EPI’s ability to handle  
84 mechanistic models using a classic model of parametric degeneracy in biology: the stomatogastric  
85 ganglion [46, 47]. Then, we show EPI’s superior scaling properties by inferring connectivities of re-  
86 current neural networks (RNNs) that exhibit stable, yet amplified responses – a hallmark of neural  
87 responses throughout the brain [48, 49, 50]. In a model of primary visual cortex (V1) with different  
88 neuron-types, we show that the equation for excitatory variability become analytically intractable  
89 as more populations are added. Strikingly, the way in which noisy inputs across neuron-types  
90 governs excitatory variability is salient in the visualized structure of the EPI inferred parameter  
91 distribution. Finally, we investigated the possible connectivities of superior colliculus (SC) that al-  
92 low execution of different tasks on interleaved trials. EPI discovered a rich distribution containing  
93 two connectivity regimes with different solution classes. We quantified parametric sensitivity and  
94 robustness in each regime by simply querying the deep probability distribution learned by EPI,  
95 which produced a mechanistic understanding of cortical responses in each regime. Intriguingly,  
96 all inferred connectivities reproduced results from optogenetic inactivation experiments in this be-  
97 havioral paradigm – emergent phenomena that EPI was not conditioned upon. These theoretical  
98 insights afforded by EPI illustrate the value of deep inference for the interrogation of neural circuit  
99 models.

## 100 3 Results

### 101 3.1 Motivating emergent property inference of theoretical models

102 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-  
103 erty inference. First, one designs or chooses an existing model that, it is hypothesized, captures  
104 the computation of interest. To ground this process in a well-known example, consider the stom-  
105 atogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic  
106 muscle activation patterns for digestion [51]. Despite full knowledge of STG connectivity and a  
107 precise characterization of its rhythmic pattern generation, biophysical models of the STG have  
108 complicated relationships between circuit parameters and computation [46, 12]. A subcircuit model  
109 of the STG [47] is shown schematically in Figure 1A. The jagged connections indicate electrical cou-  
110 pling having electrical conductance  $g_{el}$ , smooth connections in the diagram are inhibitory synaptic  
111 projections having strength  $g_{synA}$  onto the hub neuron, and  $g_{synB} = 5nS$  for mutual inhibitory con-  
112 nections. 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 typical statistical inference, this is a dataset – either collected experimentally or constructed by scientists to fit this empirical paradigm. In EPI, a different approach is taken, in which we define an emergent property: a set of mathematical criteria to be obeyed by the datasets predicted by the inferred distribution. In the STG example, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model 1B. An interesting emergent property of this stochastic model is when the hub neuron fires at an intermediate frequency between the intrinsic spiking rates of the fast and slow populations. This emergent property is shown in Figure 1C at an average frequency of 0.55Hz.

Third, the model parameters producing these outputs are inferred. Most often, brute-force parameter sweeps or rejection sampling techniques [39] are used to identify parameters whose model simulations are close to data or some desired feature. In this last step lies the opportunity for a paradigmatic shift away from empirical data-oriented representations of model output. By precisely quantifying the emergent property of interest as a statistical feature of the model, we can infer a probability distribution over parameter configurations that produce this emergent property. This unlocks the deep probabilistic modeling toolkit for treating theoretical inverse problems.

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 producing the emergent property. In the STG model, this distribution can be specifically queried to reveal the prototypical parameter configuration for intermediate hub frequency (the mode; Figure 1E yellow star), and how it decays based on changes away from the mode. Indeed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity ( $v_1$ ) and degeneracy ( $v_2$ ) (Fig. 1E, arrows) agree with error contours (Fig. 1E contours) and have diminished or preserved hub frequency, respectively (Fig. 1F activity traces) (see Section 5.2.1).



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 posterior producing intermediate hub frequency. Samples are colored by log probability density. Contours of hub neuron frequency error are shown at levels of .525, .53, ... .575 Hz (dark to light gray away from mean). Dimension of sensitivity  $\mathbf{v}_1$  (solid) and degeneracy  $\mathbf{v}_2$ . **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.

143 **3.2 A deep generative modeling approach to emergent property inference**

144 Emergent property inference (EPI) formalizes the three-step procedure of the previous section with  
 145 deep probability distributions. First, as is typical, we consider the model as a coupled set of differ-  
 146 ential equations [47]. In the running STG example, the model activity  $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$  is  
 147 the membrane potential for each neuron, which evolves according to the biophysical conductance-  
 148 based equation:

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

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

152 Second, we define the emergent property, which as above is “intermediate hub frequency” (Figure  
 153 1C). Quantifying this phenomenon is straightforward: we stipulate that the hub neuron’s spiking  
 154 frequency – denoted  $\omega_{hub}(\mathbf{x})$  is close to a frequency of 0.55Hz. Mathematically, we achieve this  
 155 with two constraints: by fixing the mean hub frequency over the inferred parameter distribution of  
 156  $\mathbf{z}$  and its resulting simulations  $\mathbf{x}$  to 0.55Hz,

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

157 and requiring that the variance of the hub frequency over the produced simulations is small

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

158 This level of variance was chosen to be low enough to exclude the fast and slow frequencies of  
 159 the two populations, but large enough to allow structural examination of the inferred parameter  
 160 distribution. By constraining the means and variances of emergent property statistics over  $\mathbf{z}$  as  
 161 well as the stochasticity of  $\mathbf{x}$ , we can precisely control the behavior that the inferred distribution  
 162 that EPI infers. In general, an emergent property

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2 \quad (4)$$

163 defines a collection of datasets with a statistic  $f(\mathbf{x}; \mathbf{z})$  (which may be comprised of multiple statis-  
 164 tics) and the means  $\boldsymbol{\mu}$  and variances  $\boldsymbol{\sigma}^2$  of those statistics over the datasets. The choice of  $\boldsymbol{\sigma}^2$   
 165 predicates the degree of variability around the mean  $\boldsymbol{\mu}$  that is consistent with the emergent prop-  
 166 erty.

167 Third, we perform emergent property inference: we find a distribution over parameter configura-  
168 tions  $\mathbf{z}$ , and insist that samples from this distribution produce the emergent property; in other  
169 words, they obey the constraints introduced in Equation 4. This distribution will be chosen from a  
170 family of probability distributions  $\mathcal{Q} = \{q_{\theta}(\mathbf{z}) : \theta \in \Theta\}$ , defined by a deep neural network [45, 52, 53]  
171 (Figure 1D, EPI box). Deep probability distributions map a simple random variable  $\mathbf{z}_0$  through a  
172 deep neural network with weights and biases  $\theta$  to parameters  $\mathbf{z} = g_{\theta}(\mathbf{z}_0)$  to a suitable complicated  
173 distribution (see Section 5.1.2 for more details). Many distributions in  $\mathcal{Q}$  will respect the emergent  
174 property constraints, so we select the most random (or “entropic”) distribution, which is the same  
175 choice made in Bayesian inference (see Section 5.1.6). In EPI optimization, stochastic gradient  
176 steps in  $\theta$  are taken such that entropy is maximized, and the emergent property  $\mathcal{X}$  is produced  
177 (see Section 5.1) The inferred EPI distribution is denoted  $q_{\theta}(\mathbf{z} | \mathcal{X})$ , to emphasize that we have  
178 conditioned our parameter distribution on emergent property  $\mathcal{X}$ .

179 The major scientific value of EPI is in the rich, queryable structure of these deep probability  
180 distributions. The probabilities of  $q_{\theta}(\mathbf{z} | \mathcal{X})$  are the densities of these parameters in the distribution  
181 producing the emergent property. The greatest probabilities (the modes) indicate prototypical  
182 parameter configurations, and the manner in which probabilities change away from the modes shows  
183 how different parameter combinations preserve or diminish the emergent property. The dimensions  
184 of greatest sensitivity (e.g. Fig. 1E solid arrow) or degeneracy (e.g. Fig. 1E dashed arrow) can be  
185 measured directly from the second order derivative of  $\log q_{\theta}(\mathbf{z} | \mathcal{X})$  called the “Hessian.” Around  
186 the mode, eigenvalues of the Hessian are negative; probabilities decrease locally in all directions  
187 away from the mode. The eigenvector with most negative eigenvalue is the parameter combination  
188 causing probability to decrease the fastest, making it the most sensitive dimension. Likewise, the  
189 flattest eigenvector, corresponding to the least negative eigenvalue, points in the most degenerate  
190 dimension. Once an EPI distribution has been inferred, this second order derivative requires trivial  
191 computation (when correct architecture class is chosen, see Section 5.1.2).

192 In the following sections, we showcase the versatility of EPI for scientific analysis on three neural  
193 circuit models across ranges of biological realism, neural system function, and network scale. First,  
194 we demonstrate the superior scalability of EPI compared to alternative techniques by inferring high-  
195 dimensional distributions of RNN connectivities that exhibit amplified, yet stable responses. Also  
196 in this RNN example, we emphasize that EPI is the only technique that controls the predictions  
197 made by the inferred parameter distribution. Next, in a model of primary visual cortex [54,  
198 55], we show how to gain insight by comparing multiple inferred distributions. Finally, we used

199 EPI to capture subtle parametric structure allowing the mechanistic characterization of multiple  
 200 parametric regimes of superior colliculus activity in a model of task switching [56]. This work is  
 201 the first to produce this level of theoretical insight via the quantification and examination of the  
 202 intricate structure captured by deep probability distributions.

203 **3.3 Scaling inference of RNN connectivity with EPI**

204 Transient amplification is a hallmark of neural activity throughout cortex, and is often thought to  
 205 be intrinsically generated by recurrent connectivity in the responding cortical area [48, 49, 50]. It  
 206 has been shown that to generate such amplified, yet stabilized responses, the connectivity of RNNs  
 207 must be non-normal [57, 48], and satisfy additional constraints [58]. In theoretical neuroscience,  
 208 RNNs are optimized and then examined to show how dynamical systems could execute a given  
 209 computation [59, 60], but such biologically realistic constraints on connectivity are ignored during  
 210 optimization for practical reasons. In general, access to distributions of connectivity adhering to  
 211 theoretical criteria like stable amplification, chaotic fluctuations [8], or low tangling [61] would add  
 212 great scientific value and contextualization to existing research with RNNs. Here, we use EPI to  
 213 learn RNN connectivities producing stable amplification, and demonstrate the superior scalability  
 214 and efficiency of EPI to alternative approaches.

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

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

216 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  $\chi_{i,j}^{(U)}, \chi_{i,j}^{(V)} \sim$   
 217  $\mathcal{N}(0, 1)$ . 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 am-  
 218 plification. Two conditions are necessary and sufficient for RNNs to exhibit stable amplification  
 219 [58]:  $\text{real}(\lambda_1) < 1$  and  $\lambda_1^s > 1$ , where  $\lambda_1$  is the eigenvalue of  $W$  with greatest real part and  $\lambda^s$  is  
 220 the maximum eigenvalue of  $W^s = \frac{W+W^\top}{2}$ . RNNs with  $\text{real}(\lambda_1) = 0.5 \pm 0.5$  and  $\lambda_1^s = 1.5 \pm 0.5$   
 221 will be stable with modest decay rate ( $\text{real}(\lambda_1)$  close to its upper bound of 1) and exhibit modest  
 222 amplification ( $\lambda_1^s$  close to its lower bound of 1). EPI can naturally condition on this emergent  
 223 property

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

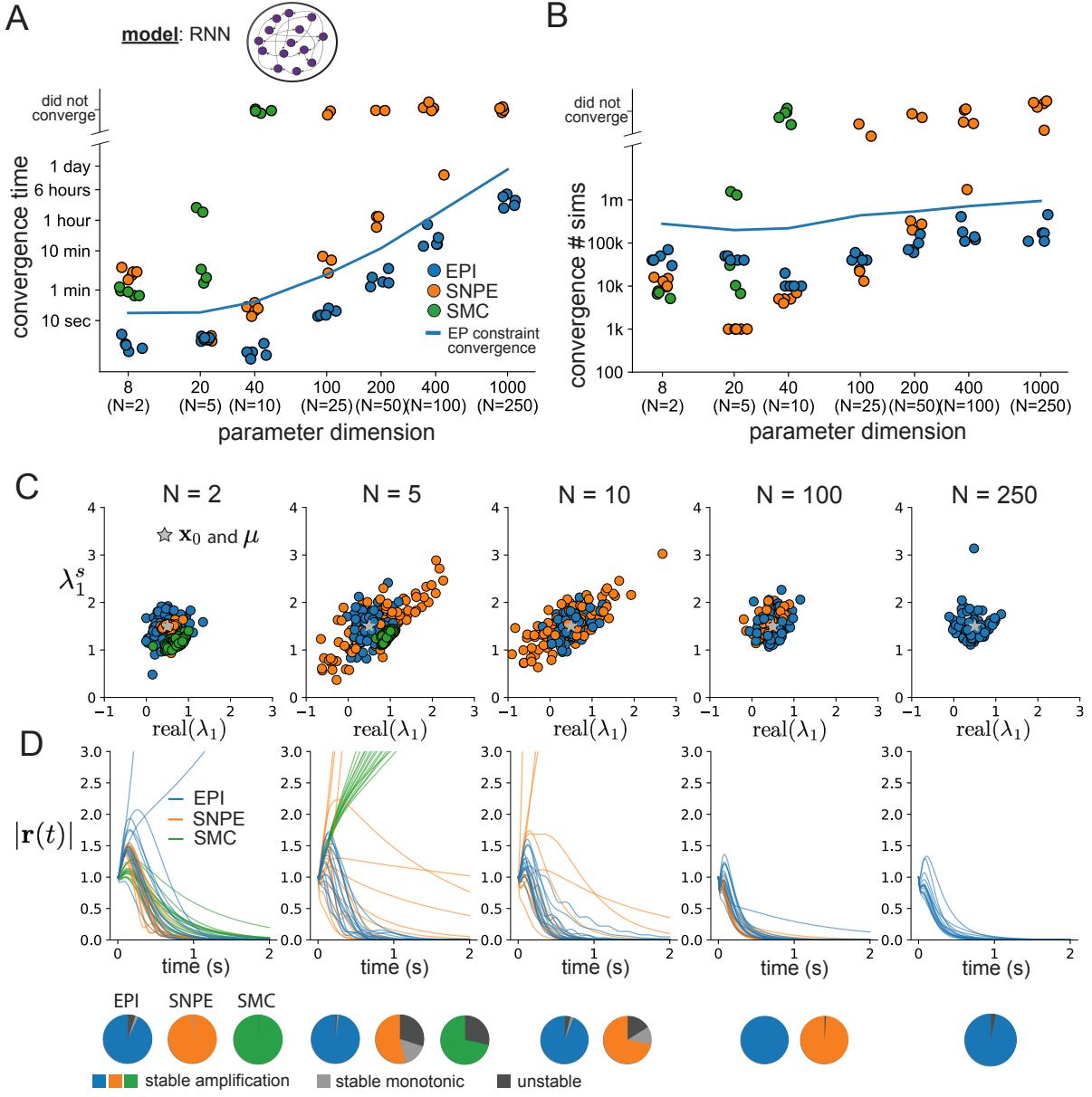


Figure 2: **A.** Wall time of EPI (blue), SNPE (orange), and SMC-ABC (green) to converge on RNN connectivities producing stable amplification. Each dot shows convergence time for an individual random seed. For reference, the mean wall time for EPI to achieve its full constraint convergence (means and 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.

