

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</sup>, Chunyu A. Duan<sup>4</sup>, Carlos D. Brody<sup>2,3,5</sup>,  
Kenneth D. Miller<sup>1</sup>, and John P. Cunningham<sup>6</sup>.

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

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

<sup>3</sup>Princeton University,

<sup>4</sup>Institute of Neuroscience, Chinese Academy of Sciences,

<sup>5</sup>Howard Hughes Medical Institute,

<sup>6</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 in terms of neural activity – and thus  
<sup>5</sup> can offer insights into neural computation. The operation of these circuits, like all models, crit-  
<sup>6</sup> ically depends on the choices of model parameters. When analytic derivation of the relationship  
<sup>7</sup> between model parameters and computational properties is intractable, approximate inference and  
<sup>8</sup> simulation-based techniques are relied upon for scientific insight. We bring the use of deep genera-  
<sup>9</sup> tive models for probabilistic inference to bear on this problem, learning distributions of parameters  
<sup>10</sup> that produce the specified properties of computation. By learning parameter distributions that  
<sup>11</sup> produce computations – an emergent property, we introduce a novel methodology for exploratory  
<sup>12</sup> analyses and hypothesis testing that is particularly well-suited to the stochastic dynamical sys-  
<sup>13</sup> tems models predominant in our field of theoretical neuroscience. We motivate this methodology  
<sup>14</sup> with a worked example analyzing sensitivity in the stomatogastric ganglion. We then use it to go  
<sup>15</sup> beyond linear theory of neuron-type input-responsivity in a model of primary visual cortex, gain  
<sup>16</sup> a mechanistic understanding of rapid task switching in superior colliculus models, and attribute  
<sup>17</sup> error to connectivity properties in recurrent neural networks solving a simple mathematical task.  
<sup>18</sup> While much use of deep learning in theoretical neuroscience focuses on drawing analogies between  
<sup>19</sup> optimized neural architectures and the brain, this work illustrates how we can further leverage the  
<sup>20</sup> power of deep learning towards solving inverse problems in theoretical neuroscience.

## 21 2 Introduction

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

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

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

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

55 Existing approaches to the inverse problem in theoretical neuroscience fall short in three key ways.  
56 First, theoretical models of neural computation aim to reflect a complex biological reality, and as a  
57 result, such models lack tractable likelihoods. Thus, standard approaches from statistical inference  
58 are unavailable. The parameter sets obtained from likelihood-free ABC lack a formalized link  
59 to Bayesian inference (except in the unrealistic 0-distance scenario), lack parameter probabilities,  
60 and only confer sensitivity analyses of an alternative likelihood to the simulator-defined likelihood  
61 of ABC [17]. Second is the undesirable trade-off between the flexibility and tractability of the  
62 approximated posterior distribution. While sampling-based approaches like ABC and Markov chain  
63 Monte Carlo (MCMC) can produce flexible posterior approximations, they must be run continually  
64 for increasing samples. While VAE approaches can result in tractable posterior sampling and  
65 sensitivity measurements post-optimization, existing approaches have relied on simplified classes  
66 of distributions, which restrict the flexibility of the posterior approximation. And third, you can  
67 never make assumptions of what inferred model parameters will predict. This is well understood  
68 when considering Box’s loop and the role of posterior predictive checks in the development and  
69 critique of scientific models [18, 19]. Uncertainty about the properties of inferred model predictions  
70 introduce a conceptual degree of freedom to the inverse problem that may be unnecessary and  
71 undesirable given the scientific motivation.

72 To address these three challenges, we developed an inference methodology – ‘emergent property  
73 inference’ – which learns a distribution over parameter configurations in a theoretical model. This  
74 distribution has two critical properties: *(i)* it is chosen such that draws from the distribution (pa-  
75 rameter configurations) correspond to systems of equations that give rise to a specified emergent  
76 property (a set of constraints); and *(ii)* it is chosen to have maximum entropy given those con-  
77 straints, such that we identify all likely parameters and can use the distribution to reason about  
78 parametric sensitivity and degeneracies [20]. First, we use stochastic gradient techniques in the  
79 spirit of likelihood-free variational inference [21] to enable inference in likelihood-free models of  
80 neural computation. Second, we stipulate a bijective deep neural network that induces a flexible  
81 family of probability distributions over model parameterizations with a probability density we can  
82 calculate [22, 23, 24], which confers fast sampling and sensitivity measurements. Third, we quan-  
83 tify the notion of emergent properties as a set of moment constraints on datasets generated by the

84 model. Thus, an emergent property is not a single data realization, but a phenomenon or a feature  
85 of the model, which is ultimately the object of interest in theoretical neuroscience. Conditioning  
86 on an emergent property requires a variant of deep probabilistic inference methods, which we have  
87 previously introduced [25]. Taken together, emergent property inference (EPI) provides a method-  
88 ology for inferring parameter configurations consistent with a particular emergent phenomena in  
89 theoretical models. We use a classic example of parametric degeneracy in a biological system, the  
90 stomatogastric ganglion [26], to motivate and clarify the technical details of EPI.

91 Equipped with this methodology, we then investigated three models of current importance in the-  
92 oretical neuroscience. These models were chosen to demonstrate generality through ranges of bi-  
93 ological realism (from conductance-based biophysics to recurrent neural networks), neural system  
94 function (from pattern generation to abstract cognitive function), and network scale (from four to  
95 infinite neurons). First, we use EPI to elucidate the mechanisms of inhibition stabilization with  
96 varying contrast in a stochastic nonlinear dynamical model of primary visual cortex with inhibitory  
97 multiplicity. Second, we discover connectivity patterns in superior colliculus resulting in resilience  
98 to optogenetic perturbation by using EPI to condition on rapid task switching. Third, we use EPI  
99 to uncover the sources of error in a low-rank recurrent neural network executing a simple math-  
100 ematical task. The novel scientific insights offered by EPI contextualize and clarify the previous  
101 studies exploring these models [27, 28, 29, 30], and more generally, these results point to the value  
102 of deep inference for the interrogation of biologically relevant models.

## 103 3 Results

### 104 3.1 Motivating emergent property inference of theoretical models

105 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-  
106 erty inference (Fig. 1A). First, one designs or chooses an existing model that, it is hypothesized,  
107 captures the computation of interest. To ground this process in a well-known example, consider  
108 the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple  
109 rhythmic muscle activation patterns for digestion [31]. Despite full knowledge of STG connectivity  
110 and a precise characterization of its rhythmic pattern generation, biophysical models of the STG  
111 have complicated relationships between circuit parameters and neural activity [26, 32]. A subcir-  
112 cuit model of the STG [27] is shown schematically in Figure 1C, and note that the behavior of this  
113 model will be critically dependent on its parameterization – the choices of conductance parameters

114  $\mathbf{z} = [g_{el}, g_{synA}]$ . Specifically, the two fast neurons ( $f1$  and  $f2$ ) mutually inhibit one another, and  
115 oscillate at a faster frequency than the mutually inhibiting slow neurons ( $s1$  and  $s2$ ). The hub  
116 neuron (hub) couples with either the fast or slow population or both.

117 Second, once the model is selected, one defines the emergent property, the measurable behavior  
118 of scientific interest. To continue our running STG example, one such emergent property is the  
119 phenomenon of *unified intermediacy* – in certain parameter regimes, the frequency of all neurons  
120 match at an intermediate firing rate. This emergent property is shown in Figure 1D at a frequency  
121 of 0.53Hz.

122 Third, qualitative parameter analysis ensues: since mathematical analysis of frequency is intractable  
123 in this model, a brute force sweep of parameters is done [27]. Subsequently, a qualitative description  
124 is formulated to describe the different parameter configurations that lead to the emergent property.  
125 In this last step lies the opportunity for a precise quantification of the emergent property as a  
126 statistical feature of the model. Once we have such a methodology, we can infer a probability  
127 distribution over parameter configurations that produce this emergent property.

128 Before presenting technical details (in the following section), let us understand emergent property  
129 inference schematically: EPI (Fig. 1A) takes, as input, the model and the specified emergent  
130 property, and as its output, produces the parameter distribution. This distribution – represented  
131 for clarity as samples from the distribution – is then a scientifically meaningful and mathematically  
132 tractable object. In the STG model, this distribution can be specifically queried to reveal the  
133 prototypical parameter configuration for unified intermediacy (the mode; Fig. 1B, yellow star),  
134 and how it decays based on changes away from the mode. The eigenvectors (of the Hessian of the  
135 distribution at the mode) quantitatively formalize the robustness of unified intermediacy (Fig. 1B  
136 solid ( $v_1$ ) and dashed ( $v_2$ ) black arrows). Indeed, samples equidistant from the mode along these  
137 EPI-identified dimensions of sensitivity ( $v_1$ ) and degeneracy ( $v_2$ ) agree with error contours (Fig.  
138 1B contours) and have diminished or preserved network syncing, respectively (Fig. 1D activity  
139 traces, Fig. S TODO) (see Section 5.2.1).

### 140 3.2 A deep generative modeling approach to emergent property inference

141 Emergent property inference (EPI) systematizes the three-step procedure of the previous section.  
142 First, we consider the model as a coupled set of differential (and potentially stochastic) equations  
143 [27]. In the running STG example, the model activity  $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$  is the membrane

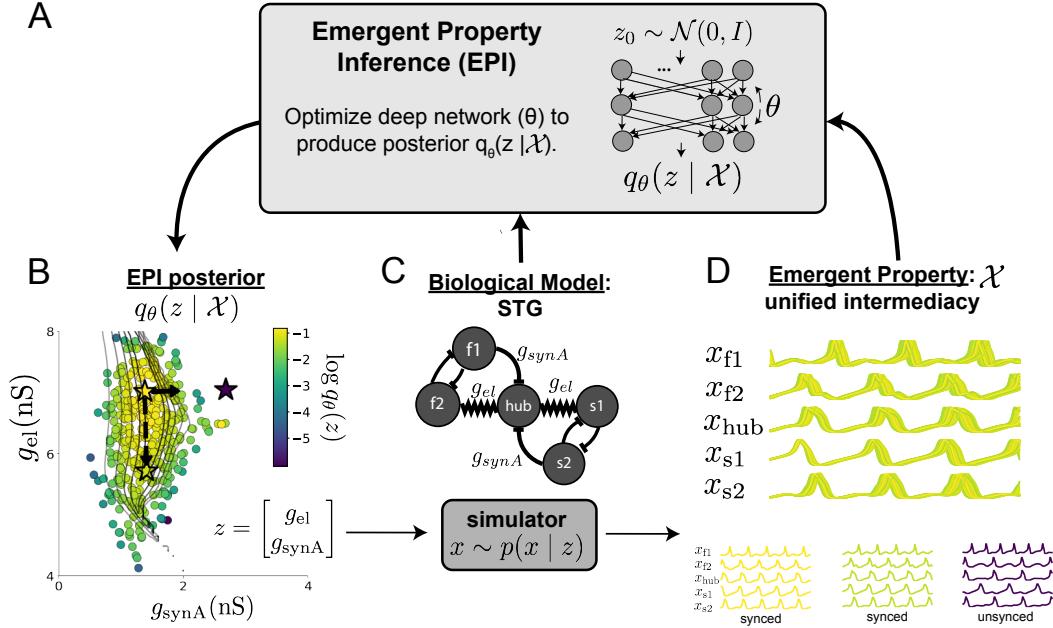


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. A. For a choice of model (STG) and emergent property (unified intermediacy), emergent property inference (EPI) learns a distribution of the model parameters  $\mathbf{z} = [g_{el}, g_{synA}]$  producing unified intermediacy. In the STG model, jagged connections indicate electrical coupling having electrical conductance  $g_{el}$ . Other connections in the diagram are inhibitory synaptic projections having strength  $g_{synA}$  onto the hub neuron, and  $g_{synB} = 5\text{nS}$  for mutual inhibitory connections. Deep probability distributions map a simple random variable  $\mathbf{w}$  through a deep neural network with weights and biases  $\boldsymbol{\theta}$  to parameters  $\mathbf{z} = f_\theta(\mathbf{w})$  distributed as  $q_\theta(\mathbf{z} | \mathcal{X})$ . B. The EPI distribution of STG model parameters producing network syncing. Samples are colored by log probability density. Distribution contours of emergent property value error are shown at levels of  $2.5 \times 10^{-5}$ ,  $5 \times 10^{-5}$ ,  $1 \times 10^{-4}$ ,  $2 \times 10^{-4}$ , and  $4 \times 10^{-4}$  (dark to light gray). Eigenvectors of the Hessian at the mode of the inferred distribution are indicated as  $\mathbf{v}_1$  (solid) and  $\mathbf{v}_2$  (dashed) with lengths scaled by the square root of the absolute value of their eigenvalues. Simulated activity is shown for three samples (stars) in panel D.  $v_1$  is sensitive to network syncing ( $p < 10^{-4}$ ), while  $v_2$  is not ( $p = 0.67$ ) (see Section 5.2.1). D. The emergent property of unified intermediacy, in which all neurons are firing close to the same intermediate frequency. Simulated activity traces are colored by log probability density of their generating parameters in the EPI-inferred distribution.

144 potential for each neuron, which evolves according to the biophysical conductance-based equation:

$$C_m \frac{d\mathbf{x}}{dt} = -\mathbf{h}(\mathbf{x}; \mathbf{z}) = -[\mathbf{h}_{leak}(\mathbf{x}; \mathbf{z}) + \mathbf{h}_{Ca}(\mathbf{x}; \mathbf{z}) + \mathbf{h}_K(\mathbf{x}; \mathbf{z}) + \mathbf{h}_{hyp}(\mathbf{x}; \mathbf{z}) + \mathbf{h}_{elec}(\mathbf{x}; \mathbf{z}) + \mathbf{h}_{syn}(\mathbf{x}; \mathbf{z})] \quad (1)$$

145 where  $C_m = 1\text{nF}$ , and  $\mathbf{h}_{leak}$ ,  $\mathbf{h}_{Ca}$ ,  $\mathbf{h}_K$ ,  $\mathbf{h}_{hyp}$ ,  $\mathbf{h}_{elec}$ , and  $\mathbf{h}_{syn}$  are the leak, calcium, potassium, hyper-  
146 polarization, electrical, and synaptic currents, all of which have their own complicated dependence  
147 on  $\mathbf{x}$  and  $\mathbf{z} = [g_{el}, g_{synA}]$  (see Section 5.2.1).

148 Second, we define the emergent property, which as above is network syncing: oscillation of the  
149 entire population at an intermediate frequency of our choosing (Figure 1A bottom). Quantifying  
150 this phenomenon is straightforward: we define network syncing to be that each neuron’s spiking  
151 frequency – denoted  $\omega_{f1}(\mathbf{x})$ ,  $\omega_{f2}(\mathbf{x})$ , etc. – is close to an intermediate frequency of 0.53Hz. Math-  
152 ematically, we achieve this via constraints on the mean and variance of  $\omega_\alpha(\mathbf{x})$  for each neuron  
153  $\alpha \in \{f1, f2, \text{hub}, s1, s2\}$ :

$$\begin{aligned} \mathcal{X} &: \mathbb{E}_{\mathbf{z}} [T(\mathbf{x}; \mathbf{z})] \triangleq \mathbb{E}_{\mathbf{z}} \begin{bmatrix} \omega_{f1}(\mathbf{x}; \mathbf{z}) \\ \omega_{f2}(\mathbf{x}; \mathbf{z}) \\ \vdots \end{bmatrix} = \begin{bmatrix} 0.53 \\ 0.53 \\ \vdots \end{bmatrix} \triangleq \boldsymbol{\mu} \\ \text{Var}_{\mathbf{z}} [T(\mathbf{x}; \mathbf{z})] &\triangleq \text{Var}_{\mathbf{z}} \begin{bmatrix} \omega_{f1}(\mathbf{x}; \mathbf{z}) \\ \omega_{f2}(\mathbf{x}; \mathbf{z}) \\ \vdots \end{bmatrix} = \begin{bmatrix} 0.025^2 \\ 0.025^2 \\ \vdots \end{bmatrix} \triangleq \boldsymbol{\sigma}^2. \end{aligned} \quad (2)$$

154 The emergent property statistics  $T(\mathbf{x}; \mathbf{z})$  along with their constrained means  $\boldsymbol{\mu}$  and variances  $\boldsymbol{\sigma}^2$   
155 define the emergent property denoted  $\mathcal{X}$ .

156 Third, we perform emergent property inference: we find a distribution over parameter configura-  
157 tions  $\mathbf{z}$ , and insist that samples from this distribution produce the emergent property; in other  
158 words, they obey the constraints introduced in Equation 2. This distribution will be chosen from a  
159 family of probability distributions  $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$ , defined by a deep generative distribution  
160 of the normalizing flow class [22, 23, 24] – neural networks which transform a simple distribution  
161 into a suitably complicated distribution (as is needed here). This deep distribution is represented  
162 in Figure 1B (see Section 5.1). Then, mathematically, we must solve the following optimization  
163 program:

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

164 where  $T(\mathbf{x}, \mathbf{z})$ ,  $\boldsymbol{\mu}$ , and  $\boldsymbol{\sigma}$  are defined as in Equation 2. Finally, we recognize that many distributions  
165 in  $\mathcal{Q}$  will respect the emergent property constraints, so we select that which has maximum entropy.  
166 This principle, captured in Equation 3 by the primal objective  $H$ , identifies parameter distributions  
167 with minimal assumptions beyond some chosen structure [33, 34, 25, 35]. Such a normative principle  
168 of maximum entropy, which is also that of Bayesian inference, naturally fits with our scientific  
169 objective of reasoning about parametric sensitivity and robustness. The recovered distribution of  
170 EPI is as variable as possible along each parametric manifold such that it produces the emergent  
171 property.

172 EPI optimizes the weights and biases  $\boldsymbol{\theta}$  of the deep neural network (which induces the probability  
173 distribution) by iteratively solving Equation 3. The optimization is complete when the sampled  
174 models with parameters  $\mathbf{z} \sim q_{\boldsymbol{\theta}}(z | \mathcal{X})$  produce activity consistent with the specified emergent  
175 property (Fig. S4). Such convergence is evaluated with a hypothesis test that the means and  
176 variances of each emergent property statistic are not different than their constrained values (see  
177 Section 5.1.2). Further validation of EPI is available in the supplementary materials, where we  
178 analyze a simpler model for which ground-truth statements can be made (Section 5.1.1).

179 In relation to broader methodology, inspection of the EPI objective reveals a natural relationship  
180 to posterior inference. Specifically, EPI executes variational inference in an exponential family  
181 model, the sufficient statistics and mean parameter of which are defined by  $T(\mathbf{x})$ ,  $\boldsymbol{\mu}$  and  $\boldsymbol{\sigma}$  (see  
182 Section 5.1.4). Equipped with this method, we may examine structure in posterior distributions or  
183 make comparisons between posteriors conditioned at different levels of the same emergent property  
184 statistic. We now prove out the value of EPI by using it to investigate and produce novel insights  
185 about three prominent models in neuroscience.

### 186 3.3 Comprehensive input-responsivity in a nonlinear sensory system

187 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output  
188 function have succeeded in explaining a host of experimentally documented phenomena. In a regime  
189 characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to  
190 paradoxical responses [4], selective amplification [36], surround suppression [37] and normalization  
191 [38]. Despite their strong predictive power, E-I circuit models rely on the assumption that inhibi-  
192 tion can be studied as an indivisible unit. However, experimental evidence shows that inhibition  
193 is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) – composing 80%  
194 of GABAergic interneurons in V1 [39, 40, 41], and that these inhibitory cell types follow specific