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

For comparison, we infer the parameters  $\mathbf{z}$  likely to produce stable amplification using two alternative likelihood-free inference approaches. We ran sequential Monte Carlo approximate Bayesian computation (SMC-ABC) [38] and sequential neural posterior estimation (SNPE) [40] with observation  $\mathbf{x}_0 = \boldsymbol{\mu}$ . SMC-ABC is a rejection sampling approach that SMC techniques to improve efficiency, and SNPE approximates posteriors with deep probability distributions using a two-network architecture (see Section 5.1.1). Unlike EPI, these statistical inference techniques do not control the mean or variance of the predictive distribution, and these predictions of the inferred posteriors are typically affected by model characteristics (e.g.  $N$  and  $g$ , Fig. 10). To compare the efficiency of these different techniques, we measured the time and number of simulations necessary for the distance of the predictive mean to be less than 0.5 from  $\boldsymbol{\mu} = \mathbf{x}_0$  (see Section 5.2.2).

As the number of neurons  $N$  in the RNN are scaled, and thus the dimension of the parameter space  $\mathbf{z} \in [-1, 1]^{4N}$ , we see that EPI converges at greater speed and at greater dimension than SMC-ABC and SNPE (Fig. 2A). It also becomes most efficient to use EPI in terms of simulation count at  $N = 50$  (Fig. 2B). It is well known that ABC techniques struggle mightily in dimensions greater than about 30 [62], yet we were careful to assess the scalability of the more comparable approach SNPE. Between EPI and SNPE, we closely controlled the number of parameters in deep probability distributions by dimensionality (Fig. 9), and tested more aggressive SNPE hyperparameterizations when SNPE failed to converge (Fig. 11). From this analysis, we see that deep inference techniques EPI and SNPE are far more amenable to inference of high dimensional parameter distributions than rejection sampling techniques like SMC-ABC, and that EPI outperforms SNPE in both criteria in high-dimensions.

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

Through this example, we have shown that EPI can be used for well-controlled insight into RNNs

256 with respect to their theoretical properties. EPI outperforms SNPE in high dimensions by using  
 257 gradient information (from  $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z}) = \nabla_{\mathbf{z}}[\text{real}(\lambda_1), \lambda_1^s]^\top$ ) on each optimization iteration. This  
 258 agrees with recent speculation that such gradient information could improve the efficiency of LFI  
 259 techniques [85]. While scaling to high dimensions is important, we show in the next two sections  
 260 how insight can be gained by inspecting structure in lower dimensional parameter distributions.

261 **3.4 EPI reveals how neuron-type specific noise governs variability in the stochastic**  
 262 **stabilized supralinear network**

263 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output  
 264 function have succeeded in explaining a host of experimentally documented phenomena. In a  
 265 regime characterized by inhibitory stabilization of strong recurrent excitation, these models give  
 266 rise to paradoxical responses [9], selective amplification [57, 48], surround suppression [63] and  
 267 normalization [64]. Despite their strong predictive power, E-I circuit models rely on the assumption  
 268 that inhibition can be studied as an indivisible unit. However, experimental evidence shows  
 269 that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) –  
 270 composing 80% of GABAergic interneurons in V1 [65, 66, 67], and that these inhibitory cell types  
 271 follow specific connectivity patterns (Fig. 3A) [68]. While research has shown that V1 only shares  
 272 specific dimensions of neuronal variability with downstream areas [69], the role played by recurrent  
 273 dynamics and the connectivity across neuron-type populations is not understood. Here, in a model  
 274 of V1 with biologically realistic connectivity, we use EPI to show how the structure of input across  
 275 neuron types affects the variability of the excitatory population – the population largely responsible  
 276 for projecting to other brain areas [70].

277 We considered response variability of a nonlinear dynamical V1 circuit model (Fig. 3A) with a  
 278 state comprised of each neuron-type population’s rate  $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$ . Each population  
 279 receives recurrent input  $W\mathbf{x}$ , where  $W$  is the effective connectivity estimated from post-synaptic  
 280 potential and connectivity rate measurements (see Section 5.2.3). Each population also experiences  
 281 an external input  $\mathbf{h}$ , which determines population rate via supralinear nonlinearity  $\phi(\cdot) = [\cdot]_+^2$ . To  
 282 build on previous work, we model visual contrast-dependent input to the E- and P-populations  
 283  $\mathbf{h} = \mathbf{b} + c\mathbf{h}_c$ . There is also an additive noisy input  $\epsilon$  parameterized by variances for each neuron  
 284 type population  $\mathbf{z} = \sigma^2 = [\sigma_E^2, \sigma_P^2, \sigma_S^2, \sigma_V^2]$ . This noise has a slower dynamical timescale  $\tau_{\text{noise}} > \tau$   
 285 then the population rate, allowing fluctuations around a stimulus-dependent steady-state

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

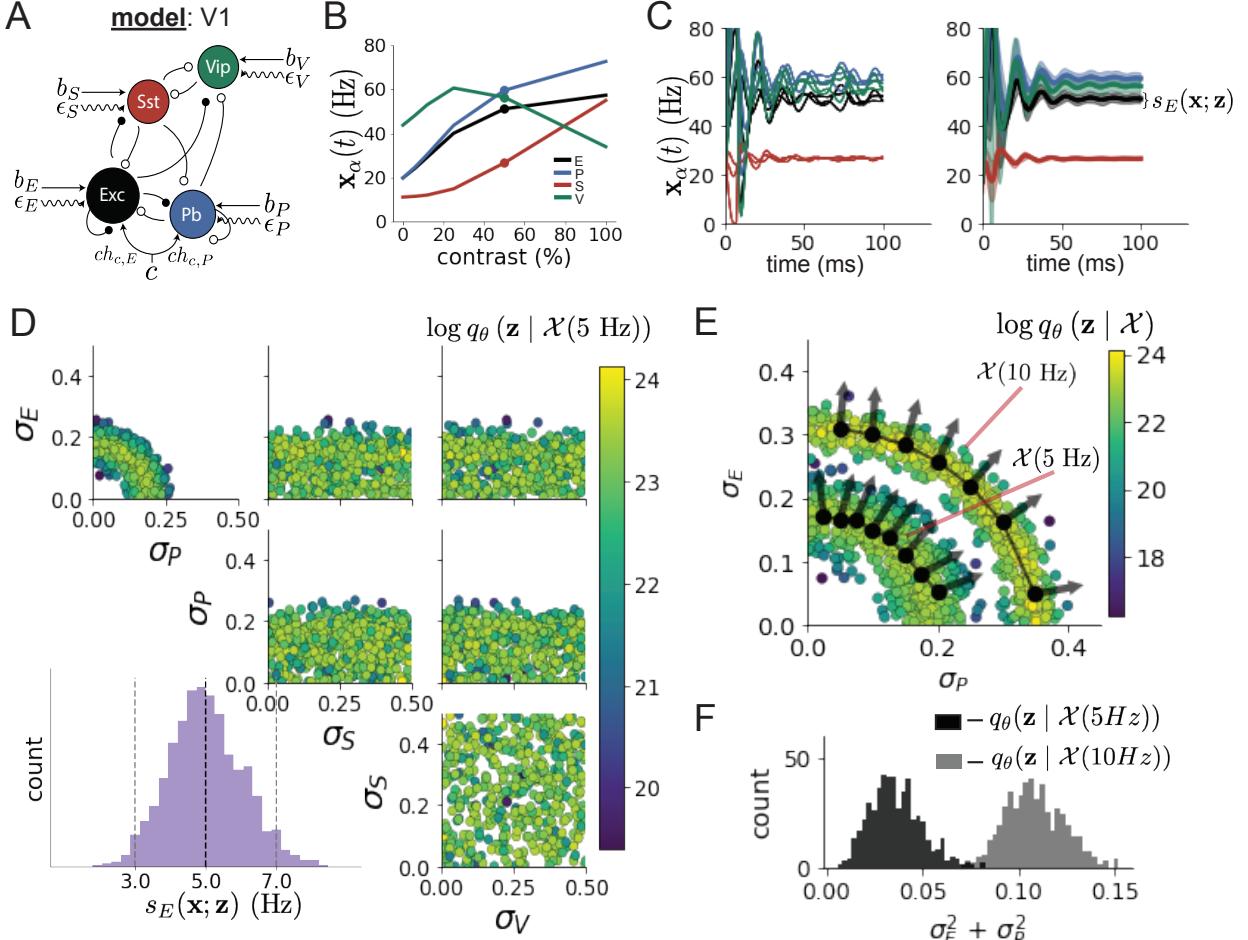


Figure 3: Emergent property inference in the stochastic stabilized supralinear network (SSSN) **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). 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.** Steady-state responses of the SSN model (deterministic,  $\sigma = \mathbf{0}$ ) to varying contrasts. The response at 50% contrast (dots) is the focus of our analysis. **C.** Transient network responses of the SSSN model at 50 % contrast. (Left) Traces are independent trials with varying initialization  $\mathbf{x}(0)$  and noise realization. (Right) Mean (solid line) and standard deviation (shading) of responses. **D.** EPI distribution of noise parameters  $\mathbf{z}$  conditioned on E-population variability. The EPI predictive distribution of  $s_E(\mathbf{x}; \mathbf{z})$  is show on the bottom-left. **E.** (Top) Enlarged visualization of the  $\sigma_E$ - $\sigma_P$  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 parameter distribution  $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$  and  $q_\theta(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$ .

286 This model is the stochastic stabilized supralinear network (SSSN) [71] generalized to have mul-  
 287 tiple inhibitory neuron types, and introduces stochasticity to previous four neuron-type models  
 288 of V1 [54]. Stochasticity and inhibitory multiplicity introduce substantial complexity to mathe-  
 289 matical derivations (see Section 5.2.4) motivating the treatment of this model with EPI. Here, we  
 290 consider fixed weights  $W$  and input  $\mathbf{h}$  [55] (Fig. 3B), and study the effect of input variability  
 291  $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$  on excitatory variability at 50% contrast.

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

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

293 where  $s_E(\mathbf{x}; \mathbf{z})$  is the standard deviation of the stochastic E-population response about its steady  
 294 state (Fig. 3C).

295 We ran EPI to obtain parameter distribution  $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$  producing E-population variability  
 296 around 5 Hz (Fig. 3D). From the marginal distribution of  $\sigma_E$  and  $\sigma_P$  (Fig. 3D, top-left), we can see  
 297 that  $s_E(\mathbf{x}; \mathbf{z})$  is sensitive to various combinations of  $\sigma_E$  and  $\sigma_P$ . Alternatively, both  $\sigma_S$  and  $\sigma_V$  are  
 298 degenerate with respect to  $s_E(\mathbf{x}; \mathbf{z})$  evidenced by the high variability in those dimensions (Fig. 3D,  
 299 bottom-right). Together, these observations imply a curved path of parametric degeneracy with  
 300 respect to  $s_E(\mathbf{x}; \mathbf{z})$  of 5 Hz, which is indicated by the modes along  $\sigma_P$  (Fig. 3E). The dimensions  
 301 of sensitivity conferred by EPI and this plain visual structure suggest a quadratic relationship  
 302 in the emergent property statistic  $s_E(\mathbf{x}; \mathbf{z})$  and parameters  $\mathbf{z}$ , which is preserved at a greater  
 303 level of variability  $\mathcal{X}(10 \text{ Hz})$  (Fig. 3E). Indeed, the sum of squares of  $\sigma_E$  and  $\sigma_P$  is larger in  
 304  $q_{\theta}(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$  than  $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$  (Fig 3F,  $p = 0$ ), while the sum of squares of  $\sigma_S$  and  $\sigma_V$  are  
 305 not significantly different in the two EPI distributions (Fig. 14,  $p = .402$ ). The strong compensatory  
 306 influence of the E- and P-population input variability on excitatory variability is intriguing, since  
 307 this circuit exhibited a paradoxical effect in the P-population (and no other inhibitory types) at  
 308 50% contrast (Fig. 14) meaning that the E-population is P-stabilized. Future research may uncover  
 309 a link between the populations of stabilizations and compensatory interactions governing excitatory  
 310 variability.

311 By defining a clear theoretical question and executing the EPI optimization, we used EPI to reveal  
 312 the quadratic relationship between  $s_E(\mathbf{x}; \mathbf{z})$  and  $\mathbf{z}$ . While this property is ultimately derivable,  
 313 we show that with each additional neuron-type population, the formula becomes quite unruly and  
 314 likely escapes comprehensible analysis in our case (see Section 5.2.4). This emphasizes the need

315 for streamlined methods for gaining understanding about theoretical models when mathematical  
316 analysis becomes prohibitive.

317 **Methods:**

318 Neural recordings in the midbrain superior colliculus (SC) exhibited two populations of neurons  
319 that simultaneously represented both task context (Pro or Anti) and motor response (contralateral  
320 or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons [56].

321 **3.5 EPI identifies multiple regimes of rapid task switching**

322 It has been shown that rats can learn to switch from one behavioral task to the next on randomly  
323 interleaved trials [72], and an important question is what types of neural connectivity allow this  
324 ability. In this experimental setup, rats were explicitly cued on each trial to either orient towards  
325 a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A) task  
326 (Fig. 4A). Neural recordings in superior colliculus (SC) exhibited two populations of neurons that  
327 represented task context (Pro or Anti). Furthermore, Pro/Anti neurons in each hemisphere were  
328 strongly correlated with the animal’s decision [56]. These results motivated a model of SC that is  
329 a four-population dynamical system with functionally-defined neuron-types. Here, our goal is to  
330 understand how connectivity in this circuit model governs the ability to switch tasks rapidly.

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

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

338 where  $\tau = 90\text{ms}$  and there is white noise of variance  $0.2^2$ . The model responds to the side with  
339 greater Pro neuron activation; e.g. the response is left if  $x_{LP} > x_{RP}$  at the end of the trial. Here,  
340 we use EPI to determine the network connectivity  $\mathbf{z} = [sW, vW, dW, hW]^\top$  that produces rapid  
341 task switching.