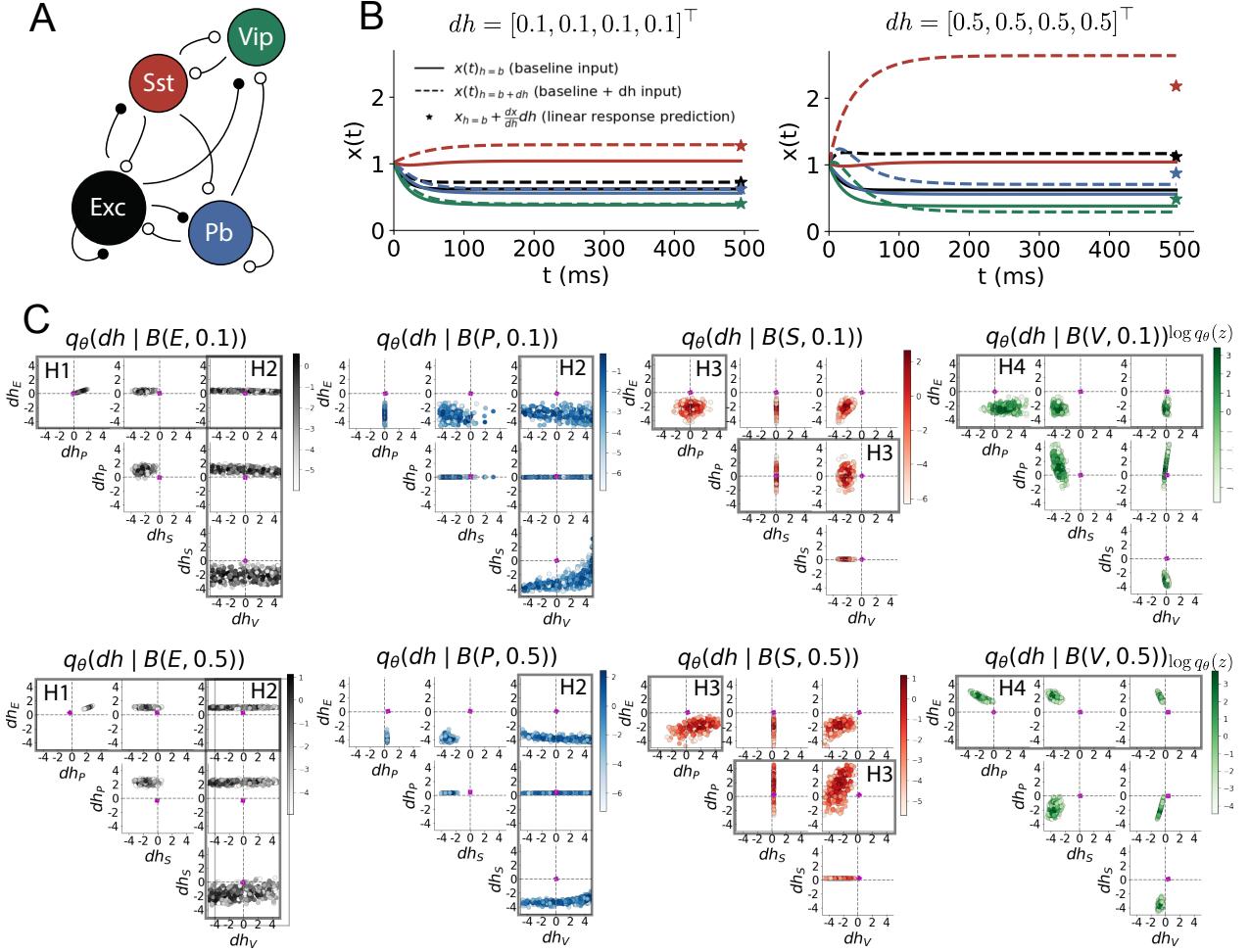


Figure 2: Hypothesis generation through EPI in a V1 model. A. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons. Some neuron-types largely do not form synaptic projections to others (excitatory and inhibitory projections filled and unfilled, respectively). B. Linear response predictions become inaccurate with greater input strength. V1 model simulations for input (solid)  $h = b$  and (dashed)  $h = b + dh$ . Stars indicate the linear response prediction. C. EPI distributions on differential input  $dh$  conditioned on differential response  $\mathcal{B}(\alpha, y)$ . Supporting evidence for the four generated hypotheses are indicated by gray boxes with labels H1, H2, H3, and H4. The linear prediction from two standard deviations away from  $y$  (from negative to positive) is overlaid in magenta (very small, near origin).

connectivity patterns (Fig. 2A) [42]. Recent theoretical advances [28, 43, 44], have only started to address the consequences of this multiplicity in the dynamics of V1, strongly relying on linear theoretical tools. Here, we go beyond linear theory by systematically generating and evaluating hypotheses of circuit model function using EPI distributions of neuron-type inputs producing various neuron-type population responses.

Specifically, we consider a four-dimensional circuit model with dynamical state given by the firing rate  $x$  of each neuron-type population  $x = [x_E, x_P, x_S, x_V]^\top$ . Given a time constant of  $\tau = 20$  ms and a power  $n = 2$ , the dynamics are driven by the rectified and exponentiated sum of recurrent ( $Wx$ ) and external  $h$  inputs:

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n. \quad (4)$$

We considered fixed effective connectivity weights  $W$  approximated from experimental recordings of publicly available datasets of mouse V1 [45, 46] (see Section 5.2.2). The input  $h = b + dh$  is comprised of a baseline input  $b = [b_E, b_P, b_S, b_V]^\top$  and a differential input  $dh = [dh_E, dh_P, dh_S, dh_V]^\top$  to each neuron-type population. Throughout subsequent analyses, the baseline input is  $b = [1, 1, 1, 1]^\top$ .

With this model, we are interested in the differential responses of each neuron-type population to changes in input  $dh$ . Initially, we studied the linearized response of the system to input  $\frac{dx_{ss}}{dh}$  at the steady state response  $x_{ss}$ , i.e. a fixed point. All analyses of this model consider the steady state response, so we drop the notation  $ss$  from here on. While this linearization accurately predicts differential responses  $dx = [dx_E, dx_P, dx_S, dx_V]^\top$  for small differential inputs to each population  $dh = [0.1, 0.1, 0.1, 0.1]^\top$  (Fig 2B left), the linearization is a poor predictor in this nonlinear model more generally (Fig. 2B right). Currently available approaches to deriving the steady state response of the system are limited.

To get a more comprehensive picture of the input-responsivity of each neuron-type beyond linear theory, we used EPI to learn a distribution of the differential inputs to each population  $dh$  that produce an increase of  $y$  in the rate of each neuron-type population  $\alpha \in \{E, P, S, V\}$ . We want to know the differential inputs  $dh$  that result in a differential steady state  $dx_\alpha$  (the change in  $x_\alpha$  when receiving input  $h = b + dh$  with respect to the baseline  $h = b$ ) of value  $y$  with some small, arbitrarily chosen amount of variance  $0.01^2$ . These statements amount to the emergent property

$$\mathcal{B}(\alpha, y) \triangleq \mathbb{E} \begin{bmatrix} dx_\alpha \\ (dx_\alpha - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix}. \quad (5)$$

222 We maintain the notation  $\mathcal{B}(\cdot)$  throughout the rest of the study as short hand for emergent property,  
 223 which represents a different signature of computation in each application.

224 Using EPI, we inferred the distribution of  $dh$  shown in Figure 2C producing  $\mathcal{B}(\alpha, y)$ . Columns  
 225 correspond to inferred distributions of excitatory ( $\alpha = E$ , red), parvalbumin ( $\alpha = P$ , blue), so-  
 226 matostatin ( $\alpha = S$ , red) and VIP ( $\alpha = V$ , green) neuron-type response increases, while each  
 227 row corresponds to increase amounts of  $y \in \{0.1, 0.5\}$ . For each pair of parameters, we show the  
 228 two-dimensional marginal distribution of samples colored by  $\log q_\theta(dh | \mathcal{B}(\alpha, y))$ . The inferred dis-  
 229 tributions immediately suggest four hypotheses:

230

231 H1: as is intuitive, each neuron-type's firing rate should be sensitive to that neuron-type's  
 232 direct input (e.g. Fig. 2C H1 gray boxes indicate low variance in  $dh_E$  when  $\alpha = E$ . Same  
 233 observation in all inferred distributions);

234 H2: the E- and P-populations should be largely unaffected by input to the V-population (Fig.  
 235 2C H2 gray boxes indicate high variance in  $dh_V$  when  $\alpha \in \{E, P\}$ );

236 H3: the S-population should be largely unaffected by input to the P-population (Fig. 2C H3  
 237 gray boxes indicate high variance in  $dh_P$  when  $\alpha = S$ );

238 H4: there should be a nonmonotonic response of the V-population with input to the E-  
 239 population (Fig. 2C H4 gray boxes indicate that negative  $dh_E$  should result in small  $dx_V$ ,  
 240 but positive  $dh_E$  should elicit a larger  $dx_V$ );

241 We evaluate these hypotheses by taking perturbations in individual neuron-type input  $\delta h_\alpha$  away  
 242 from the modes of the inferred distributions at  $y = 0.1$

$$dh^* = z^* = \underset{z}{\operatorname{argmax}} \log q_\theta(z | \mathcal{B}(\alpha, 0.1)). \quad (6)$$

243 Here  $\delta x_\alpha$  is the change in steady state response of the system with input  $h = b + dh^* + \delta h_\alpha \hat{u}_\alpha$   
 244 compared to  $h = b + dh^*$ , where  $\hat{u}_\alpha$  is a unit vector in the dimension of  $\alpha$ . The EPI-generated  
 245 hypotheses are confirmed (for details, see Section 5.2.2):

246 H1: the neuron-type responses are sensitive to their direct inputs (Fig. 3A black, 3B blue,  
 247 3C red, 3D green);

248 H2: the E- and P-populations are not affected by  $\delta h_V$  (Fig. 3A green, 3B green);

249 H3: the S-population is not affected by  $\delta h_P$  (Fig. 3C blue);

250 H4: the V-population exhibits a nonmonotonic response to  $\delta h_E$  (Fig. 3D black), and is in  
 251 fact the only population to do so (Fig. 3A-C black).

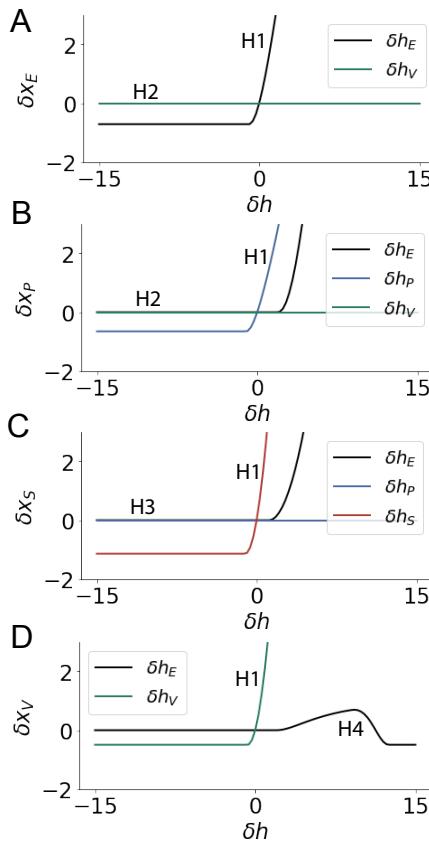


Figure 3: Confirming EPI generated hypotheses in V1. A. Differential responses  $\delta x_E$  by the E-population to changes in individual input  $\delta h_\alpha \hat{u}_\alpha$  away from the mode of the EPI distribution  $dh^*$ . B-D Same plots for the P-, S-, and V-populations. Labels H1, H2, H3, and H4 indicate which curves confirm which hypotheses.

252 These hypotheses were in stark contrast to what was available to us via traditional analytical linear  
 253 prediction (Fig. 2C, magenta, see Section 5.2.2).

254 Here, we examined the neuron-type responsivity of this model of V1 with scientifically motivated  
 255 choice of connectivity  $W$ . With EPI, we could just as easily have examined the distribution of such  
 256  $W$ 's consistent with some response characteristics for a fixed input  $h$  or another emergent property  
 257 such as inhibition stabilization. Most importantly, this analysis is a proof-of-concept demonstrating  
 258 the valuable ability to condition parameters of interest of a neural circuit model on some chosen  
 259 emergent property. To this point, we have shown the utility of EPI on relatively low-level emergent  
 260 properties like network syncing and differential neuron-type population responses. In the remainder  
 261 of the study, we focus on using EPI to understand models of more abstract cognitive function.

### 262 3.4 Identifying neural mechanisms of flexible task switching

263 In a rapid task switching experiment [47], rats were explicitly cued on each trial to either orient  
 264 towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti  
 265 (A) task (Fig. 4a). Neural recordings in the midbrain superior colliculus (SC) exhibited two

266 populations of neurons that simultaneously represented both task context (Pro or Anti) and motor  
 267 response (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons  
 268 [29]. Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is  
 269 a four-population dynamical system. We analyzed this model, where the neuron-type populations  
 270 are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right  
 271 (R)), their connectivity is parameterized geometrically (Fig. 4B). The input-output function of  
 272 this model is chosen such that the population responses  $x = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$  are bounded  
 273 from 0 to 1 giving rise to high (1) or low (0) responses at the end of the trial:

$$x_\alpha = \left( \frac{1}{2} \tanh \left( \frac{u_\alpha - \epsilon}{\zeta} \right) + \frac{1}{2} \right) \quad (7)$$

274 where  $\epsilon = 0.05$  and  $\zeta = 0.5$ . The dynamics evolve with timescale  $\tau = 0.09$  via an internal variable  
 275  $u$  governed by connectivity weights  $W$

$$\tau \frac{du}{dt} = -u + Wx + h + \sigma dB \quad (8)$$

276 with gaussian noise of variance  $\sigma^2 = 1$ . The input  $h$  is comprised of a cue-dependent input to the  
 277 Pro or Anti populations, a stimulus orientation input to either the Left or Right populations, and  
 278 a choice-period input to the entire network (see Section 5.2.3). Here, we use EPI to determine the  
 279 changes in network connectivity  $z = [sW_P, sW_A, vW_{PA}, vW_{AP}, dW_{PA}, dW_{AP}, hW_P, hW_A]$  resulting  
 280 in greater levels of rapid task switching accuracy.

281 To quantify the emergent property of rapid task switching at various levels of accuracy, we consid-  
 282 ered the requirements of this model in this behavioral paradigm. At the end of successful trials,  
 283 the response of the Pro population in the hemisphere of the correct choice must have a value near  
 284 1, while the Pro population in the opposite hemisphere must have a value near 0. Constraining a  
 285 population response  $x_\alpha \in [0, 1]$  to be either 0 or 1 can be achieved by requiring that it has Bernoulli  
 286 variance (see Section 5.2.3). Thus, we can formulate rapid task switching at a level of accuracy  
 287  $p \in [0, 1]$  in both tasks in terms of the average steady response of the Pro population  $\hat{p}$  of the  
 288 correct choice, the error in Bernoulli variance of that Pro neuron  $\sigma_{\text{err}}^2$ , and the average difference

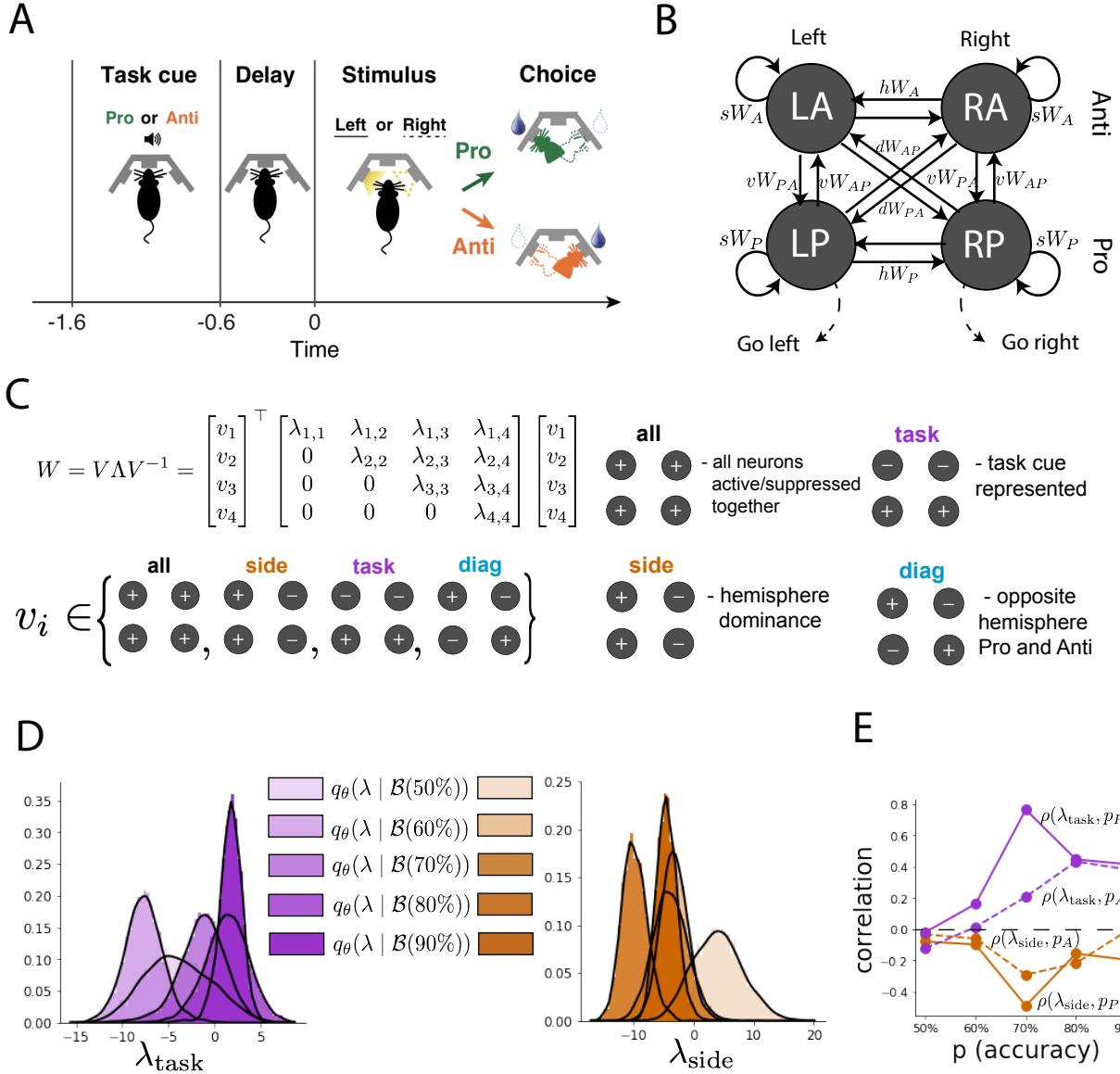


Figure 4: EPI reveals changes in SC [29] connectivity that control task accuracy. A. Rapid task switching behavioral paradigm (see text). B. Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters:  $sW$  - self,  $hW$  - horizontal,  $vW$  - vertical,  $dW$  - diagonal weights. Subscripts  $P$  and  $A$  of connectivity weights indicate Pro or Anti populations, and e.g.  $vW_{PA}$  is a vertical weight from an Anti to a Pro population. C. The Schur decomposition of the weight matrix  $W = V \Lambda V^{-1}$  is a unique decomposition with orthogonal  $V$  and upper triangular  $\Lambda$ . Schur modes:  $v_{\text{all}}$ ,  $v_{\text{task}}$ ,  $v_{\text{side}}$ , and  $v_{\text{diag}}$ . D. The marginal EPI distributions of the Schur eigenvalues at each level of task accuracy. E. The correlation of Schur eigenvalue with task performance in each learned EPI distribution.

289 in Pro neuron responses  $d$  in both Pro and Anti trials:

$$\mathcal{B}(p) \triangleq \mathbb{E} \begin{bmatrix} \hat{p}_P \\ \hat{p}_A \\ (\hat{p}_P - p)^2 \\ (\hat{p}_A - p)^2 \\ \sigma_{P,err}^2 \\ \sigma_{A,err}^2 \\ d_P \\ d_A \end{bmatrix} = \begin{bmatrix} p \\ p \\ 0.15^2 \\ 0.15^2 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}. \quad (9)$$

290 Thus,  $\mathcal{B}(p)$  denotes Bernoulli, winner-take-all responses between Pro neurons in a model executing  
291 rapid task switching near accuracy level  $p$ .

292 We used EPI to learn distributions of the SC weight matrix parameters  $z$  conditioned on of various  
293 levels of rapid task switching accuracy  $\mathcal{B}(p)$  for  $p \in \{50\%, 60\%, 70\%, 80\%, 90\%\}$ . To make sense  
294 of these inferred distributions, we followed the approach of Duan et al. by decomposing the con-  
295 nectivity matrix  $W = V\Lambda V^{-1}$  in such a way (the Schur decomposition) that the basis vectors  $v_i$   
296 are the same for all  $W$  (Fig. 4C). These basis vectors have intuitive roles in processing for this  
297 task, and are accordingly named the *all* mode - all neurons co-fluctuate, *side* mode - one side  
298 dominates the other, *task* mode - the Pro or Anti populations dominate the other, and *diag* mode -  
299 Pro- and Anti-populations of opposite hemispheres dominate the opposite pair. The corresponding  
300 eigenvalues (e.g.  $\lambda_{task}$ , which change according to  $W$ ) indicate the degree to which activity along  
301 that mode is increased or decreased by  $W$ .

302 We found that for greater task accuracies, the task mode eigenvalue increases, indicating the  
303 importance of  $W$  to the task representation (Fig. 4D, purple; adjacent distributions from 60%  
304 to 90% have  $p < 10^{-4}$ , Mann-Whitney test with 50 estimates and 100 samples). Stepping from  
305 random chance (50%) networks to marginally task-performing (60%) networks, there is a marked  
306 decrease of the side mode eigenvalues (Fig. 4D, orange;  $p < 10^{-4}$ ). Such side mode suppression  
307 relative to 50% remains in the models achieving greater accuracy, revealing its importance towards  
308 task performance. There were no interesting trends with task accuracy in the all or diag mode  
309 (hence not shown in Fig. 4). Importantly, we can conclude from our methodology that side  
310 mode suppression in  $W$  allows rapid task switching, and that greater task-mode representations  
311 in  $W$  increase accuracy. These hypotheses are confirmed by forward simulation of the SC model  
312 (Fig. 4E, see Section 5.2.3) suggesting experimentally testable predictions: increase in rapid task

313 switching performance should be correlated with changes in effective connectivity corresponding to  
 314 an increase in task mode and decrease in side mode eigenvalues.

315 **3.5 Linking RNN connectivity to error**

316 So far, each model we have studied was designed from fundamental biophysical principles, genetically-  
 317 or functionally-defined neuron types. At a more abstract level of modeling, recurrent neural net-  
 318 works (RNNs) are high-dimensional dynamical models of computation that are becoming increas-  
 319 ingly popular in neuroscience research [48]. In theoretical neuroscience, RNN dynamics usually  
 320 follow the equation

$$\frac{dx}{dt} = -x + W\phi(x) + h, \quad (10)$$

321 where  $x$  is the network activity,  $W$  is the network connectivity,  $\phi(\cdot) = \tanh(\cdot)$ , and  $h$  is the input to  
 322 the system. Such RNNs are trained to do a task from a systems neuroscience experiment, and then  
 323 the unit activations of the trained RNN are compared to recorded neural activity. Fully-connected  
 324 RNNs with tens of thousands of parameters are challenging to characterize [49], especially making  
 325 statistical inferences about their parameterization. Alternatively, we considered a rank-1,  $N$ -neuron  
 326 RNN with connectivity consisting of the sum of a random and a structured component:

$$W = g\chi + \frac{1}{N}mn^\top. \quad (11)$$

327 The random component  $g\chi$  has strength  $g$ , and random component weights are Gaussian dis-  
 328 tributed  $\chi_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$ . The structured component  $\frac{1}{N}mn^\top$  has entries of  $m$  and  $n$  drawn from  
 329 Gaussian distributions  $m_i \sim \mathcal{N}(M_m, 1)$  and  $n_i \sim \mathcal{N}(M_n, 1)$ . Recent theoretical work derives the  
 330 low-dimensional response properties of low-rank networks from statistical parameterizations of their  
 331 connectivity, such as  $z = [g, M_m, M_n]$  [30]. We used EPI to infer the parameterizations of rank-  
 332 1 RNNs solving an example task, enabling discovery of properties of connectivity that result in  
 333 different types of error in the computation.