342 We define the computation of rapid task switching as accurate execution of each task. Inferred

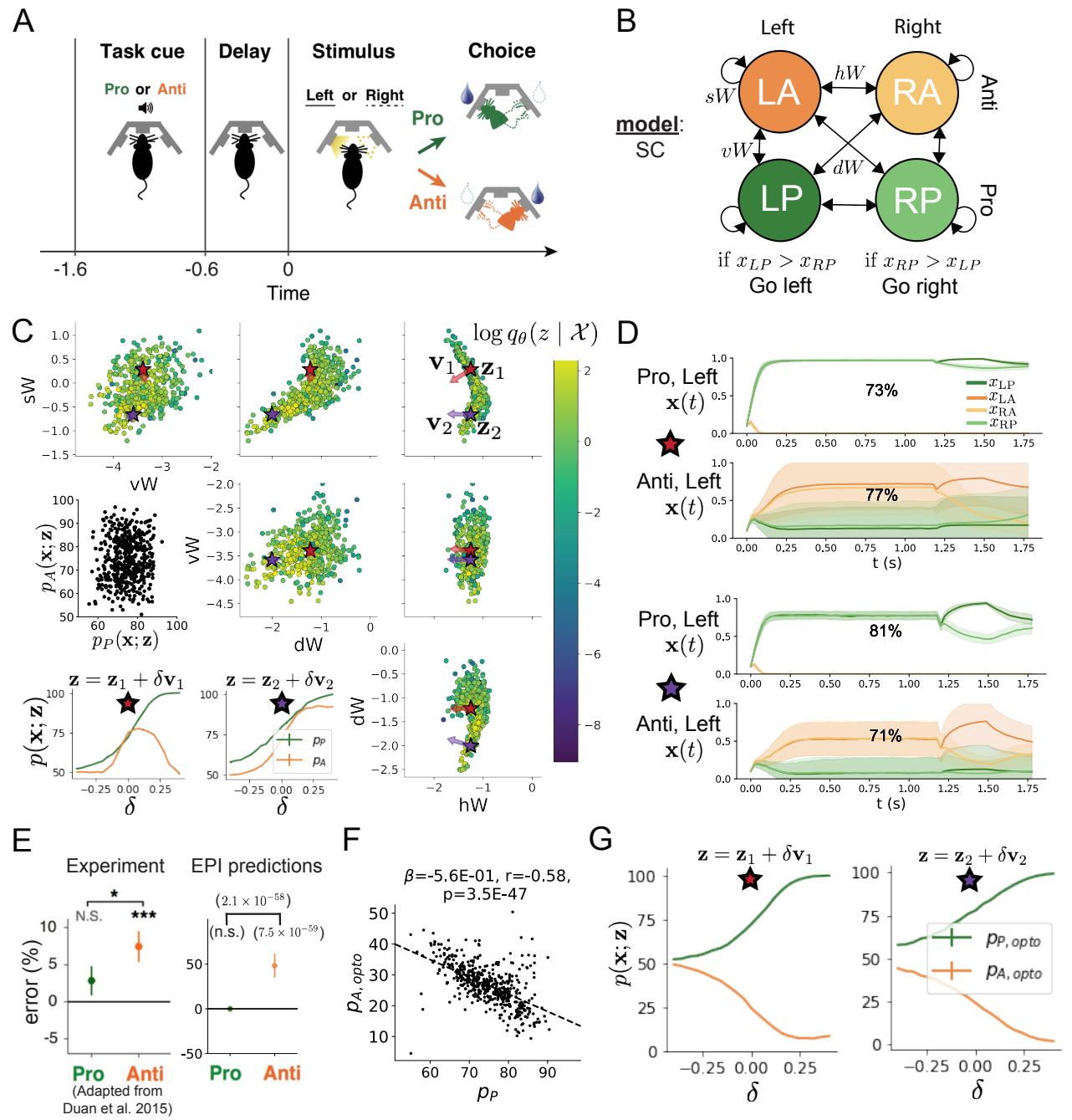


Figure 4: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters:  $sW$  - self,  $hW$  - horizontal,  $vW$  - vertical,  $dW$  - diagonal weights. **C.** The EPI inferred distribution of rapid task switching networks. Red and purple stars ( $\mathbf{z}_1$  and  $\mathbf{z}_2$ ) indicate different connectivity regimes with different sensitivity vectors  $\mathbf{v}_1$  and  $\mathbf{v}_2$ . (Middle-left) EPI predictive distribution of task accuracies. (Bottom-left) Task accuracy along dimensions of sensitivity in each connectivity regime. **D.** Means (solid) and standard deviations (shaded) of each population across random simulated trials. Top plots show Pro (top) and Anti (bottom) responses for connectivity  $\mathbf{z}_1$ . Bottom rows show the same  $\mathbf{z}_2$ . **E.** The EPI distribution predicts experimental results (left) showing no change in the Pro task, but larger error in the Anti task (right). **F.** Accuracy in the Anti task during delay period optogenetic inactivation  $p_{A, opto}$  is strongly anticorrelated with accuracy in the Pro task. **G.** Accuracy with delay period inactivation along each connectivity regime's dimension of sensitivity.

343 models should not exhibit fully random responses (50%), or perfect performance (100%), since  
 344 perfection is never attained by even the best trained rats. We formulate rapid task switching as an  
 345 emergent property by stipulating that the average accuracy in the Pro task  $p_P(\mathbf{x}; \mathbf{z})$  and Anti task  
 346  $p_A(\mathbf{x}; \mathbf{z})$  be 75% with variance  $7.5\%^2$ .

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 75\% \\ 75\% \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 7.5\%^2 \\ 7.5\%^2 \end{bmatrix} \end{aligned} \quad (10)$$

347 The EPI inferred distribution (Fig. 4C) produces task accuracies (Fig. 4C, middle-left) according  
 348 to our mathematical definition of rapid task switching (Equation 10). The patterns of connectivity  
 349 that govern each task accuracy are nonlinear (Fig. 16A-B); they are not captured well by linear  
 350 prediction (Fig. 16C). For example, the patterns in connectivity increasing Pro accuracy change  
 351 dramatically after crossing a threshold of  $sW$  (Fig. 16A  $sW-hW$  marginal distribution). Not only  
 352 has EPI captured this intricate, nonlinear posterior, it offers probabilistic tools for understanding  
 353 the different regimes of model behavior.

354 To establish these two regimes of connectivity, we took gradient steps along  $q_{\theta}(\mathbf{z} | \mathcal{X})$  to produce  
 355 modes  $\mathbf{z}_1$  and  $\mathbf{z}_2$  (Fig. 4C red and purple stars, Section 5.2.5). Simulations from these two regimes  
 356 reveal different responses in each task (Fig. 4D). We characterized these regimes by identifying  
 357 the dimensions of connectivity that rapid task switching is most sensitive to. The sensitivity  
 358 dimensions  $\mathbf{v}_1$  and  $\mathbf{v}_2$  (Fig. 4C, red and purple arrows) point in different directions, resulting in  
 359 different changes to task accuracy (Fig. 4D, bottom-left, Fig. 17). In regime 1, Anti accuracy  
 360 diminishes in either direction of sensitivity away from the mode, while in regime 2, Anti accuracy  
 361 matches monotonic increases in Pro accuracy.

362 To understand why these distinct connectivity regimes have different failure modes, we can analyze  
 363 the properties of connectivity in each regime. By taking the eigendecomposition of  $W$ , we can  
 364 quantify how strongly different modes of processing are reflected in the connectivity matrix (see  
 365 Section 5.2.5). **Note SB: I need to alter these plots a tad to confirm and support the**  
 366 **mechanistic explanation.**

367 During the delay period of this task, the circuit must prepare to execute the correct task based  
 368 on the cue input. Experimental results from Duan et al. found that optogenetic inactivation of  
 369 SC during the delay period consistently decreased performance in the Anti task, but had no effect

370 on the Pro task (Fig. 4E)). All network connectivities inferred by EPI exhibited this same effect,  
371 when network activities were silenced during the delay period (see Section 5.2.5). Notably, EPI  
372 inferred connectivities were only conditioned upon the emergent property of rapid task switching,  
373 not on Anti task failure during delay period silencing.

374 Similarities across Pro and Anti trials in choice period responses following delay period inactivation  
375 (Fig. 21A) suggested that connectivity patterns inducing greater Pro task accuracy increase error  
376 in delay period inactivated Anti trials (Fig. 4F). The strong anticorrelation between  $p_P$  and  $p_{A,opto}$   
377 across EPI inferred connectivities led to the following hypothesis about each connectivity regime:  
378 the sensitivity dimension of each regime decreases  $p_{A,opto}$  irrespective of its effect on  $p_A$ , since  
379 both  $\mathbf{v}_1$  and  $\mathbf{v}_2$  increase  $p_P$ . Indeed, in regimes 1 and 2 where sensitivity dimensions elicit different  
380 responses in  $p_A$ ,  $p_{A,opto}$  decreases since the connectivity changes enhancing  $p_P$  exacerbate Anti trial  
381 error (Fig. 4F). Thus, the altered state caused by delay period silencing makes the  $p_P$  connectivity  
382 better than the  $p_A$  down connectivity.

383 In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task  
384 switching. This EPI distribution revealed multiple regimes of rapid task switching, which we  
385 characterized using the probabilistic toolkit EPI seemlessly affords. EPI allowed us to conclude  
386 that since *all* parameters of this model producing rapid task switching make an experimentally  
387 verified prediction, the model is well-chosen in that regard. Finally, we used our knowledge about  
388 how  $\mathbf{z}$  governs  $p_{A,opto}$  to make accurate predictions about each identified regime of connectivity.

## 389 4 Discussion

390 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets  
391 [37]. Such careful inference procedures are developed for these statistical models allowing precise,  
392 quantitative reasoning, which clarifies the way data informs beliefs about the model parameters.  
393 However, these statistical models lack resemblance to the underlying biology, making it unclear  
394 how to go from the structure revealed by these methods, to the neural mechanisms giving rise  
395 to it. In contrast, theoretical neuroscience has focused on careful mechanistic modeling and the  
396 production of emergent properties of computation. The careful steps of *i.*) model design and  
397 *ii.*) emergent property definition, are followed by *iii.*) practical inference methods resulting in an  
398 opaque characterization of the way model parameters govern computation. In this work, we improve  
399 upon parameter inference techniques in theoretical neuroscience with emergent property inference,

400 harnessing deep learning towards careful inference in careful models of neural computation (see  
401 Section 5.1.1).

402 Specifically, approximate Bayesian computation [73, 74, 38] has been the standard approach to  
403 parameter inference in neural circuit models lacking tractable likelihoods. ABC methods do not  
404 confer probabilities on accepted parameters, require an acceptance threshold chosen to trade-off  
405 inference quality with tractability, do not scale efficiently to high-dimensional parameter spaces, and  
406 require independent techniques to analyze sensitivity for local parameter choices [62]. In contrast,  
407 EPI allows probability evaluations at any point in parameter space, conditions posteriors on the  
408 natural quantification of emergent properties, scales to high dimensional parameter spaces, and  
409 naturally admits sensitivity quantification via fast evaluations of the posterior Hessian.

410 Technically, EPI is a maximum entropy method, which learns parameter distributions that are  
411 as random as possible given that they produce the emergent property. Conceptually, maximally  
412 random distributions given some constraints are useful for understanding parametric sensitivity.  
413 This is well understood in Bayesian inference, where maximum entropy is the chosen normative  
414 principle. This is emphasized by an innovative formalism unifying top-down maximum entropy  
415 normative models with bottom-up statistical models [75]. Indeed, EPI is an adaptive variational  
416 inference program, and may be considered to have a Bayesian uniform prior (see Section 5.1.6).

417 Biologically realistic models of neural circuits often prove formidable to analyze for two main rea-  
418 sons. A primary challenge is that the number of parameters scales dramatically with the number of  
419 neurons, limiting analysis of its parameter space. We see in Section 3.3 that EPI scales seemlessly  
420 to high dimensional parameter spaces of RNN connectivities, while maintaining the production  
421 of the specified emergent property. EPI strongly outperforms the standard likelihood-free infer-  
422 ence technique (SMC-ABC [38]), and a recently developed deep likelihood-free inference technique  
423 (SNPE [40]), most likely because of it's ability to leverage the gradient information of the emer-  
424 gent property statistics and to adapt it's parameter sampling distribution at every step of gradient  
425 descent.

426 A secondary challenge is that the structure of the parametric regimes governing emergent properties  
427 is intricate. For example, even in low dimensional circuits, models can support more than one steady  
428 state [76] and non-trivial dynamics on strange attractors [77]. With EPI, we use deep probabillity  
429 distributions to capture the complex nonlinear parameter distributions governing model behavior.  
430 In Section 3.4, we used EPI to reveal a curved parametric manifolds governing curcuit variability  
431 in the stochastic stabilized supralinear network, and used hypothesis testing techniques to validate

432 our findings. In Section 3.5, we identified two regimes of SC connectivity resulting in rapid task  
433 switching, and found that the full distribution of rapid task switching networks reproduced an  
434 experimental result.

435 EPI leverages deep learning technology for neuroscientific inquiry in a categorically different way  
436 than approaches focused on training neural networks to execute behavioral tasks [78]. These works  
437 focus on examining optimized deep neural networks while considering the objective function, learn-  
438 ing rule, and architecture used. This endeavor efficiently obtains sets of parameters that can be  
439 reasoned about with respect to such considerations, but lacks the careful probabilistic treatment of  
440 parameter inference in EPI. All of these approaches can be used complementarily to enhance the  
441 practice of theoretical neuroscience.

442 **Acknowledgements:**

443 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-  
444 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-  
445 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-  
446 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program  
447 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James  
448 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

449 **Data availability statement:**

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

452 **Code availability statement:**

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

454 **References**

- 455 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern  
456 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 457 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,  
458 21(1):25–45, 1998.
- 459 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.

- 460 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in  
461 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 462 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R  
463 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology  
464 models. *PLoS Comput Biol*, 3(10):e189, 2007.
- 465 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network  
466 homeostasis, and pathological compensation from a biologically plausible ion channel expres-  
467 sion model. *Neuron*, 82(4):809–821, 2014.
- 468 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-  
469 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 470 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural  
471 networks. *Physical review letters*, 61(3):259, 1988.
- 472 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-  
473 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,  
474 17(11):4382–4388, 1997.
- 475 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration  
476 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 477 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley  
478 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 479 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit  
480 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 481 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron  
482 model. *PLoS computational biology*, 2(7):e94, 2006.
- 483 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar  
484 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 485 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,  
486 13(8):1713–1720, 2001.

- 487 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.  
488     A statistical paradigm for neural spike train decoding applied to position prediction from  
489     ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–  
490     7425, 1998.
- 491 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding  
492     models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 493 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.  
494     A point process framework for relating neural spiking activity to spiking history, neural  
495     ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 496 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correla-  
497     tions imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–  
498     1012, 2006.
- 499 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan  
500     Segev. A novel multiple objective optimization framework for constraining conductance-based  
501     neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 502 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature  
503     analysis. *Neural computation*, 19(4):1022–1038, 2007.
- 504 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and  
505     Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of  
506     neural population activity. In *Advances in neural information processing systems*, pages  
507     1881–1888, 2009.
- 508 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and  
509     Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural  
510     information processing systems*, 24:1350–1358, 2011.
- 511 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In  
512     *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 513 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-  
514     dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,  
515     9(3):e1002922, 2013.

- 516 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W  
517 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.  
518 *Science*, 349(6244):184–187, 2015.
- 519 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq  
520 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in  
521 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 522 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable  
523 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th In-*  
524 *ternational Conference on Machine Learning*, 2019.
- 525 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.  
526 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.  
527 *Nature Communications*, 10(4933), 2019.
- 528 [30] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical  
529 neural population models through nonlinear embeddings. In *Advances in neural information*  
530 *processing systems*, pages 163–171, 2016.
- 531 [31] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-  
532 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 533 [32] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-  
534 encoders. *bioRxiv*, page 399246, 2018.
- 535 [33] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,  
536 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R  
537 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-  
538 encoders. *Nature methods*, page 1, 2018.
- 539 [34] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M  
540 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.  
541 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 542 [35] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R  
543 Datta. Composing graphical models with neural networks for structured representations and  
544 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

- 545 [36] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,  
546 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:  
547 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural*  
548 *Information Processing Systems*, 2019.
- 549 [37] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-  
550 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,  
551 2018.
- 552 [38] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.  
553 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 554 [39] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.  
555 A framework for parameter estimation and model selection from experimental data in systems  
556 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 557 [40] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan  
558 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P  
559 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural  
560 dynamics. *bioRxiv*, page 838383, 2019.
- 561 [41] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast  
562 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on*  
563 *Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 564 [42] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized  
565 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–  
566 4248. PMLR, 2020.
- 567 [43] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural  
568 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 569 [44] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow  
570 networks. *International Conference on Learning Representations*, 2017.
- 571 [45] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.  
572 *International Conference on Machine Learning*, 2015.

- 573 [46] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-  
574 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,  
575 2001.
- 576 [47] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an  
577 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.  
578 *Neuron*, 77(5):845–858, 2013.
- 579 [48] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of  
580 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 581 [49] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dy-  
582 namics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–  
583 1406, 2014.
- 584 [50] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding  
585 and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.
- 586 [51] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-  
587 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 588 [52] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
589 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 590 [53] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for  
591 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,  
592 2017.
- 593 [54] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and  
594 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,  
595 115(3):1399–1409, 2016.
- 596 [55] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel  
597 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the  
598 operating regime of cortex. *bioRxiv*, 2020.
- 599 [56] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J  
600 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor  
601 routing. *bioRxiv*, page 245613, 2018.