334 The task we consider is Gaussian posterior conditioning: calculate the parameters of a posterior  
 335 distribution induced by a prior  $p(\mu_y) = \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 1)$  and a likelihood  $p(y|\mu_y) = \mathcal{N}(\mu_y, \sigma_y^2 =$   
 336 1), given a single observation  $y$ . Conjugacy offers the result analytically;  $p(\mu_y|y) = \mathcal{N}(\mu_{post}, \sigma_{post}^2)$ ,  
 337 where:

$$\mu_{post} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y}{\sigma_y^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_y^2}} \quad \sigma_{post}^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_y^2}}. \quad (12)$$

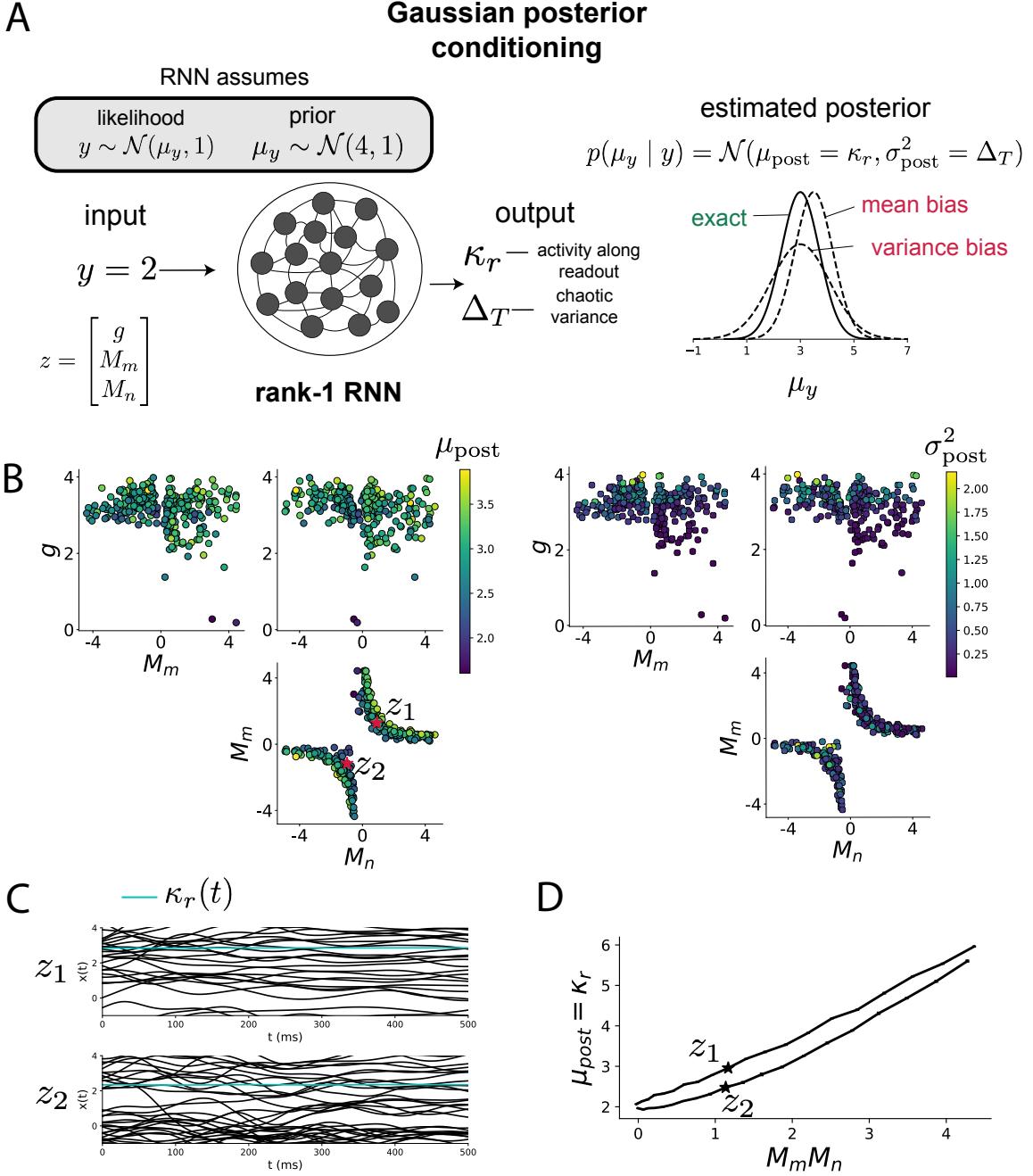


Figure 5: Sources of error in an RNN solving a simple task. A. (left) A rank-1 RNN executing a Gaussian posterior conditioning computation on  $\mu_y$ . (right) Error in this computation can come from over- or underestimating the posterior mean or variance. B. EPI distribution of rank-1 RNNs executing Gaussian posterior conditioning. Samples are colored by (left) posterior mean  $\mu_{\text{post}} = \kappa_r$  and (right) posterior variance  $\sigma_{\text{post}}^2 = \Delta_T$ . C. Finite-size network simulations of 2,000 neurons with parameters  $z_1$  and  $z_2$  sampled from the inferred distribution. Activity along readout  $\kappa_r$  (cyan) is stable despite chaotic fluctuations. D. The posterior mean computed by RNNs parameterized by  $z_1$  and  $z_2$  perturbed in the dimension of the product of  $M_m$  and  $M_n$ . Means and standard errors are shown across 10 realizations of 2,000-neuron networks.

338 To solve this Gaussian posterior conditioning task, the RNN response to a constant input  $h =$   
 339  $yr + (n - M_n)$  must equal the posterior mean along readout vector  $r$ , where

$$\kappa_r = \frac{1}{N} \sum_{j=1}^N r_j \phi(x_j). \quad (13)$$

340 Additionally, the amount of chaotic variance  $\Delta_T$  must equal the posterior variance. Theory for  
 341 low-rank RNNs allows us to express  $\kappa_r$  and  $\Delta_T$  in terms of each other through a solvable system of  
 342 nonlinear equations (see Section 5.2.4) [30]. This theory facilitates the mathematical formalization  
 343 of task execution into an emergent property, where the emergent property statistics of the RNN  
 344 activity are  $\kappa_r$  and  $\Delta_T$ , and the emergent property values are the ground truth posterior mean  
 345  $\mu_{\text{post}}$  and variance  $\sigma_{\text{post}}^2$ :

$$\mathbb{E} \begin{bmatrix} \kappa_r \\ \Delta_T \\ (\kappa_r - \mu_{\text{post}})^2 \\ (\Delta_T - \sigma_{\text{post}}^2)^2 \end{bmatrix} = \begin{bmatrix} \mu_{\text{post}} \\ \sigma_{\text{post}}^2 \\ 0.1 \\ 0.1 \end{bmatrix}. \quad (14)$$

346 We chose a substantial amount of variance in these emergent property statistics, so that the inferred  
 347 distribution resulted in RNNs with a variety of errors in their solutions to the gaussian posterior  
 348 conditioning problem.

349 EPI was used to learn distributions of RNN connectivity properties  $z = [g, M_m, M_n]$  executing  
 350 Gaussian posterior conditioning given an input of  $y = 2$ , where the true posterior is  $\mu_{\text{post}} = 3$  and  
 351  $\sigma_{\text{post}} = 0.5$  (Fig. 5A). We examined the nature of the over- and under-estimation of the posterior  
 352 means (Fig. 5B left) and variances (Fig. 5B right) in the inferred distributions (300 samples).  
 353 The symmetry in the  $M_m$ - $M_n$  plane, suggests a degeneracy in the product of  $M_m$  and  $M_n$  (Fig.  
 354 5B). Indeed,  $M_m M_n$  strongly determines the posterior mean ( $r = 0.62, p < 10^{-4}$ ). Furthermore,  
 355 the random strength  $g$  strongly determines the chaotic variance ( $r = 0.56, p < 10^{-4}$ ). Neither of  
 356 these observations were obvious from what mathematical analysis is available in networks of this  
 357 type (see Section 5.2.4). While the link between random strength  $g$  and chaotic variance  $\Delta_T$  (and  
 358 resultingly posterior variance in this problem) is well-known [3], the distribution admits a novel  
 359 hypothesis: the estimation of the posterior mean by the RNN increases with  $M_m M_n$ .

360 We tested this prediction by taking parameters  $z_1$  and  $z_2$  as representative samples from the positive  
 361 and negative  $M_m$ - $M_n$  quadrants, respectively. Instead of using the theoretical predictions shown in  
 362 Figure 5B, we simulated finite-size realizations of these networks with 2,000 neurons (e.g. Fig. 5C).  
 363 We perturbed these parameter choices by  $M_m M_n$  clarifying that the posterior mean can be directly

364 controlled in this way (Fig. 5D;  $p < 10^{-4}$ ), see Section 5.2.4). Thus, EPI confers a clear picture  
365 of error in this computation: the product of the low rank vector means  $M_m$  and  $M_n$  modulates  
366 the estimated posterior mean while the random strength  $g$  modulates the estimated posterior  
367 variance. This novel procedure of inference on reduced parameterizations of RNNs conditioned on  
368 the emergent property of task execution is generalizable to other settings modeled in [30] like noisy  
369 integration and context-dependent decision making (Fig. S5).

## 370 4 Discussion

### 371 4.1 EPI is a general tool for theoretical neuroscience

372 Biologically realistic models of neural circuits are comprised of complex nonlinear differential equa-  
373 tions, making traditional theoretical analysis and statistical inference intractable. We advance the  
374 capabilities of statistical inference in theoretical neuroscience by presenting EPI, a deep inference  
375 methodology for learning parameter distributions of theoretical models performing neural compu-  
376 tation. We have demonstrated the utility of EPI on biological models (STG), intermediate-level  
377 models of interacting genetically- and functionally-defined neuron-types (V1, SC), and the most  
378 abstract of models (RNNs). We are able to condition both deterministic and stochastic models on  
379 low-level emergent properties like spiking frequency of membrane potentials, as well as high-level  
380 cognitive function like posterior conditioning. Technically, EPI is tractable when the emergent  
381 property statistics are continuously differentiable with respect to the model parameters, which is  
382 very often the case; this emphasizes the general applicability of EPI.