- 602 [57] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,  
603 2009.
- 604 [58] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural  
605 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 606 [59] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in*  
607 *neurobiology*, 25:156–163, 2014.
- 608 [60] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*  
609 *opinion in neurobiology*, 46:1–6, 2017.
- 610 [61] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H  
611 Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex  
612 embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966,  
613 2018.
- 614 [62] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian com-*  
615 *putation*. CRC Press, 2018.
- 616 [63] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory  
617 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–  
618 592, 2009.
- 619 [64] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralin-  
620 ear network: a unifying circuit motif underlying multi-input integration in sensory cortex.  
621 *Neuron*, 85(2):402–417, 2015.
- 622 [65] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and  
623 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,  
624 5(10):793, 2004.
- 625 [66] Bernardo Rudy, Gordon Fishell, Soohyun Lee, and Jens Hjerling-Leffler. Three groups of  
626 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-*  
627 *robiology*, 71(1):45–61, 2011.
- 628 [67] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-  
629 tex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.

- 630 [68] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-  
631 bition of inhibition in visual cortex: the logic of connections between molecularly distinct  
632 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 633 [69] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn.  
634 Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.
- 635 [70] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate  
636 cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
- 637 [71] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D  
638 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-  
639 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 640 [72] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and  
641 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,  
642 2015.
- 643 [73] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-  
644 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 645 [74] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo  
646 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,  
647 2003.
- 648 [75] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical  
649 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.
- 650 [76] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give  
651 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of  
652 Sciences*, 115(13):3464–3469, 2018.
- 653 [77] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emerg-  
654 ent dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint  
655 arXiv:1605.04463*, 2016.
- 656 [78] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,  
657 2019.

- 658 [79] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,  
659 1957.
- 660 [80] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:  
661 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 662 [81] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and  
663 Edward Teller. Equation of state calculations by fast computing machines. *The journal of  
664 chemical physics*, 21(6):1087–1092, 1953.
- 665 [82] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.  
666 1970.
- 667 [83] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte  
668 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,  
669 73(2):123–214, 2011.
- 670 [84] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-  
671 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–  
672 820, 2011.
- 673 [85] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based infer-  
674 ence. *Proceedings of the National Academy of Sciences*, 2020.
- 675 [86] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-  
676 free variational inference. In *Advances in Neural Information Processing Systems*, pages  
677 5523–5533, 2017.
- 678 [87] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-  
679 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience  
680 Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 681 [88] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,  
682 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience  
683 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 684 [89] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and  
685 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In  
686 *Bernstein Conference 2018, Berlin, Germany*, 2018.

- 687 [90] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-  
688 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,*  
689 *Germany*, 2019.
- 690 [91] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-  
691 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural  
692 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 693 [92] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and  
694 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 695 [93] Sean R Bittner and John P Cunningham. Approximating exponential family models (not  
696 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 697 [94] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural  
698 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,  
699 2012.
- 700 [95] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary  
701 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,  
702 2018.
- 703 [96] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable  
704 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 705 [97] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula  
706 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially  
707 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–  
708 1929, 2009.
- 709 [98] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter  
710 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 711 [99] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of  
712 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 713 [100] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-  
714 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv*  
715 *preprint arXiv:1912.02762*, 2019.

- 716 [101] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-  
717 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 718 [102] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.  
719 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-*  
720 *tion processing systems*, 29:4743–4751, 2016.
- 721 [103] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Internation-*  
722 *al Conference on Learning Representations*, 2015.
- 723 [104] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for  
724 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- 725 [105] Emmanuel Klinger, Dennis Rickert, and Jan Hasenauer. pyabc: distributed, likelihood-free  
726 inference. *Bioinformatics*, 34(20):3591–3593, 2018.
- 727 [106] David S Greenberg, Marcel Nonnenmacher, and Jakob H Macke. Automatic posterior trans-  
728 formation for likelihood-free inference. *International Conference on Machine Learning*, 2019.

729 **5 Methods**

730 **5.1 Emergent property inference (EPI)**

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

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

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

751 Precisely, the emergent property statistics  $f(\mathbf{x}; \mathbf{z})$  must have means  $\boldsymbol{\mu}$  and variances  $\boldsymbol{\sigma}^2$  over the  
 752 EPI distribution of parameters and stochasticity of the data given the parameters.

753 In EPI, deep probability distributions are used as posterior approximations  $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X})$ . In deep  
 754 probability distributions, a simple random variable  $\mathbf{z}_0 \sim q_0(\mathbf{z}_0)$  is mapped deterministically via a  
 755 sequence of deep neural network layers  $(g_1, \dots, g_l)$  parameterized by weights and biases  $\boldsymbol{\theta}$  to the  
 756 support of the distribution of interest:

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

757 Such deep probability distributions embed the posterior distribution in a deep network. Once  
 758 optimized, this deep network representation has remarkably useful properties: immediate posterior  
 759 sampling, and immediate probability, gradient, and Hessian evaluation at any parameter choice.  
 760 Given a choice of model  $p(\mathbf{x} \mid \mathbf{z})$  and emergent property of interest  $\mathcal{X}$ ,  $q_{\theta}(\mathbf{z})$  is optimized via  
 761 the neural network parameters  $\theta$  to find a maximally entropic distribution  $q_{\theta}^*$  within the deep  
 762 variational family  $\mathcal{Q}$  producing the emergent property  $\mathcal{X}$ :

$$q_{\theta}(\mathbf{z} \mid \mathcal{X}) = q_{\theta}^*(\mathbf{z}) = \operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}(\mathbf{z})) \quad (13)$$

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

763 Entropy is chosen as the normative selection principle, since we want the posterior to only contain  
 764 structure predicated by the emergent property [79, 80]. This choice of selection principle is also  
 765 that of standard Bayesian inference, and we derive an exact relation between EPI and variational  
 766 inference (see Section 5.1.5). However, a key difference is that variational inference and other  
 767 Bayesian methods do not constrain the predictions of their inferred posteriors. This optimization  
 768 is executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [44].

769 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.  
 770 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-  
 771 ond, we describe the special class of deep probability distributions used in EPI called normalizing  
 772 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-  
 773 tion 13 (Section 5.1.3). Then, we demonstrate the details of this optimization in a toy example  
 774 (Section 5.1.4). Finally, we establish the known relationship between maximum entropy distribu-  
 775 tions and exponential families (Section 5.1.5), which is used to explain the relation between EPI  
 776 and variational inference (Section 5.1.6).

### 777 5.1.1 Related approaches

778 When Bayesian inference problems lack conjugacy, scientists use approximate inference methods  
 779 like variational inference (VI) [43] and Markov chain Monte Carlo (MCMC) [81, 82]. After opti-  
 780 mization, variational methods return a parameterized posterior distribution, which we can analyze.  
 781 Also, the variational approximating distribution class is often chosen such that it permits fast  
 782 sampling. In contrast MCMC methods only produce samples from the approximated posterior dis-  
 783 tribution. No parameterized distribution is estimated, and additional samples are always generated  
 784 with the same sampling complexity. Inference in models defined by systems of differential has been

785 demonstrated with MCMC [83], although this approach requires tractable likelihoods. Advances  
786 have leveraged structure in stochastic differential equation models to improve likelihood  
787 approximations, thus expanding the domain of applicable models [84].

788 Likelihood-free (or “simulation-based”) inference (LFI) [85] is model parameter inference in the  
789 absence of a tractable likelihood function. The most prevalent approach to LFI is approximate  
790 Bayesian computation [73], in which satisfactory parameter samples are kept from random prior  
791 sampling according to a rejection heuristic. The obtained set of parameters do not have a prob-  
792 abilities, and further insight about the model must be gained from examination of the parameter  
793 set and their generated activity. Methodological advances to ABC methods have come through  
794 the use of Markov chain Monte Carlo (MCMC-ABC) [74] and sequential Monte Carlo (SMC-ABC)  
795 [38] sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still  
796 struggles to scale in dimensionality (cf. Fig. 2). Furthermore, once a parameter set has been  
797 obtained by SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again  
798 with a new population of initialized particles to obtain additional samples.

799 For scientific model analysis, we seek a posterior distribution exhibiting the properties of a well-  
800 chosen variational approximation: a parametric form conferring analytic calculations, and trivial  
801 sampling time. For this reason, ABC and MCMC techniques are unattractive, since they only  
802 produce a set of parameter samples and have unchanging sampling rate. EPI executes likelihood-  
803 free inference using the MEFN [44] algorithm using a deep variational posterior approximation.  
804 The deep neural network of EPI defines the parametric form of the posterior approximation. Fur-  
805 thermore, the EPI distribution is constrained to produce an emergent property. In other words,  
806 the summary statistics of the posterior predictive distribution are fixed to have certain first and  
807 second moments. EPI optimization is enabled using stochastic gradient techniques in the spirit  
808 of likelihood-free variational inference [86]. The analytic relationship between EPI and variational  
809 inference is explained in Secton 5.1.6.

810 We note that, during our preparation and early presentation of this work [87, 88], another work  
811 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural  
812 circuits ([89, 90, 40]). We are encouraged by this general problem being recognized by others in the  
813 community, and we emphasize that these works offer complementary neuroscientific contributions  
814 (different theoretical models of focus) and use different technical methodologies (ours is built on  
815 our prior work [44], theirs similarly [91]).

816 The method EPI differs from SNPE in some key ways. SNPE belongs to a “sequential” class of

817 recently developed LFI methods in which two neural networks are used for posterior inference.  
818 This first neural network is a normalizing flow used to estimate the posterior  $p(\mathbf{z} | \mathbf{x})$  (SNPE)  
819 or the likelihood  $p(\mathbf{x} | \mathbf{z})$  (sequential neural likelihood (SNL [41])). A recent advance uses an  
820 unconstrained neural network to estimate the likelihood ratio (sequential neural ratio estimation  
821 (SNRE [42])). In SNL and SNRE, MCMC sampling techniques are used to obtain samples from  
822 the approximated posterior. This contrasts with EPI and SNPE, which afford a normalizing flow  
823 approximation to the posterior, which facilitates immediate measurements of sample probability,  
824 gradient, or Hessian for system analysis. The second neural network in this sequential class of  
825 methods is the amortizer. This network maps data  $\mathbf{x}$  (or statistics  $f(\mathbf{x}; \mathbf{z})$  or model parameters  $\mathbf{z}$ )  
826 to the weights and biases of the first neural network. These methods are optimized on a conditional  
827 density (or ratio) estimation objective on a sequentially adapting finite sample-based approximation  
828 to the posterior.

829 The approximating fidelity of the first neural network in sequential approaches is optimized to  
830 generalize across the entire distribution it is conditioned upon. This optimization towards gen-  
831 eralization of sequential methods can reduce the accuracy at the singular posterior of interest.  
832 Whereas in EPI, the entire expressivity of the normalizing flow is dedicated to learning a single  
833 distribution as well as possible. While amortization is not possible in EPI parameterized by the  
834 mean parameter  $\mu$  (due to the inverse mapping problem [92]), we have shown this two-network  
835 amortization approach to be effective in exponential family distributions defined by their natural  
836 parameterization [93].

837 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in  
838 natural models. Around a point, one can measure the Jacobian. One approach that scales well is  
839 EAR [94]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [95] and  
840 its stochastic adaptation [96]. Casting identifiability as a statistical estimation problem, the profile  
841 likelihood can assess via iterated optimization while holding parameters fixed [97]. An exciting  
842 recent method is capable of recovering the functional form of such unidentifiabilities away from a  
843 point by following degenerate dimensions of the fisher information matrix [98]. Global structural  
844 non-identifiabilities can be found for models with polynomial or rational dynamics equations using  
845 DAISY [99]. With EPI, we have all the benefits given by a statistical inference method plus the  
846 ability to query the gradient or Hessian of the inferred distribution at any chosen parameter value.

847 **5.1.2 Normalizing flows**

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

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

852 However, this computation has cubic complexity in dimensionality for fully connected layers. By  
 853 restricting our layers to normalizing flows [45, 100] – bijective functions with fast log determinant  
 854 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log  
 855 probability calculation confers efficient optimization of the maximum entropy objective (see Section  
 856 5.1.3). We use the Real NVP [52] normalizing flow class, because its coupling architecture confers  
 857 both fast sampling (forward) and fast log probability evaluation (backward). Fast probability  
 858 evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout  
 859 parameter space. Glow permutations were used in between coupling stages [101]. This is in contrast  
 860 to autoregressive architectures [53, 102], in which only forward or backward passes are efficient. In  
 861 this work, normalizing flows are used as flexible posterior approximations  $q_{\theta}(\mathbf{z})$  having weights and  
 862 biases  $\theta$ . We specify the architecture used in each application by the number of Real-NVP affine  
 863 coupling stages, and the number of neural network layers and units per layer of the conditioning  
 864 functions.

865 **5.1.3 Augmented Lagrangian optimization**

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

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

868 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  
 869 Lagrange multipliers where  $m = |\mu_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})| = 2|f(\mathbf{x}; \mathbf{z})|$ , and  $c$  is the penalty coefficient.  
 870 The sufficient statistics  $T(\mathbf{x}; \mathbf{z})$  and mean parameter  $\mu_{\text{opt}}$  are determined by the means  $\mu$  and  
 871 variances  $\sigma^2$  of emergent property statistics  $f(\mathbf{x}; \mathbf{z})$  defined in Equation 13. Specifically,  $T(\mathbf{x}; \mathbf{z})$  is  
 872 a concatenation of the first and second moments,  $\mu_{\text{opt}}$  is a concatenation of  $\mu$  and  $\sigma^2$  (see section  
 873 5.1.5), and the Lagrange multipliers are closely related to the natural parameters  $\eta$  of exponential

874 families (see Section 5.1.6). Weights and biases  $\boldsymbol{\theta}$  of the deep probability distribution are optimized  
875 according to Equation 15 using the Adam optimizer with learning rate  $10^{-3}$  [103].

876 To take gradients with respect to the entropy  $H(q_{\boldsymbol{\theta}}(\mathbf{z}))$ , it can be expressed using the reparam-  
877 eterization trick as an expectation of the negative log density of parameter samples  $\mathbf{z}$  over the  
878 randomness in the parameterless initial distribution  $q_0(\mathbf{z}_0)$ :

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

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

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

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

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

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

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

911 When assessing the suitability of EPI for a particular modeling question, there are some important  
 912 technical considerations. First and foremost, as in any optimization problem, the defined emergent  
 913 property should always be appropriately conditioned (constraints should not have wildly different  
 914 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution  
 915 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-  
 916 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,  
 917 one should consider the computational cost of the gradient calculations. In the best circumstance,  
 918 there is a simple, closed form expression (e.g. Section 5.2.2) for the emergent property statistic  
 919 given the model parameters. On the other end of the spectrum, many forward simulation iterations  
 920 may be required before a high quality measurement of the emergent property statistic is available  
 921 (e.g. Section 5.2.1). In such cases, backpropagating gradients through the SDE evolution will be  
 922 expensive.

#### 923 5.1.4 Example: 2D LDS

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

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

926 with

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

927 To run EPI with the dynamics matrix elements as the free parameters  $\mathbf{z} = [a_1, a_2, a_3, a_4]$  (fix-  
 928 ing  $\tau = 1$ ), the emergent property statistics  $T(\mathbf{x})$  were chosen to contain the first and second  
 929 moments of the oscillatory frequency,  $\frac{\text{imag}(\lambda_1)}{2\pi}$ , and the growth/decay factor,  $\text{real}(\lambda_1)$ , of the oscil-  
 930 lating system.  $\lambda_1$  is the eigenvalue of greatest real part when the imaginary component is zero, and

931 alternatively of positive imaginary component when the eigenvalues are complex conjugate pairs.  
 932 To learn the distribution of real entries of  $A$  that produce a band of oscillating systems around  
 933 1Hz, we formalized this emergent property as  $\text{real}(\lambda_1)$  having mean zero with variance  $0.25^2$ , and  
 934 the oscillation frequency  $2\pi\text{imag}(\lambda_1)$  having mean  $\omega = 1$  Hz with variance  $(0.1\text{Hz})^2$ :

$$\mathbb{E}[T(\mathbf{x})] \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}. \quad (20)$$

935

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

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

939 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  
 940 does not require simulation, we cannot derive the distribution  $q_{\theta}^*$  directly. This fact is due to the  
 941 formally hard problem of the backward mapping: finding the natural parameters  $\eta$  from the mean  
 942 parameters  $\boldsymbol{\mu}$  of an exponential family distribution [92]. Instead, we used EPI to approximate this  
 943 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two  
 944 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped  
 945 onto a support of  $z_i \in [-10, 10]$ . (see Section 5.1.2).

946 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the  
 947 parameter distribution. To validate our method, we analytically derived the contours of the prob-  
 948 ability density from the emergent property statistics and values. In the  $a_1$ - $a_4$  plane, the black  
 949 line at  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ , dotted black line at the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$ ,  
 950 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  
 951 of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired  
 952 statistical constraints and model degeneracy in the sum of  $a_1$  and  $a_4$ . Intuitively, the parameters  
 953 equivalent with respect to emergent property statistic  $\text{real}(\lambda_1)$  have similar log densities.

954 To explain the bimodality of the EPI distribution, we examined the imaginary component of  $\lambda_1$ .

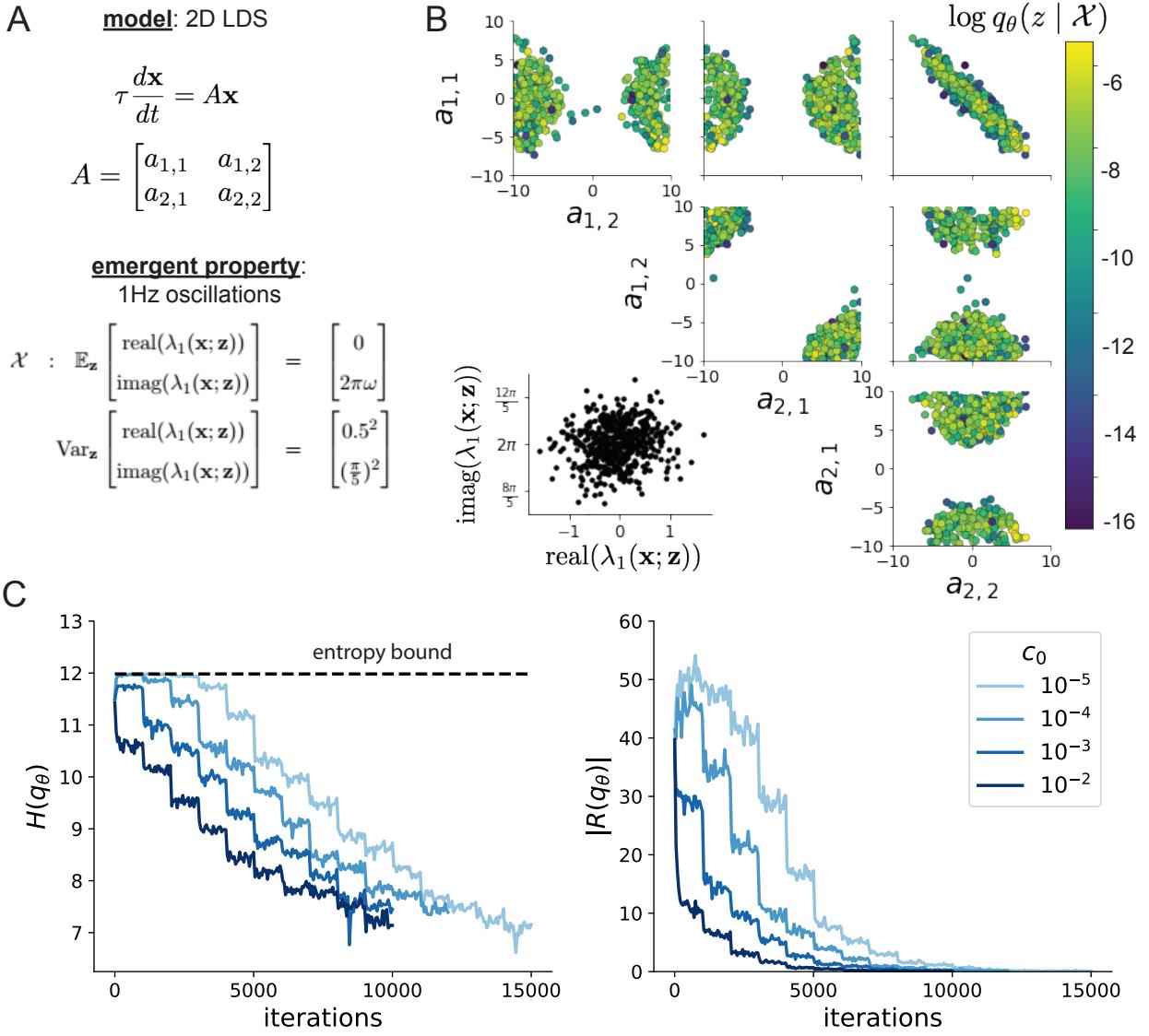


Figure 5: (LDS1): **A.** Two-dimensional linear dynamical system model, where real entries of the dynamics matrix  $A$  are the parameters. **B.** The EPI distribution for a two-dimensional linear dynamical system with  $\tau = 1$  that produces an average of 1Hz oscillations with some small amount of variance. Dashed lines indicate the parameter axes. **C.** Entropy throughout the optimization. At the beginning of each augmented Lagrangian epoch (2,000 iterations), the entropy dipped due to the shifted optimization manifold where emergent property constraint satisfaction is increasingly weighted. **D.** Emergent property moments throughout optimization. At the beginning of each augmented Lagrangian epoch, the emergent property moments adjust closer to their constraints.

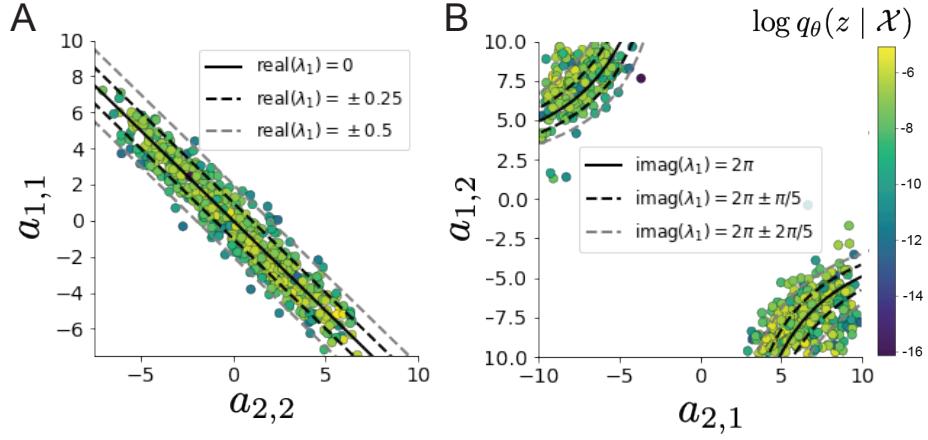


Figure 6: (LDS2): **A.** Probability contours in the  $a_1$ - $a_4$  plane were derived from the relationship to emergent property statistic of growth/decay factor  $\text{real}(\lambda_1)$ . **B.** Probability contours in the  $a_2$ - $a_3$  plane were derived from the emergent property statistic of oscillation frequency  $2\pi\text{imag}(\lambda_1)$ .

955 When  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ , we have

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

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

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

958 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  
959  $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$  for various  $a_1a_4$  (Fig. S2B). With  $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_\theta} [|a_1a_4 - E_{q_\theta}[a_1a_4]|]$ ,  
960 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  
961 dotted). This validates the curved structure of the inferred distribution learned through EPI. We  
962 took steps in negative standard deviation of  $a_1a_4$  (dotted and gray lines), since there are few positive  
963 values  $a_1a_4$  in the learned distribution. Subtler combinations of model and emergent property will  
964 have more complexity, further motivating the use of EPI for understanding these systems. As we  
965 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz  
966 (Fig. S3).

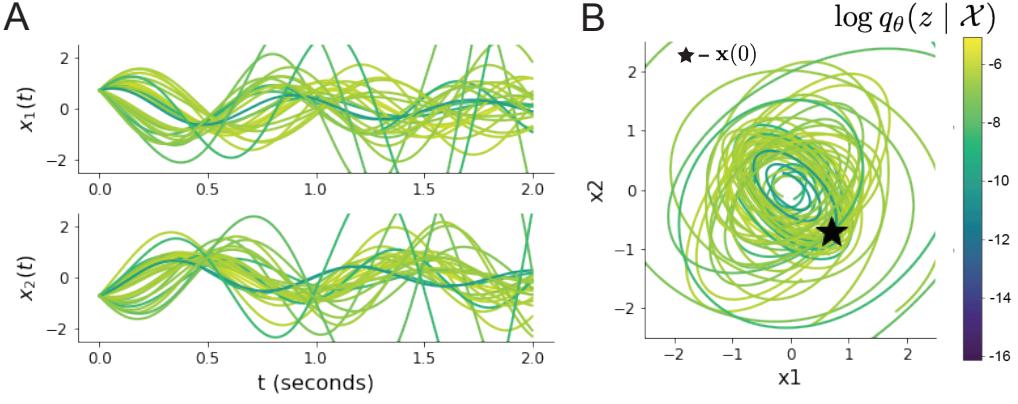


Figure 7: (LDS3): Sampled dynamical systems  $\mathbf{z} \sim q_\theta(\mathbf{z})$  and their simulated activity from  $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$  colored by log probability. **A.** Each dimension of the simulated trajectories throughout time. **B.** The simulated trajectories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

#### 967 5.1.5 Maximum entropy distributions and exponential families

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

$$\begin{aligned} p^*(\mathbf{z}) &= \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \\ \text{s.t. } \mathbb{E}_{\mathbf{z} \sim p}[T(\mathbf{z})] &= \boldsymbol{\mu}_{\text{opt}}. \end{aligned} \quad (24)$$

970 will have probability density in the exponential family:

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

971 The mappings between the mean parameterization  $\boldsymbol{\mu}_{\text{opt}}$  and the natural parameterization  $\boldsymbol{\eta}$  are  
972 formally hard to identify [92].

973 In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equation  
974 4

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (26)$$

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

$$\operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \mathbb{E}_{\mathbf{z}, \mathbf{x}}[(f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2] \quad (27)$$

976 As a general maximum entropy distribution (Equation 24), the sufficient statistics vector contains

977 both first and second order moments of  $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (28)$$

978 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (29)$$

### 979 5.1.6 EPI as variational inference

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

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \quad (30)$$

984 The posterior distribution is analytically available when the prior is conjugate with the likelihood.  
 985 However, conjugacy is rare in practice, and alternative methods, such as variational inference [104],  
 986 are utilized.

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

$$q_{\boldsymbol{\theta}}^*(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})). \quad (31)$$

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

$$KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})) = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] - \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{z} | \mathbf{x}))] \quad (32)$$

$$= -H(q_{\boldsymbol{\theta}}) - \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z})) - \log(p(\mathbf{x}))] \quad (33)$$

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

$$\underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}} || p(\mathbf{z} | \mathbf{x})) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmax}} H(q_{\boldsymbol{\theta}}) + \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z}))]. \quad (34)$$

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

$$p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) | \boldsymbol{\mu}_f, \Sigma_f), \quad (35)$$

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

$$\log p(\mathbf{x} | \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \quad (36)$$

998 where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\boldsymbol{\mu}_f}{\sigma_f^2} \\ \frac{-1}{2\sigma_f^2} \end{bmatrix}, \text{ and} \quad (37)$$

$$999 \quad T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \quad (38)$$

1000 The variational objective is then

$$\underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta) + \boldsymbol{\eta}_f^\top \mathbb{E}_{\mathbf{z} \sim q_\theta} [T(\mathbf{x}; \mathbf{z})] \quad (39)$$

1001 Comparing this to the Lagrangian objective (without augmentation) of EPI, we see they are the  
 1002 same

$$\begin{aligned} q_\theta^*(\mathbf{z}) &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top (\mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}_{\text{opt}}) \\ &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})]. \end{aligned} \quad (40)$$

1003 where  $T(\mathbf{x}; \mathbf{z})$  consists of the first and second moments of the emergent property statistic  $f(\mathbf{x}; \mathbf{z})$   
 1004 (Equation 28). Thus, EPI is implicitly executing variational inference with a uniform prior and a  
 1005 mean-field gaussian likelihood on the emergent property statistics. The data  $\mathbf{x}$  used by this implicit  
 1006 variational inference program would be that generated by the adapting variational approximation  
 1007  $\mathbf{x} \sim p(\mathbf{x} | \mathbf{z})q_\theta(\mathbf{z})$ , and the likelihood parameters  $\boldsymbol{\eta}_f$  of EPI optimization epoch  $k$  are predicated  
 1008 by  $\boldsymbol{\eta}_{\text{opt},k}$ . However, in EPI we have not specified a prior distribution, or collected data, which can  
 1009 inform us about model parameters. Instead we have a mathematical specification of an emergent  
 1010 property, which the model must produce, and a maximum entropy selection principle. Accordingly,  
 1011 we replace the notation of  $p(\mathbf{z} | \mathbf{x})$  with  $p(\mathbf{z} | \mathcal{X})$  conceptualizing an inferred distribution that obeys  
 1012 emergent property  $\mathcal{X}$  (see Section 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) [47] 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:

$$C_m \frac{dx_\alpha}{dt} = -[h_{leak}(\mathbf{x}; \mathbf{z}) + h_{Ca}(\mathbf{x}; \mathbf{z}) + h_K(\mathbf{x}; \mathbf{z}) + h_{hyp}(\mathbf{x}; \mathbf{z}) + h_{elec}(\mathbf{x}; \mathbf{z}) + h_{syn}(\mathbf{x}; \mathbf{z})] + dB. \quad (41)$$

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:

$$h_{leak}(\mathbf{x}; \mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (42)$$

$$h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{post} - x_\alpha^{pre}) \quad (43)$$

$$h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (44)$$

$$h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (45)$$

$$h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (46)$$

$$h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (47)$$

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:

$$M_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_\alpha - v_1}{v_2} \right) \right) \quad (48)$$

1041           
$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (49)$$

1042           
$$N_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_\alpha - v_3}{v_4} \right) \right) \quad (50)$$

1043           
$$\lambda_N = \phi_N \cosh \left( \frac{x_\alpha - v_3}{2v_4} \right) \quad (51)$$

1044           
$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (52)$$

1045           
$$H_\infty = \frac{1}{1 + \exp \left( \frac{x_\alpha + v_5}{v_6} \right)} \quad (53)$$

1046           
$$\tau_h = 272 - \left( \frac{-1499}{1 + \exp \left( \frac{-x_\alpha + v_7}{v_8} \right)} \right). \quad (54)$$

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:

$$S_\infty = \frac{1}{1 + \exp \left( \frac{v_{th} - x_\alpha}{v_9} \right)}. \quad (55)$$