383 In this study, we have focused on applying EPI to low dimensional parameter spaces of models  
384 with low dimensional dynamical states. These choices were made to present the reader with a  
385 series of interpretable conclusions, which is more challenging in high dimensional spaces. In fact,  
386 EPI should scale reasonably to high dimensional parameter spaces, as the underlying technology has  
387 produced state-of-the-art performance on high-dimensional tasks such as texture generation [25]. Of  
388 course, increasing the dimensionality of the dynamical state of the model makes optimization more  
389 expensive, and there is a practical limit there as with any machine learning approach. Although,  
390 theoretical approaches (e.g. [30]) can be used to reason about the wholistic activity of such high  
391 dimensional systems by introducing some degree of additional structure into the model.

392 **4.2 Novel hypotheses from EPI**

393 In neuroscience, machine learning has primarily been used to reveal structure in large-scale neural  
394 datasets [50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60] (see review, [16]). Such careful inference procedures  
395 are developed for these statistical models allowing precise, quantitative reasoning, which clarifies  
396 the way data informs beliefs about the model parameters. However, these statistical models lack  
397 resemblance to the underlying biology, making it unclear how to go from the structure revealed by  
398 these methods, to the neural mechanisms giving rise to it. In contrast, theoretical neuroscience has  
399 focused on careful mechanistic modeling and the production of emergent properties of computation.  
400 The careful steps of *i.*) model design and *ii.*) emergent property definition, are followed by *iii.)*  
401 practical inference methods resulting in an opaque characterization of the way model parameters  
402 govern computation. In this work, we replaced this opaque procedure of parameter identification  
403 in theoretical neuroscience with emergent property inference, opening the door to careful inference  
404 in careful models of neural computation.

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

411 In Section 3.3, we advanced the tractability of low-dimensional neural circuit models by showing  
412 that EPI offers insights about cell-type specific input-responsivity that cannot be afforded through  
413 the available linear analytical methods [28, 43, 44]. By flexibly conditioning this V1 model on  
414 different emergent properties, we performed an exploratory analysis of a *model* rather than a  
415 dataset, generating a set of testable hypotheses, which were proved out. Furthermore, exploratory  
416 analyses can be directed towards formulating hypotheses of a specific form. For example, model  
417 parameter dependencies on behavioral performance can be assessed by using EPI to condition on  
418 various levels of task accuracy (See Section 3.4). This analysis identified experimentally testable  
419 predictions (proved out *in-silico*) of patterns of effective connectivity in SC that should be correlated  
420 with increased performance.

421 In our final analysis, we presented a novel procedure for doing statistical inference on interpretable  
422 parameterizations of RNNs executing simple tasks. Specifically, we analyzed RNNs solving a pos-

423 terior conditioning problem in the spirit of [63, 64]. This methodology relies on recently extended  
424 theory of responses in random neural networks with low-rank structure [30]. While we focused  
425 on rank-1 RNNs, which were sufficient for solving this task, this inference procedure generalizes  
426 to RNNs of greater rank necessary for more complex tasks. The ability to apply the probabilistic  
427 model selection toolkit to RNNs should prove invaluable as their use in neuroscience increases.

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

435 **Acknowledgements:**

436 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-  
437 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-  
438 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-  
439 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program  
440 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James  
441 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

442 **Data availability statement:**

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

445 **Code availability statement:**

446 The software written for the current study is available from the corresponding author upon rea-  
447 sonable request.

448 **References**

- 449 [1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.  
450 [2] John J Hopfield. Neural networks and physical systems with emergent collective computational  
451 abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

- 452 [3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural  
453 networks. *Physical review letters*, 61(3):259, 1988.
- 454 [4] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-  
455 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,  
456 17(11):4382–4388, 1997.
- 457 [5] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in  
458 perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 459 [6] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.  
460 A framework for parameter estimation and model selection from experimental data in systems  
461 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 462 [7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Confer-  
463 ence on Learning Representations*, 2014.
- 464 [8] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation  
465 and variational inference in deep latent gaussian models. *International Conference on Machine  
466 Learning*, 2014.
- 467 [9] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical  
468 neural population models through nonlinear embeddings. In *Advances in neural information  
469 processing systems*, pages 163–171, 2016.
- 470 [10] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear  
471 dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 472 [11] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders.  
473 *bioRxiv*, page 399246, 2018.
- 474 [12] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,  
475 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg,  
476 et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature  
477 methods*, page 1, 2018.
- 478 [13] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M  
479 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.  
480 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

- 481 [14] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R  
482 Datta. Composing graphical models with neural networks for structured representations and  
483 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 484 [15] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,  
485 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:  
486 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural  
487 Information Processing Systems*, 2019.
- 488 [16] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-  
489 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,  
490 2018.
- 491 [17] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula  
492 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially  
493 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–  
494 1929, 2009.
- 495 [18] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics.  
496 *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.
- 497 [19] David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models.  
498 2014.
- 499 [20] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R  
500 Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology,  
501 and beyond. *The Journal of chemical physics*, 143(1):07B201\_1, 2015.
- 502 [21] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-  
503 free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–  
504 5533, 2017.
- 505 [22] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.  
506 *International Conference on Machine Learning*, 2015.
- 507 [23] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
508 *arXiv preprint arXiv:1605.08803*, 2016.

- 509 [24] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density  
510 estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.
- 511 [25] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow  
512 networks. *International Conference on Learning Representations*, 2017.
- 513 [26] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness,  
514 and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.
- 515 [27] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an  
516 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.  
517 *Neuron*, 77(5):845–858, 2013.
- 518 [28] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and vi-  
519 sual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,  
520 115(3):1399–1409, 2016.
- 521 [29] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J  
522 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor  
523 routing. *bioRxiv*, page 245613, 2018.
- 524 [30] Francesca Mastrogiovanni and Srdjan Ostojic. Linking connectivity, dynamics, and computa-  
525 tions in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- 526 [31] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-  
527 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 528 [32] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit  
529 parameters. *Nature neuroscience*, 7(12):1345, 2004.
- 530 [33] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,  
531 1957.
- 532 [34] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:  
533 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 534 [35] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise  
535 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.

- 536 [36] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of  
537 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 538 [37] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory  
539 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–  
540 592, 2009.
- 541 [38] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear  
542 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,  
543 85(2):402–417, 2015.
- 544 [39] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and  
545 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,  
546 5(10):793, 2004.
- 547 [40] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of  
548 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-  
549 biology*, 71(1):45–61, 2011.
- 550 [41] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex:  
551 From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 552 [42] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-  
553 bition of inhibition in visual cortex: the logic of connections between molecularly distinct  
554 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 555 [43] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing Wang.  
556 Paradoxical response reversal of top- down modulation in cortical circuits with three interneu-  
557 ron types. *Elife*, 6:1–15, 2017.
- 558 [44] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying  
559 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 560 [45] (2018) Allen Institute for Brain Science. Layer 4 model of v1. available from:  
561 <https://portal.brain-map.org/explore/models/l4-mv1>.
- 562 [46] Yazan N Billeh, Binghuang Cai, Sergey L Gratiy, Kael Dai, Ramakrishnan Iyer, Nathan W  
563 Gouwens, Reza Abbasi-Asl, Xiaoxuan Jia, Joshua H Siegle, Shawn R Olsen, et al. Systematic

- 564 integration of structural and functional data into multi-scale models of mouse primary visual  
565 cortex. *bioRxiv*, page 662189, 2019.
- 566 [47] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain  
567 regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.
- 568 [48] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current  
569 opinion in neurobiology*, 46:1–6, 2017.
- 570 [49] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-  
571 dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.
- 572 [50] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,  
573 13(8):1713–1720, 2001.
- 574 [51] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.  
575 A statistical paradigm for neural spike train decoding applied to position prediction from  
576 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–  
577 7425, 1998.
- 578 [52] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding  
579 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 580 [53] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A  
581 point process framework for relating neural spiking activity to spiking history, neural ensemble,  
582 and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 583 [54] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan  
584 Segev. A novel multiple objective optimization framework for constraining conductance-based  
585 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 586 [55] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and  
587 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis  
588 of neural population activity. In *Advances in neural information processing systems*, pages  
589 1881–1888, 2009.
- 590 [56] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In  
591 *Advances in neural information processing systems*, pages 1692–1700, 2011.

- 592 [57] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W  
593 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.  
594 *Science*, 349(6244):184–187, 2015.
- 595 [58] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq  
596 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in  
597 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 598 [59] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable  
599 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th Inter-*  
600 *national Conference on Machine Learning*, 2019.
- 601 [60] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.  
602 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.  
603 *Nature Communications*, 10(4933), 2019.
- 604 [61] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give  
605 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of*  
606 *Sciences*, 115(13):3464–3469, 2018.
- 607 [62] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emer-  
608 gent dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint*  
609 *arXiv:1605.04463*, 2016.
- 610 [63] Xaq Pitkow and Dora E Angelaki. Inference in the brain: statistics flowing in redundant  
611 population codes. *Neuron*, 94(5):943–953, 2017.
- 612 [64] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like  
613 dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*,  
614 page 696088, 2019.
- 615 [65] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,  
616 2019.
- 617 [66] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for  
618 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- 619 [67] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial  
620 Intelligence and Statistics*, pages 814–822, 2014.

- 621 [68] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-  
622 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience*  
623 *Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 624 [69] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,  
625 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with  
626 degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 627 [70] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and  
628 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In  
629 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 630 [71] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical inference for  
631 analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin, Germany*,  
632 2019.
- 633 [72] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan  
634 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P  
635 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural  
636 dynamics. *bioRxiv*, page 838383, 2019.
- 637 [73] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-  
638 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural  
639 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 640 [74] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and  
641 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 642 [75] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International  
643 Conference on Learning Representations*, 2015.
- 644 [76] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
645 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 646 [77] Nicolas Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking  
647 neurons. *Journal of computational neuroscience*, 8(3):183–208, 2000.
- 648 [78] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and  
649 saving energy in wireless communication. *science*, 304(5667):78–80, 2004.

- 650 [79] David Sussillo and Larry F Abbott. Generating coherent patterns of activity from chaotic  
651 neural networks. *Neuron*, 63(4):544–557, 2009.

652 **5 Methods**

653 **5.1 Emergent property inference (EPI)**

654 Emergent property inference (EPI) learns distributions of theoretical model parameters that pro-  
 655 duce emergent properties of interest by combining ideas from maximum entropy flow networks  
 656 (MEFNs) [25] and likelihood-free variational inference (LFVI) [21]. Consider model parameteri-  
 657 zation  $z$  and data  $x$  which has an intractable likelihood  $p(x | z)$  defined by a model simulator of  
 658 which samples are available  $x \sim p(x | z)$ . EPI optimizes a distribution  $q_\theta(z)$  (itself parameterized  
 659 by  $\theta$ ) of model parameters  $z$  to produce an emergent property of interest  $\mathcal{B}$ ,

$$\mathcal{B} \triangleq \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] = \mu. \quad (15)$$

660 Precisely, the emergent property statistics  $T(x)$  must equal the emergent property values  $\mu$ , in  
 661 expectation over the EPI distribution of parameters  $q_\theta(z)$  and the distribution of simulated activity  
 662  $p(x | z)$ . This is a viable way to represent emergent properties in theoretical models, as we have  
 663 demonstrated in the main text, and enables the EPI optimization.

664 With EPI, we use deep probability distributions to learn flexible approximations to model parameter  
 665 distributions  $q_\theta(z)$ . In deep probability distributions, a simple random variable  $w \sim q_0(w)$  is  
 666 mapped deterministically via a sequence of deep neural network layers ( $f_1, \dots, f_l$ ) parameterized by  
 667 weights and biases  $\theta$  to the support of the distribution of interest:

$$z = f_\theta(\omega) = f_l(\dots f_1(w)). \quad (16)$$

668 Given a simulator defined by a theoretical model  $x \sim p(x | z)$  and some emergent property of  
 669 interest  $\mathcal{B}$ ,  $q_\theta(z)$  is optimized via the neural network parameters  $\theta$  to find a maximally entropic  
 670 distribution  $q_\theta^*$  within the deep variational family  $\mathcal{Q}$  producing the emergent property:

$$\begin{aligned} q_\theta^*(z) &= \operatorname{argmax}_{q_\theta \in \mathcal{Q}} H(q_\theta(z)) \\ &\text{s.t. } \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] = \mu. \end{aligned} \quad (17)$$

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

$$H(q_\theta(z)) = \int -q_\theta(z) \log(q_\theta(z)) dz = \mathbb{E}_{z \sim q_\theta} [-\log(q_\theta(z))] = \mathbb{E}_{w \sim q_0} [-\log(q_\theta(f_\theta(w)))]. \quad (18)$$

676 Thus, the gradient of the entropy of the deep probability distribution can be estimated as an  
 677 average of gradients of the log density of samples  $z$ :

$$\nabla_{\theta} H(q_{\theta}(z)) = \mathbb{E}_{w \sim q_0} [-\nabla_{\theta} \log(q_{\theta}(f_{\theta}(w)))]. \quad (19)$$

678 In EPI, MEFNs are purposed towards variational learning of model parameter distributions. A  
 679 closely related methodology, variational inference, uses optimization to approximate posterior dis-  
 680 tributions [66]. Standard methods like stochastic gradient variational Bayes [7] or black box varia-  
 681 tional inference [67] simply do not work for inference in theoretical models of neural circuits, since  
 682 they require tractable likelihoods  $p(x | z)$ . Work on likelihood-free variational inference (LFVI) [21],  
 683 which like EPI seeks to do inference in models with intractable likelihoods, employs an additional  
 684 deep neural network as a ratio estimator, enabling an estimation of the optimization objective for  
 685 variational inference. Like LFVI, EPI can be framed as variational inference (see Section 5.1.4).  
 686 But, unlike LFVI, EPI uses a single deep network to learn a distribution and is optimized to pro-  
 687 duce an emergent property, rather than condition on data points. Optimizing the EPI objective is  
 688 a technological challenge, the details of which we elaborate in Section 5.1.2. Before going through  
 689 those details, we ground this optimization in a toy example.

690 We note that, during our preparation and early presentation of this work [68, 69], another work  
 691 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural  
 692 circuits ([70, 71, 72], preprint posted simultaneously with this preprint). We are encouraged by  
 693 this general problem being recognized by others in the community, and we emphasize that these  
 694 works offer complementary neuroscientific contributions (different theoretical models of focus) and  
 695 use different technical methodologies (ours is built on our prior work [25], theirs similarly [73]).  
 696 These distinct methodologies and scientific investigations emphasize the increased importance and  
 697 timeliness of both works.

### 698 5.1.1 Example: 2D LDS

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

$$\tau \frac{dx}{dt} = Ax \quad (20)$$

702 with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (21)$$

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

$$\mathbb{E}[T(x)] \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \mu. \quad (22)$$

710

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

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (23)$$

714 Importantly, even though  $\mathbb{E}_{x \sim p(x|z)}[T(x)]$  is calculable directly via a closed form function and  
 715 does not require simulation, we cannot derive the distribution  $q_\theta^*$  directly. This fact is due to the  
 716 formally hard problem of the backward mapping: finding the natural parameters  $\eta$  from the mean  
 717 parameters  $\mu$  of an exponential family distribution [74]. Instead, we used EPI to approximate this  
 718 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two  
 719 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped  
 720 onto a support of  $z_i \in [-10, 10]$ . (see Section 5.1.3).

721 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the  
 722 parameter distribution. To validate our method, we analytically derived the contours of the prob-  
 723 ability density from the emergent property statistics and values. In the  $a_1$ - $a_4$  plane, the black  
 724 line at  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ , dotted black line at the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$ ,  
 725 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  
 726 of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired  
 727 statistical constraints and model degeneracy in the sum of  $a_1$  and  $a_4$ . Intuitively, the parameters  
 728 equivalent with respect to emergent property statistic  $\text{real}(\lambda_1)$  have similar log densities.

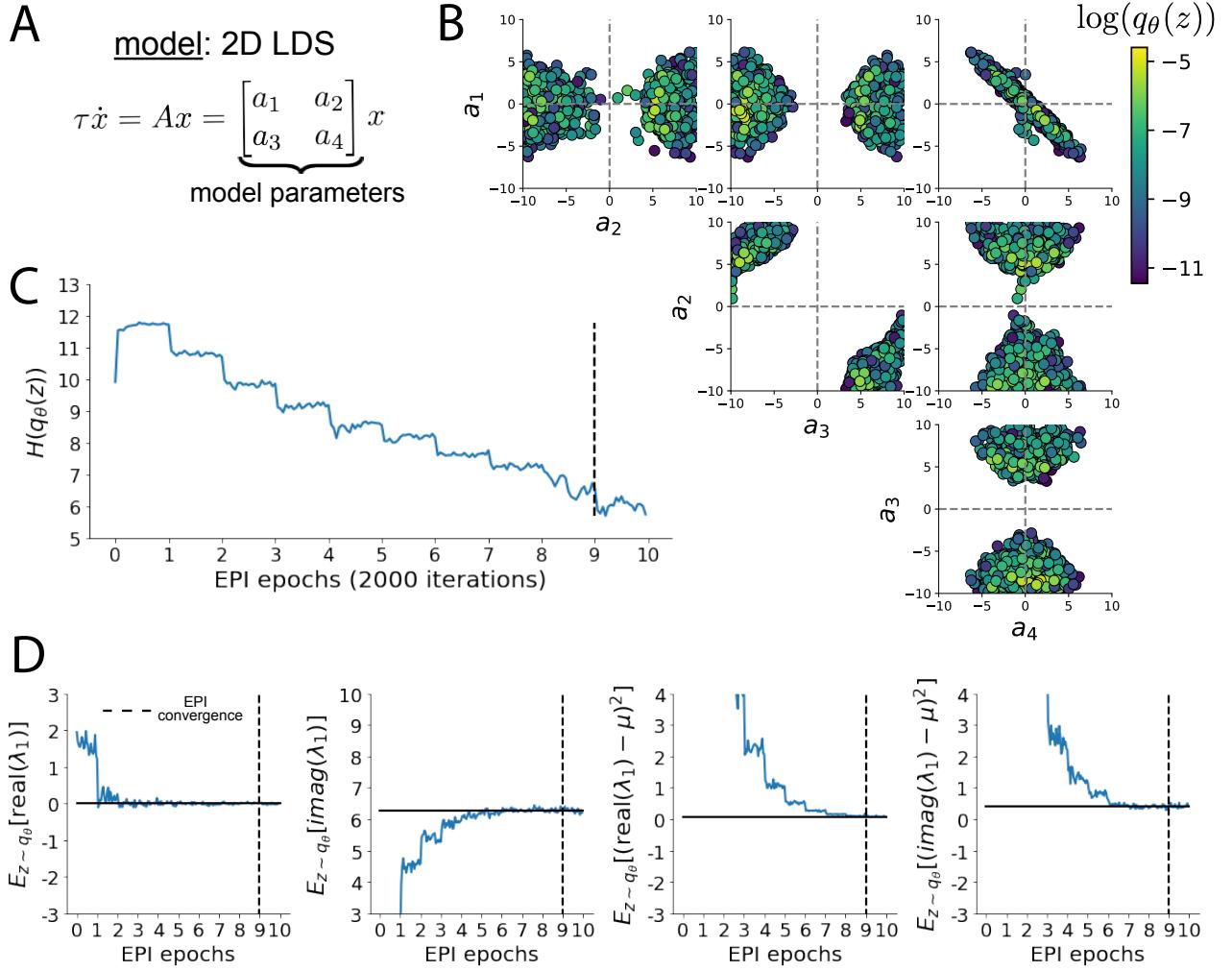


Fig. S1: A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix  $A$  are the parameters. B. The EPI distribution for a two-dimensional linear dynamical system with  $\tau = 1$  that produces an average of 1Hz oscillations with some small amount of variance. Dashed 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.

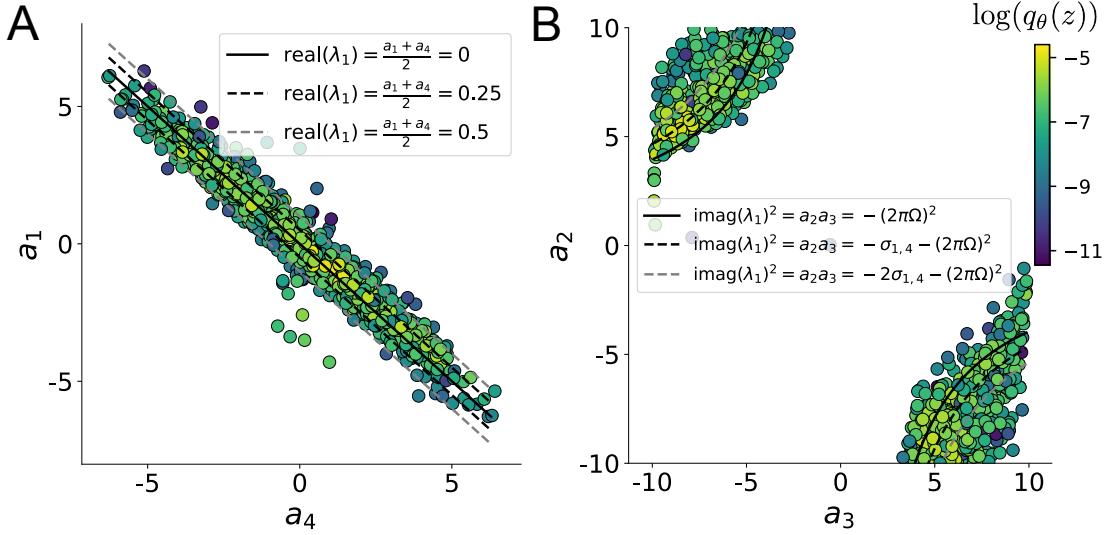


Fig. S2: A. Probability contours in the  $a_1$ - $a_4$  plane were derived from the relationship to emergent property statistic of growth/decay factor  $\text{real}(\lambda_1)$ . B. Probability contours in the  $a_2$ - $a_3$  plane were derived from the emergent property statistic of oscillation frequency  $2\pi\text{imag}(\lambda_1)$ .