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

$$\Phi = [0.0, 0.01, \dots, 1.0]^\top .. \quad (56)$$

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  $\alpha \in$

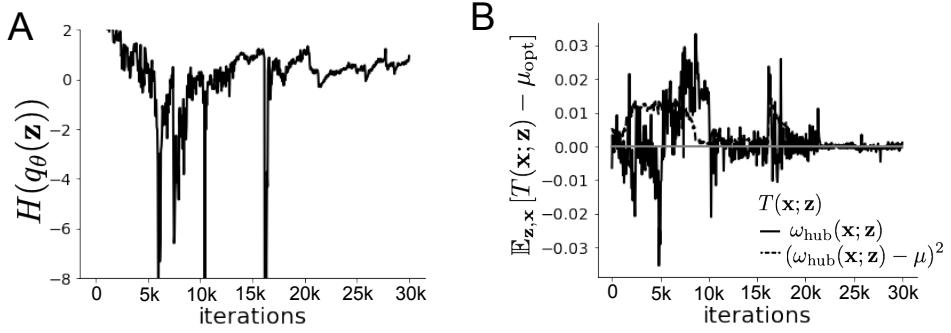


Figure 8: (STG1): EPI optimization of the STG model producing network syncing. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

1063  $\{f_1, f_2, \text{hub}, s_1, s_2\}$ . The soft-argmax is then calculated using temperature parameter  $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta|X_\alpha| \odot i), \quad (57)$$

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

$$\omega_\alpha = 0.01\psi_\alpha \text{Hz}. \quad (58)$$

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  
 1070 (as in Equations 28 and 29)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \quad (59)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \quad (60)$$

1071 Throughout optimization, the augmented Lagrangian parameters  $\eta$  and  $c$ , were updated after each  
 1072 epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

1073 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-  
 1074 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  $\mathbf{z}$ :  $\frac{\partial^2 \log q_{\boldsymbol{\theta}}(\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^\top}$ . 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

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}. \quad (61)$$

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 [38], and SNPE [40] (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 [58]. 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

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \lambda_1^s(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0.5)^2 \\ (\lambda_1^s(\mathbf{x}; \mathbf{z}) - 1.5)^2 \end{bmatrix}, \quad (62)$$

1109

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.25^2 \\ 0.25^2 \end{bmatrix}. \quad (63)$$

1110 Gradients of maximum eigenvalues of Hermitian matrices like  $W^s$  are available with modern auto-  
 1111 matic 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  $W^r = V^\top U$

$$\lambda_{\pm} = \frac{\text{Tr}(W^r) \pm \sqrt{\text{Tr}(W^r)^2 - 4\text{Det}(W^r)}}{2}. \quad (64)$$

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 likelihood-free inference (LFI) techniques, since the likelihood  
 1120 of these eigenvalues given  $\mathbf{z}$  is not available. Approximate Bayesian computation (ABC) [73] 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

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}. \quad (65)$$

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.

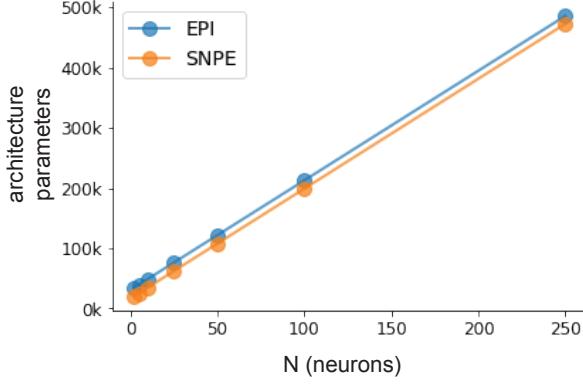


Figure 9: (RNN1): Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size ( $N$ ).

1128 SNPE, the next LFI approach in our comparison, is far more similar to EPI. Like EPI, SNPE  
 1129 treats parameters in mechanistic models with deep probability distributions, yet the two learning  
 1130 algorithms are categorically different. SNPE uses a two-network architecture to approximate the  
 1131 posterior distribution of the model conditioned on observed data  $\mathbf{x}_0$ . The amortizing network maps  
 1132 observations  $\mathbf{x}_i$  to the parameters of the deep probability distribution. The weights and biases of the  
 1133 parameter network are optimized by sequentially augmenting the training data with additional pairs  
 1134  $(\mathbf{z}_i, \mathbf{x}_i)$  based on the most recent posterior approximation. This sequential procedure is important  
 1135 to get training data  $\mathbf{z}_i$  to be closer to the true posterior, and  $\mathbf{x}_i$  to be closer to the observed data.  
 1136 For 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  
 1138 to the support as in EPI. This architectural choice closely tracked the size of the architecture used  
 1139 by EPI (Fig. 9). As in SMC-ABC, we ran SNPE with  $\mathbf{x}_0 = \mu$ . All SNPE optimizations were run  
 1140 for a limit of 1.5 days on a Tesla V100 GPU, or until two consecutive rounds resulted in a validation  
 1141 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. 10A). 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. 10B top, orange). At both levels of  $g$  when  $N = 2$ , the posterior of SNPE  
 1148 has lower entropy than EPI at convergence (Fig. 10B top). However at  $N = 10$ , SNPE results in

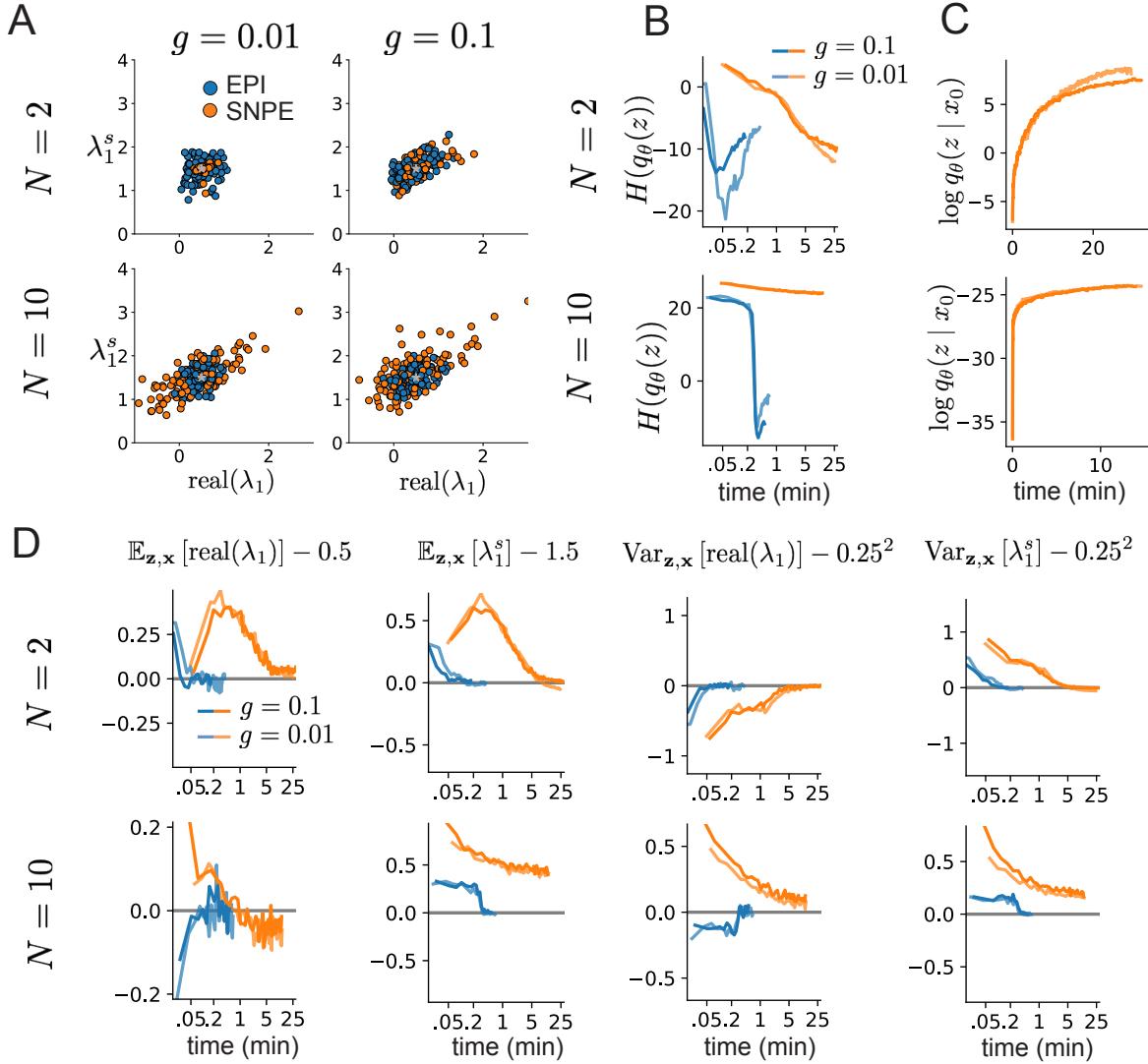


Figure 10: (RNN2): Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A.** Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with  $N = 2$  (top),  $N = 10$  (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.

1149 a predictive distribution of more widely dispersed eigenvalues (Fig. 10A bottom), and an inferred  
1150 posterior with greater entropy than EPI (Fig. 10B 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. 10C), and that EPI converges after several epochs of its own optimization  
1155 to enforce the emergent property constraints (Fig. 10D 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. 10D 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  
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. 10D 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

$$d(q_\theta(z)) = |\mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}|_2 \quad (66)$$

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-  
1170 izations 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. 11A). 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

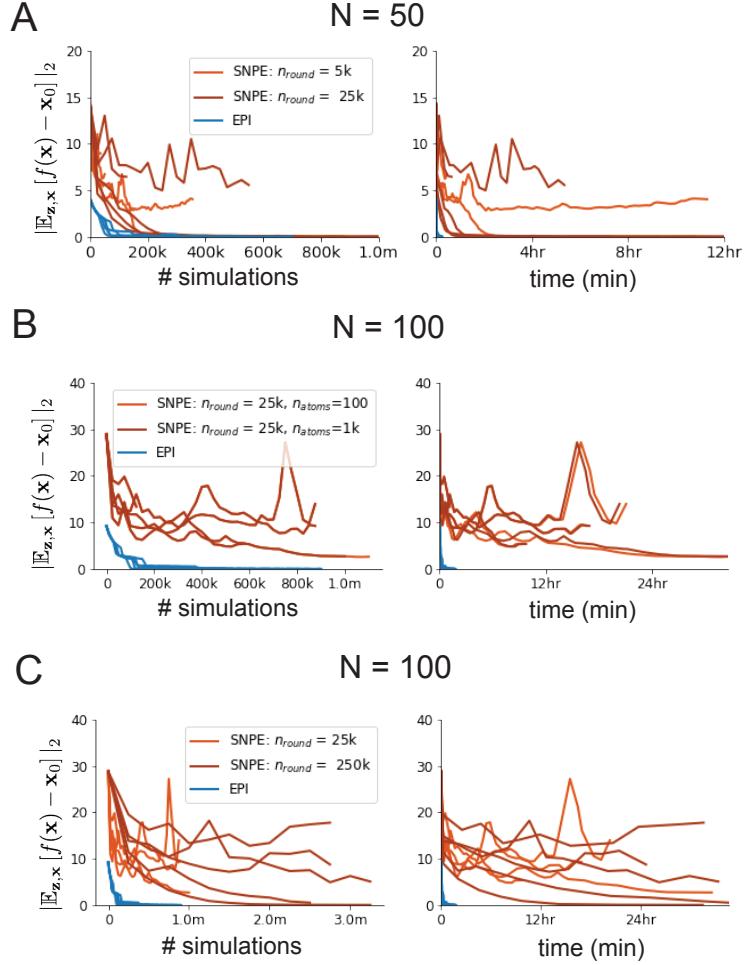


Figure 11: (RNN3): SNPE convergence was enabled by increasing  $n_{\text{round}}$ , not  $n_{\text{atom}}$ . **A.** Difference of mean predictions  $\mathbf{x}_0$  throughout optimization at  $N = 50$  with by simulation count (left) and wall time (right) of SNPE with  $n_{\text{round}} = 5,000$  (light 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).

1179 seeds) (Fig. 11B). 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).

1185 **5.2.3 Primary visual cortex**

1186 In the stochastic stabilized supralinear network [71], population rate responses  $\mathbf{x}$  to input  $\mathbf{h}$ , recur-  
 1187 rent input  $W\mathbf{x}$  and slow noise  $\epsilon$  are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \epsilon), \quad (67)$$

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

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (68)$$

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

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (69)$$

1190 so that  $\sigma$  parameterizes the variance of the noisy input in the absence of recurrent connectivity  
 1191 ( $W = \mathbf{0}$ ). As contrast increases, input to the E- and P-populations increases relative to a baseline  
 1192 input  $\mathbf{h} = \mathbf{h}_b + c\mathbf{h}_c$ . Connectivity ( $W_{\text{fit}}$ ) and input ( $\mathbf{h}_{b,\text{fit}}$  and  $\mathbf{h}_{c,\text{fit}}$ ) parameters were fit using the  
 1193 deterministic V1 circuit model [55]

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (70)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (71)$$

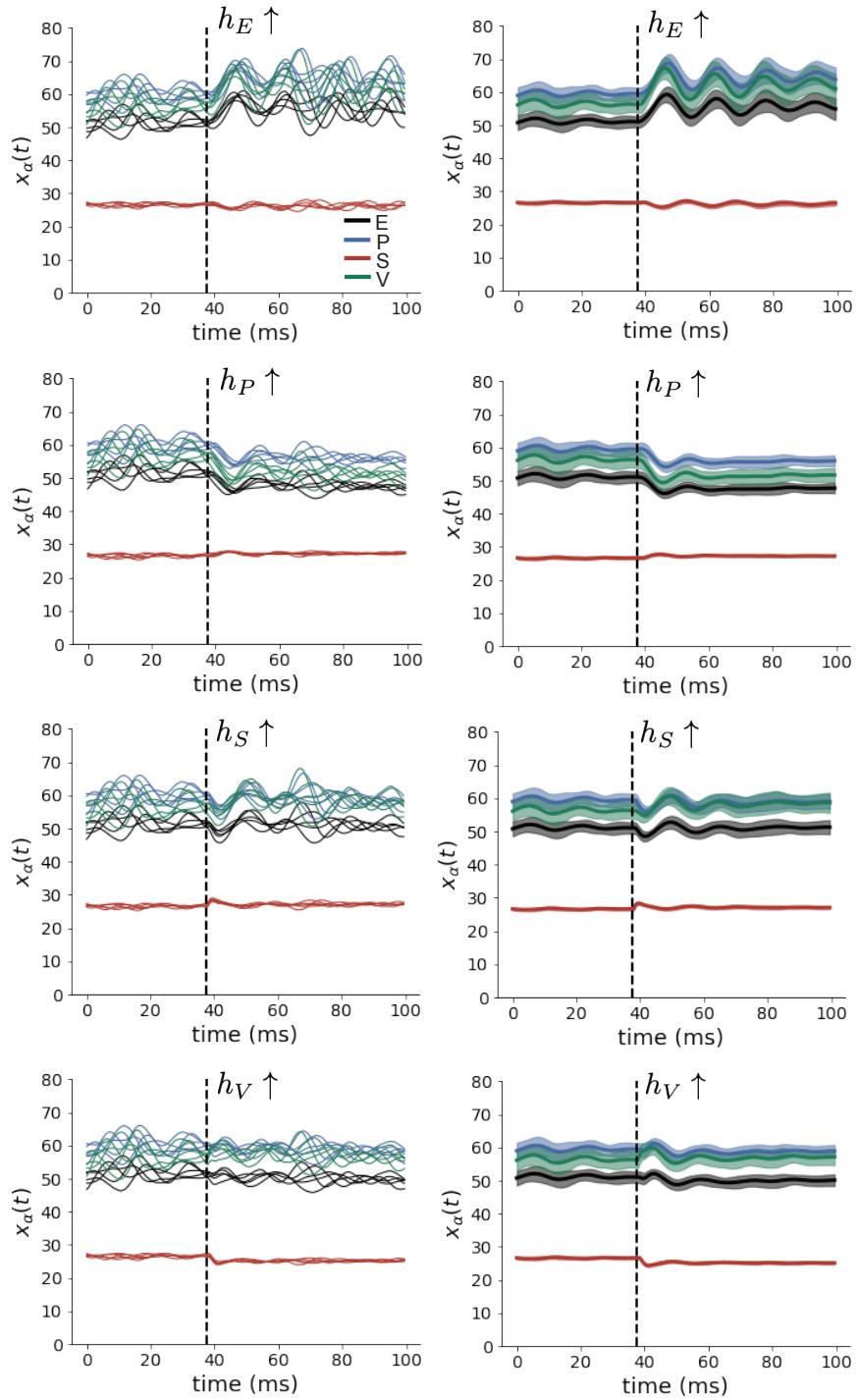


Figure 12: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

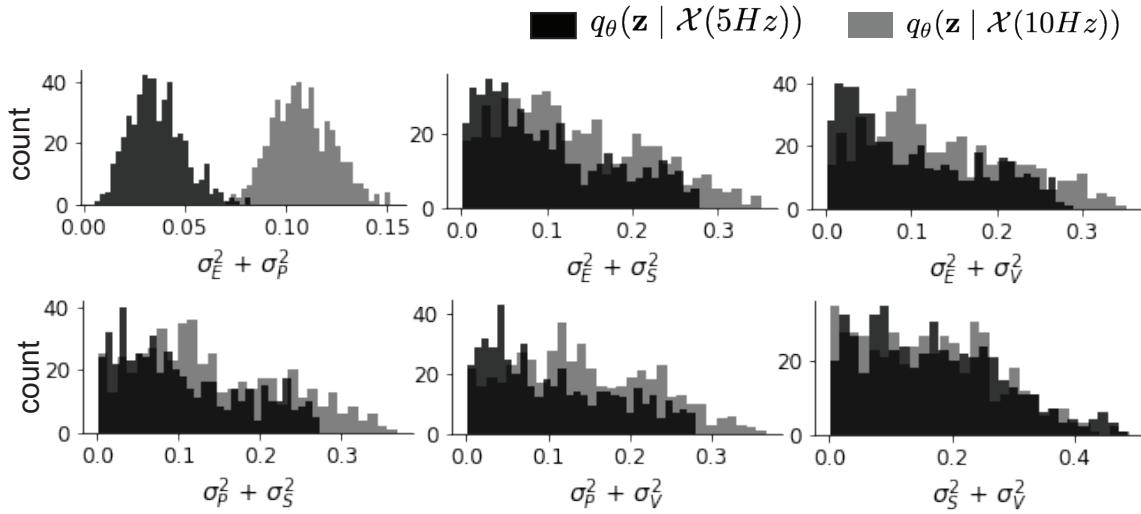


Figure 13: (V1 2) Posterior predictive distributions of the sum of squares of each pair of noise parameters.

1194 and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (72)$$

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

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (73)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (74)$$

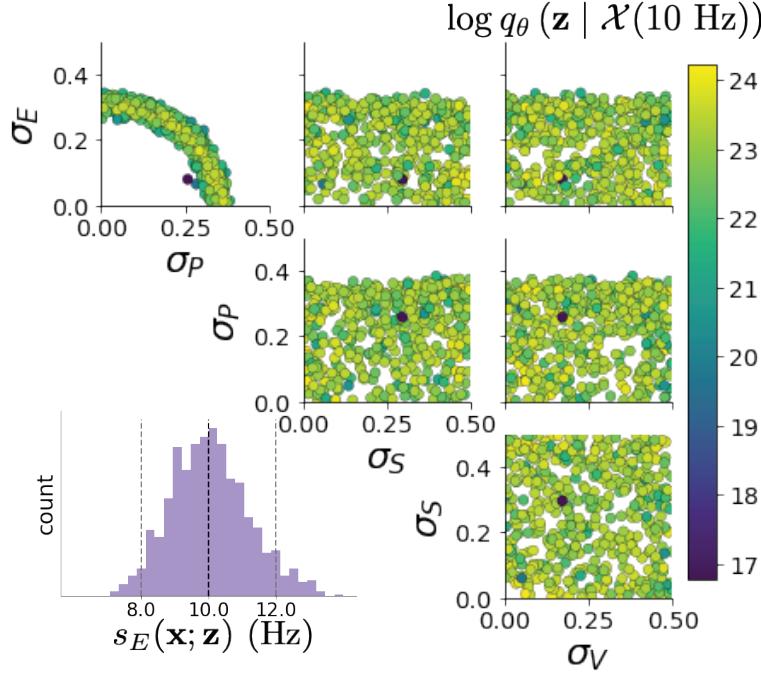


Figure 14: (V1 3) EPI posterior for  $\mathcal{X}(10 \text{ Hz})$ .

1197 and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (75)$$

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  
 1201 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[ \sqrt{\mathbb{E}_{t > t_{ss}} [(x_E(t) - \mathbb{E}_{t > t_{ss}} [x_E(t)])^2]} \right] \quad (76)$$

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  $q_{\theta}(\mathbf{z} \mid \mathcal{X})$  throughout the  $\sigma_E$ - $\sigma_P$  marginal. Specifically, we  
 1209 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_{\theta}(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (77)$$

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

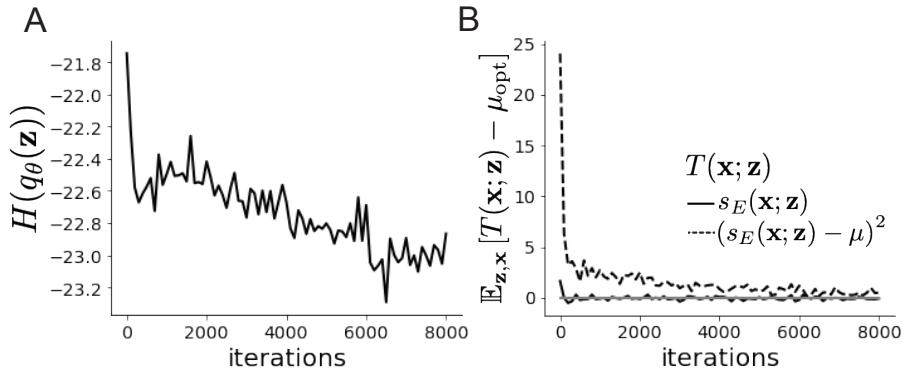


Figure 15: (V1 4) Optimization for V1

#### 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 In the model of Duan et al [56], there are four total units: two in each hemisphere corresponding to  
 1216 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right  
 1217 Pro (RP) and right Anti (RA). Each unit has an activity ( $x_{\alpha}$ ) and internal variable ( $u_{\alpha}$ ) related  
 1218 by

$$x_{\alpha} = \phi(u_{\alpha}) = \left( \frac{1}{2} \tanh \left( \frac{u_{\alpha} - a}{b} \right) + \frac{1}{2} \right) \quad (78)$$

1219 where  $\alpha \in \{LP, LA, RA, RP\}$ ,  $a = 0.05$  and  $b = 0.5$  control the position and shape of the nonlin-  
 1220 earity, respectively. During periods of optogenetic inactivation, activity was decreased proportional  
 1221 to the optogenetic strength  $\gamma$

$$x_{\alpha} = (1 - \gamma)\phi(u_{\alpha}). \quad (79)$$

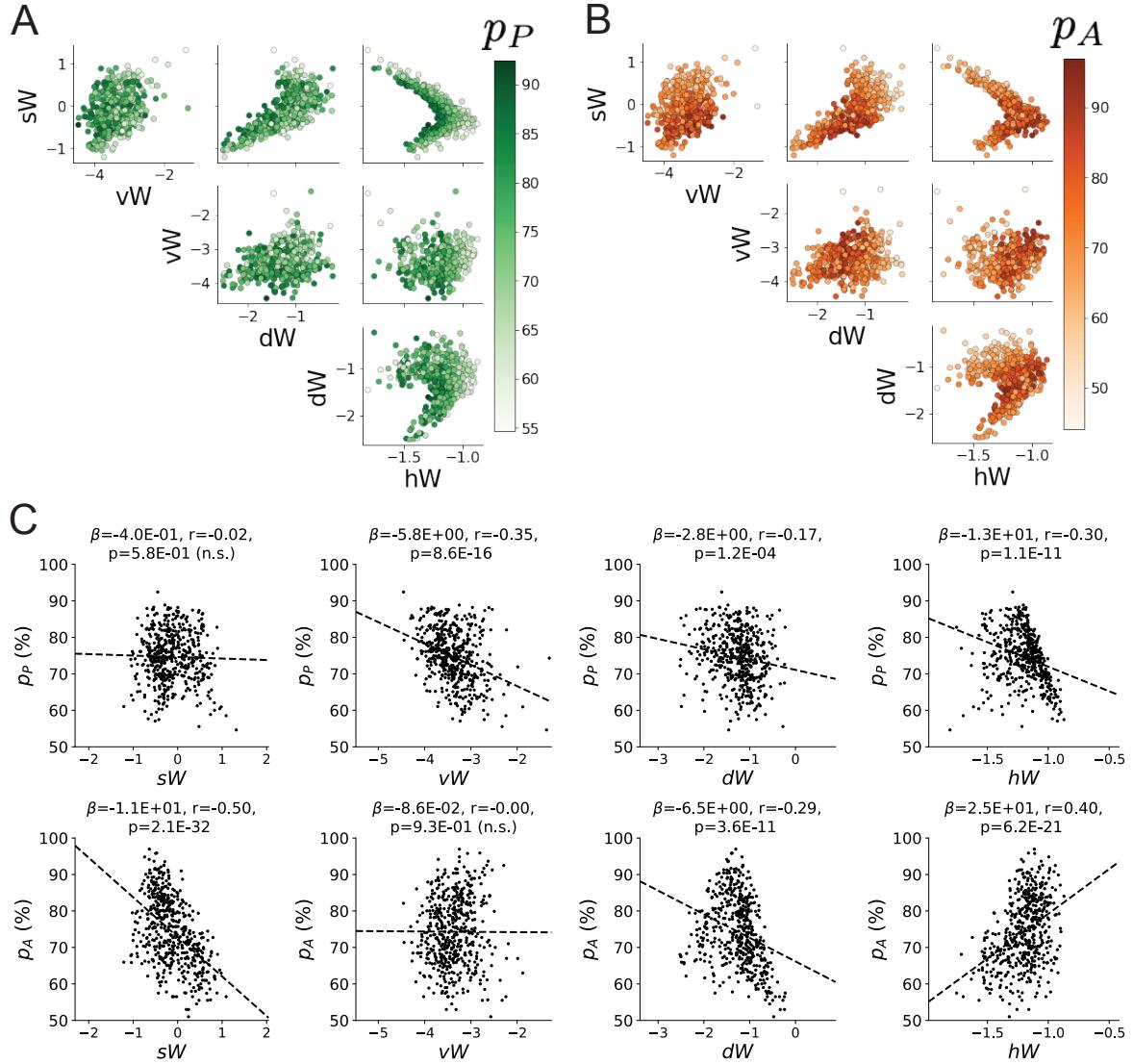


Figure 16: (SC1): **A.** Same pairplot as Fig. 4C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies.  $\beta$  is slope coefficient of linear regression,  $r$  is correlation, and  $p$  is the two-tailed p-value.

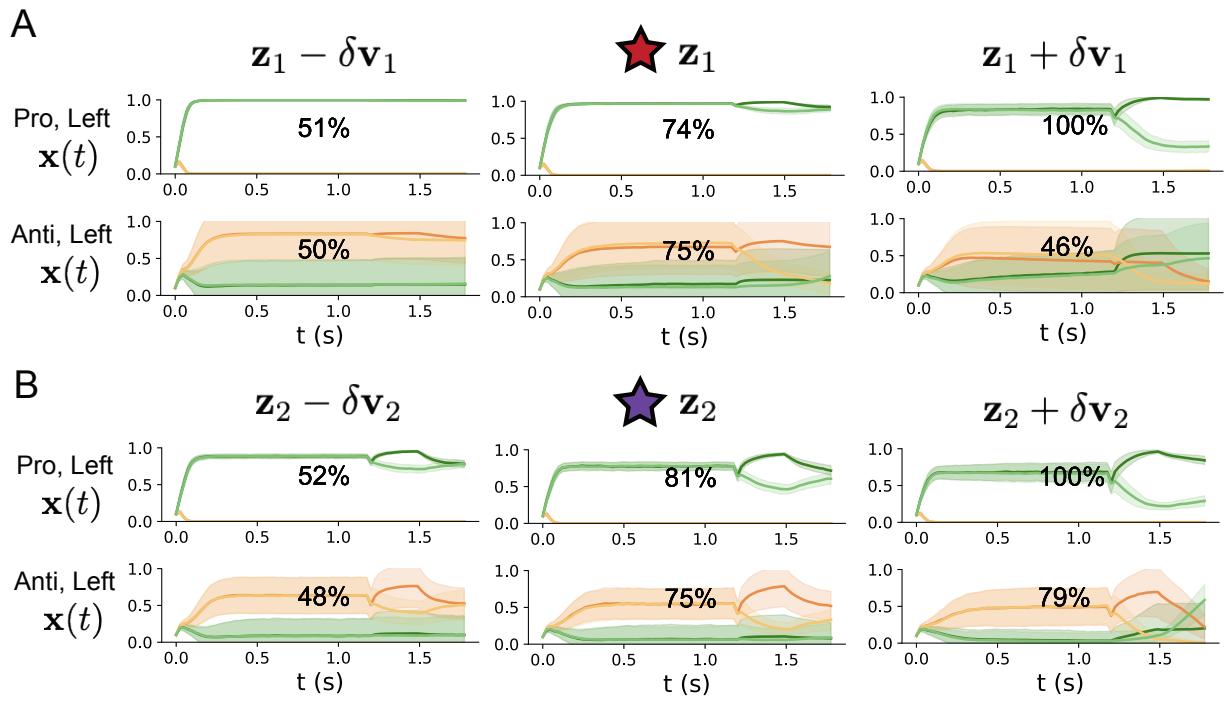


Figure 17: (SC2): **A.** Simulations in network regime  $\mathbf{z}_1$  (center) with simulations given connectivity perturbations in the negative direction of the sensitivity vector  $\mathbf{v}_1$  (left) and positive direction (right). **B.** Same as A for network regime  $\mathbf{z}_2$ .

<sub>1222</sub> We order the neural populations of  $x$  and  $u$  in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (80)$$

<sub>1223</sub> which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (81)$$

<sub>1224</sub> with time constant  $\tau = 0.09s$ , step size 24ms and Gaussian noise  $d\mathbf{B}$  of variance  $0.2^2$ . The weight  
<sub>1225</sub> matrix has 4 parameters  $sW$ ,  $vW$ ,  $hW$ , and  $dW$ :

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (82)$$

<sub>1226</sub> The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (83)$$

<sub>1227</sub> There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (84)$$

<sub>1228</sub> a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (85)$$

<sub>1229</sub> rule-based input depending on the condition

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

<sub>1230</sub>

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (87)$$

<sub>1231</sub> a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (88)$$

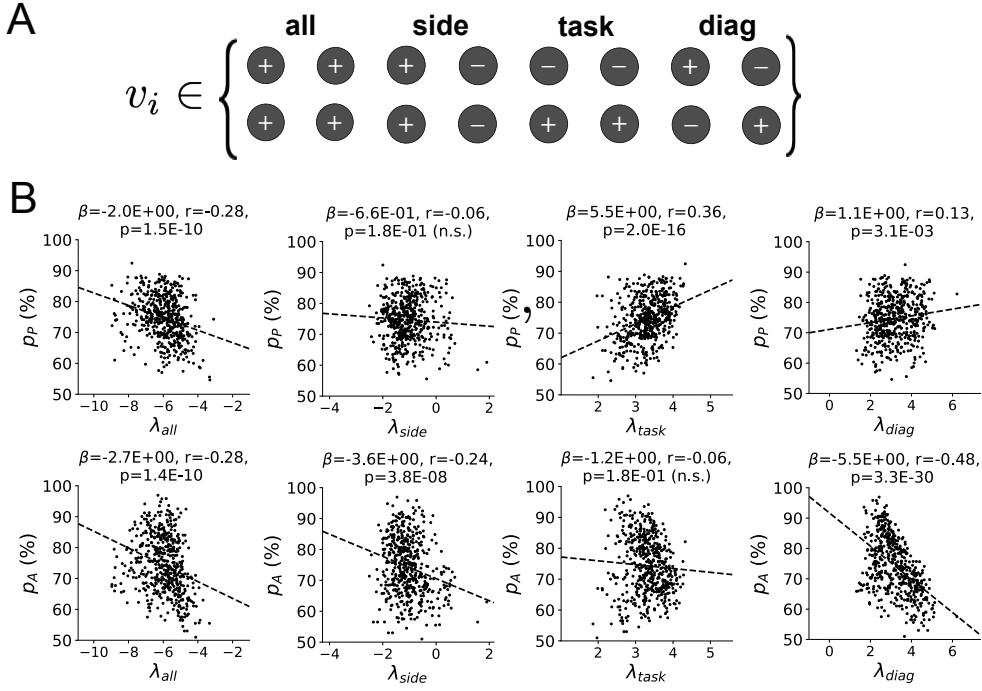


Figure 18: (SC3): **A.** Invariant eigenvectors of connectivity matrix  $W$ . **B.** Eigenvalues of connectivities of EPI distribution versus task accuracies.

1232 and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases} \quad (89)$$

1233 The input parameterization was fixed to  $I_{\text{constant}} = 0.75$ ,  $I_{P,\text{bias}} = 0.5$ ,  $I_{P,\text{rule}} = 0.6$ ,  $I_{A,\text{rule}} = 0.6$ ,

1234  $I_{\text{choice}} = 0.25$ , and  $I_{\text{light}} = 0.5$ .

1235 The accuracies of  $p_P$  and  $p_A$  are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (90)$$

1236 and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (91)$$

1237 given that the stimulus is on the left side, where  $\Theta$  is the Heaviside step function, and the accuracy

1238 is averaged over 200 independent trials. The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (92)$$

1239 where  $\beta = 100$ .

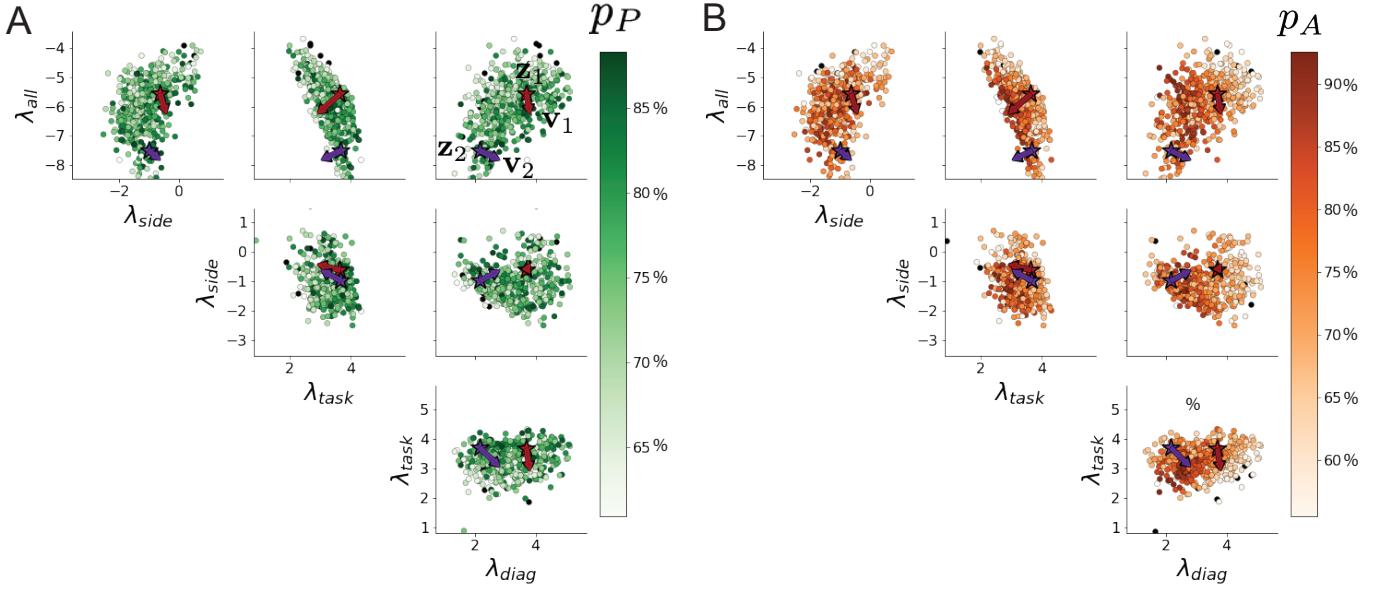


Figure 19: (SC4): **A.** Pairplots of eigenvalues of connectivity matrices in EPI distribution colored by Pro task accuracy. Red and purple stars and arrows correspond to eigenvalues and sensitivity directions  $\mathbf{z}_1$ ,  $\mathbf{z}_2$ ,  $\mathbf{v}_1$ , and  $\mathbf{v}_2$ . **B.** Same colored by Anti task accuracy.

1240 Writing the EPI posterior as a maximum entropy distribution,  $T(\mathbf{x}; \mathbf{z})$  is comprised of both these  
 1241 first and second moments of the accuracy in each task (as in Equations 28 and 29)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p(\mathbf{x}; \mathbf{z})_P \\ p(\mathbf{x}; \mathbf{z})_A \\ (p(\mathbf{x}; \mathbf{z})_P - 75\%)^2 \\ (p(\mathbf{x}; \mathbf{z})_A - 75\%)^2 \end{bmatrix}, \quad (93)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 7.5\%^2 \\ 7.5\%^2 \end{bmatrix}. \quad (94)$$

1242 Throughout optimization, the augmented Lagrangian parameters  $\eta$  and  $c$ , were updated after each  
 1243 epoch of 2,000 iterations(see Section 5.1.3). The optimization converged after six epochs (Fig. 21).

1244 For EPI in Fig. 4C, we used a real NVP architecture with three coupling layers of affine transfor-  
 1245 mations parameterized by two-layer neural networks of 50 units per layer. The initial distribution  
 1246 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  
 1247 augmented Lagrangian coefficient of  $c_0 = 10^2$ , a batch size  $n = 100$ , and  $\beta = 2$ . The distribution  
 1248

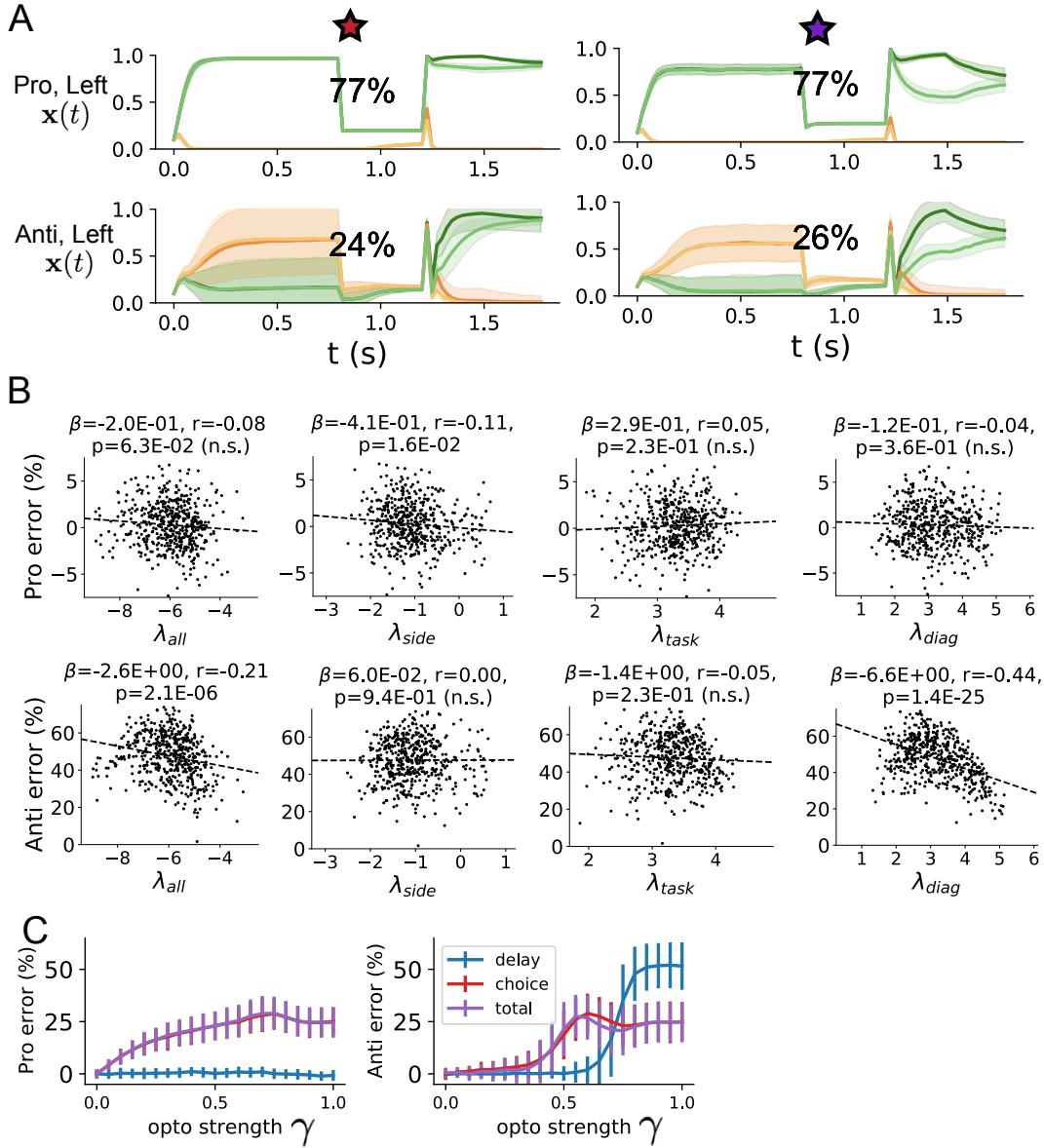


Figure 20: (SC5): **A.** Response of each parameter regime to optogenetic silencing during the delay period. **B.** Connectivity eigenvalues versus the task error induced by delay period inactivation. **C.** Error induced by delay period inactivation with increasing optogenetic strength. Means and standard deviations are calculated across the entire EPI posterior.

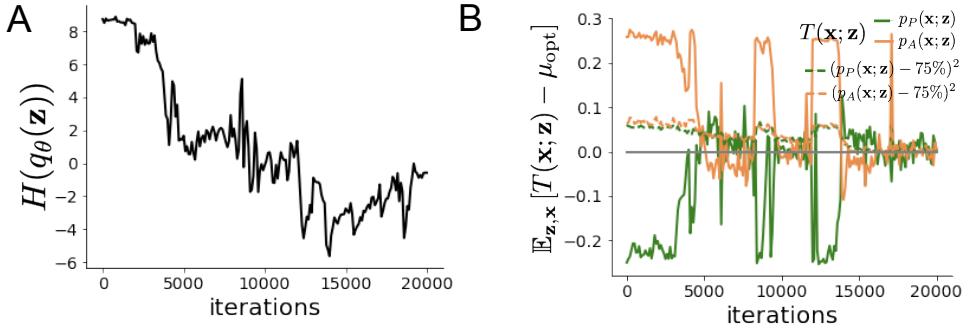


Figure 21: (SC6): **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

1249 shown is that of the architecture converging with criteria  $N_{\text{test}} = 25$  at greatest entropy across  
1250 random seeds.

1251 To make sense of this inferred distribution, we identified two modes used to represent the two  
1252 regimes of connectivity in this posterior:

$$\begin{aligned} \mathbf{z}_1 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW > 0 \end{aligned} \tag{95}$$

1253 and

$$\begin{aligned} \mathbf{z}_2 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW < 0 \end{aligned} \tag{96}$$

1254 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-  
1255 tion of the symmetric connectivity matrices  $W = V \Lambda V^{-1}$ , which results in the same basis vectors  
1256  $\mathbf{v}_i$  for all  $W$  parameterized by  $\mathbf{z}$  (Fig. 18A). These basis vectors have intuitive roles in processing  
1257 for this task, and are accordingly named the *all* mode - all neurons co-fluctuate, *side* mode - one  
1258 side dominates the other, *task* mode - the Pro or Anti populations dominate the other, and *diag*  
1259 mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair. We found  
1260 significant trends across the EPI posterior connectivities: the eigenvalues  $\lambda_{\text{task}}$  and  $\lambda_{\text{diag}}$  were cor-  
1261 related with  $p_P$ , while  $\lambda_{\text{all}}$  was anticorrelated with  $p_P$ .  $\lambda_{\text{all}}$ ,  $\lambda_{\text{side}}$ , and  $\lambda_{\text{diag}}$  were all significantly  
1262 anticorrelated with  $p_A$ .

1263 Under this decomposition, we can re-visualize the posterior in eigenvalue space (Fig. 19). Fur-  
1264 thermore, we can project the dimensions of sensitivity into eigenvalue space as well, giving us a  
1265 more intuitive sense of how connectivity affects computation in each regime. We see that sensitivity

1266 dimensions  $\mathbf{v}_1$  and  $\mathbf{v}_2$ , which cause  $p_P$  to increase and a regime dependent change in  $p_A$ , both point  
1267 in the direction of increasing  $\lambda_{\text{side}}$  and decreasing  $\lambda_{\text{task}}$ . These eigenvalue changes are evident in  
1268 the simulations of connectivity perturbations away from the modes (Fig. 17). As the component  
1269 of connectivity along  $\mathbf{v}_1$  and  $\mathbf{v}_2$  becomes stronger (left-to-right), there is less separation between  
1270 Pro an Anti populations (lower  $\lambda_{\text{task}}$ ) and greater separation between Left and Right populations  
1271 following stimulus presentation (greater  $\lambda_{\text{side}}$ ). A key differentiating factor is that  $\mathbf{v}_1$  substantially  
1272 increases  $\lambda_{\text{diag}}$ , while  $\mathbf{v}_2$  does not.

1273 During optogenetic silencing simulations, activations  $x_\alpha(t)$  were set to a fraction of their values ( $1 -$   
1274  $\gamma$ ), where  $\gamma$  is the optogenetic perturbation strength. We found that  $\lambda_{\text{all}}$  and  $\lambda_{\text{diag}}$  were significantly  
1275 anticorrelated with Anti error during delay period inactivation. Delay period inactivation was from  
1276  $0.8 < t < 1.2$ , choice period inactivation was for  $t > 1.2$  and total inactivation was for the entire  
1277 trial.