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1 a_4 - a_2 a_3}{\tau}}, & \text{if } a_1 a_4 < a_2 a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (24)$$

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

$$\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 \quad (25)$$

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

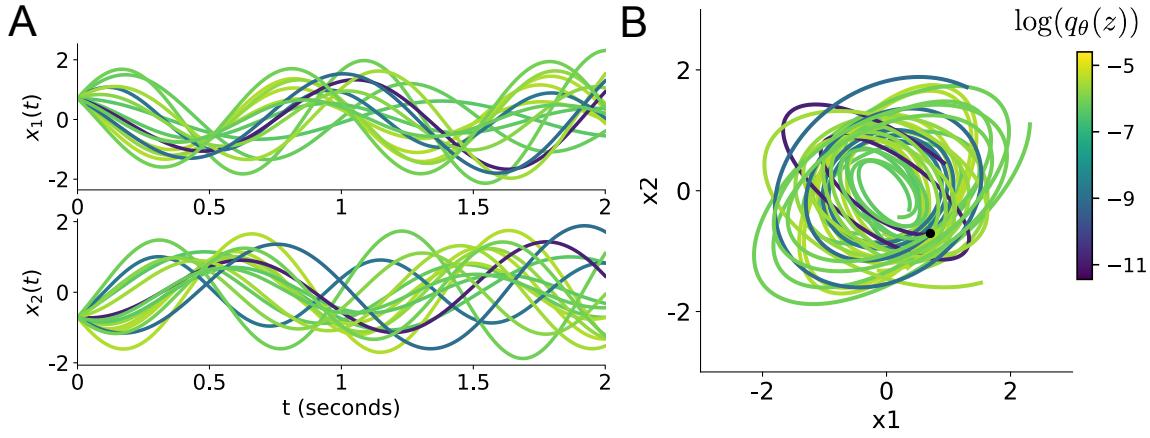


Fig. S3: Sampled dynamical systems  $z \sim q_\theta(z)$  and their simulated activity from  $x(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$  colored by log probability. A. Each dimension of the simulated trajectories throughout time. B The simulated trajectories in phase space.

#### 742 5.1.2 Augmented Lagrangian optimization

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

$$L(\theta; \eta, c) = -H(q_\theta) + \eta^\top R(\theta) + \frac{c}{2} \|R(\theta)\|^2 \quad (26)$$

745 where  $R(\theta) = \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x) - \mu]]$ ,  $\eta \in \mathbb{R}^m$  are the Lagrange multipliers where  $m = |\mu| = |T(x)|$ , and  $c$  is the penalty coefficient. These Lagrange multipliers are closely related to the natural  
 746 parameters of exponential families (see Section 5.1.4). Deep neural network weights and biases  $\theta$  of  
 747 the deep probability distribution are optimized according to Equation 26 using the Adam optimizer  
 748 with its standard parameterization [75].  $\eta$  is initialized to the zero vector and adapted following  
 749 each augmented Lagrangian epoch, which is a period of optimization with fixed  $(\eta, c)$  for a given  
 750 number of stochastic optimization iterations. A low value of  $c$  is used initially, and conditionally  
 751 increased after each epoch based on constraint error reduction. For example, the initial value of  
 752  $c$  was  $c_0 = 10^{-3}$  during EPI with the oscillating 2D LDS (Fig. S1C). The penalty coefficient is  
 753 updated based on the result of a hypothesis test regarding the reduction in constraint violation. The  
 754 p-value of  $\mathbb{E}[|R(\theta_{k+1})|] > \gamma \mathbb{E}[|R(\theta_k)|]$  is computed, and  $c_{k+1}$  is updated to  $\beta c_k$  with probability  
 755  $1-p$ . The other update rule is  $\eta_{k+1} = \eta_k + c_k \frac{1}{n} \sum_{i=1}^n (T(x^{(i)}) - \mu)$  given a batch size  $n$ . Throughout  
 756 the study,  $\beta = 4.0$ ,  $\gamma = 0.25$ , and the batch size was a hyperparameter, which varied according to  
 757 the application of EPI.

759 The intention is that  $c$  and  $\eta$  start at values encouraging entropic growth early in optimization.  
760 With each training epoch in which the update rule for  $c$  is invoked by unsatisfactory constraint  
761 error reduction, the constraint satisfaction terms are increasingly weighted, resulting in a decreased  
762 entropy. This encourages the discovery of suitable regions of parameter space, and the subsequent  
763 refinement of the distribution to produce the emergent property. In the oscillating 2D LDS example,  
764 each augmented Lagrangian epoch ran for 2,000 iterations (Fig. S1C-D). Notice the initial entropic  
765 growth, and subsequent reduction upon each update of  $\eta$  and  $c$ . The momentum parameters of the  
766 Adam optimizer were reset at the end of each augmented Lagrangian epoch.

767 Rather than starting optimization from some  $\theta$  drawn from a randomized distribution, we found  
768 that initializing  $q_\theta(z)$  to approximate an isotropic Gaussian distribution conferred more stable, con-  
769 sistent optimization. The parameters of the Gaussian initialization were chosen on an application-  
770 specific basis. Throughout the study, we chose isotropic Gaussian initializations with mean  $\mu_{\text{init}}$  at  
771 the center of the distribution support and some standard deviation  $\sigma_{\text{init}}$ , except for one case, where  
772 an initialization informed by random search was used (see Section 5.2.2).

773 To assess whether EPI distribution  $q_\theta(z)$  produces the emergent property, we defined a hypothesis  
774 testing convergence criteria. The algorithm has converged when a null hypothesis test of constraint  
775 violations  $R(\theta)_i$  being zero is accepted for all constraints  $i \in \{1, \dots, m\}$  at a significance threshold  
776  $\alpha = 0.05$ . This significance threshold is adjusted through Bonferroni correction according to the  
777 number of constraints  $m$ . The p-values for each constraint are calculated according to a two-tailed  
778 nonparametric test, where 200 estimations of the sample mean  $R(\theta)^i$  are made from  $k$  resamplings  
779 of  $z$  from a finite sample of size  $n$  taken at the end of the augmented Lagrangian epoch.  $k$  is  
780 determined by a fraction of the batch size  $\nu$ , which varies according to the application. In the  
781 linear two-dimensional system example, we used a batch size of  $n = 1000$  and set  $\nu = 0.1$  resulting  
782 in convergence after the ninth epoch of optimization. (Fig. S1C-D black dotted line).

783 When assessing the suitability of EPI for a particular modeling question, there are some important  
784 technical considerations. First and foremost, as in any optimization problem, the defined emergent  
785 property should always be appropriately conditioned (constraints should not have wildly different  
786 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution  
787 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-  
788 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,  
789 one should consider the computational cost of the gradient calculations. In the best circumstance,  
790 there is a simple, closed form expression (e.g. Section 5.1.1) for the emergent property statistic

791 given the model parameters. On the other end of the spectrum, many forward simulation iterations  
 792 may be required before a high quality measurement of the emergent property statistic is available  
 793 (e.g. Section 5.2.1). In such cases, optimization will be expensive.

794 **5.1.3 Normalizing flows**

795 Deep probability models typically consist of several layers of fully connected neural networks.  
 796 When each neural network layer is restricted to be a bijective function, the sample density can be  
 797 calculated using the change of variables formula at each layer of the network. For  $z' = f(z)$ ,

$$q(z') = q(f^{-1}(z')) \left| \det \frac{\partial f^{-1}(z')}{\partial z'} \right| = q(z) \left| \det \frac{\partial f(z)}{\partial z} \right|^{-1}. \quad (27)$$

798 However, this computation has cubic complexity in dimensionality for fully connected layers. By  
 799 restricting our layers to normalizing flows [22] – bijective functions with fast log determinant Ja-  
 800 cobian computations, we can tractably optimize deep generative models with objectives that are a  
 801 function of sample density, like entropy. Most of our analyses use either a planar flow [22] or real  
 802 NVP [76], which have proven effective in our architecture searches. Planar flow architectures are  
 803 specified by the number of planar bijection layers used, while real NVP architectures are specified  
 804 by the number of masks, neural network layers per mask, units per layer, and batch normalization  
 805 momentum parameter.

806 **5.1.4 Emergent property inference as variational inference in an exponential family**

807 Now that we have fully described the EPI method, we consider its broader contextualization as a  
 808 statistical method and its relation to Bayesian inference. In Bayesian inference a prior belief about  
 809 model parameters  $z$  is formalized into a prior distribution  $p(z)$ , and the statistical model capturing  
 810 the effect of  $z$  on observed data points  $x$  is formalized in the likelihood distribution  $p(x | z)$ . In  
 811 Bayesian inference, we obtain a posterior distribution  $p(z | x)$ , which captures how the data inform  
 812 our knowledge of model parameters using Bayes’ rule:

$$p(z | x) = \frac{p(x | z)p(z)}{p(x)}. \quad (28)$$

813 The posterior distribution is analytically available when the prior is conjugate with the likelihood.  
 814 However, conjugacy is rare in practice, and alternative methods, such as variational inference [66],  
 815 are utilized.

816 As we compare EPI to variational inference, it is important to consider that EPI is a maximum  
 817 entropy method, and that maximum entropy methods have a fundamental relationship with expo-  
 818 nential family distributions. A maximum entropy distribution of form:

$$\begin{aligned} p^*(z) &= \operatorname{argmax}_{p \in \mathcal{P}} H(p(z)) \\ \text{s.t. } \mathbb{E}_{z \sim p}[T(z)] &= \mu. \end{aligned} \quad (29)$$

819 will have probability density in the exponential family:

$$p^*(z) \propto \exp(\eta^\top T(z)). \quad (30)$$

820 The mappings between the mean parameterization  $\mu$  and the natural parameterization  $\eta$  are for-  
 821 mally hard to identify [74].

822 Now, consider the goal of doing variational inference with an exponential family posterior dis-  
 823 tribution  $p(z | x)$ . We use the following abbreviated notation to collect the base measure  $b(z)$   
 824 and sufficient statistics  $T(z)$  into  $\tilde{T}(z)$  and likewise concatenate a 1 onto the end of the natural  
 825 parameter  $\tilde{\eta}(x)$ . The log normalizing constant  $A(\eta(x))$  remains unchanged:

$$\begin{aligned} p(z | x) &= b(z) \exp\left(\eta(x)^\top T(z) - A(\eta(x))\right) = \exp\left(\begin{bmatrix} \eta(x) \\ 1 \end{bmatrix}^\top \begin{bmatrix} T(z) \\ b(z) \end{bmatrix} - A(\eta(x))\right). \\ &= \exp\left(\tilde{\eta}(x)^\top \tilde{T}(z) - A(\eta(x))\right) \end{aligned} \quad (31)$$

826 Variational inference with an exponential family posterior distribution uses optimization to mini-  
 827 mize the following divergence [66]:

$$q_\theta^* = \operatorname{argmin}_{q_\theta \in Q} KL(q_\theta || p(z | x)). \quad (32)$$

828  $q_\theta(z)$  is the variational approximation to the posterior with variational parameters  $\theta$ . We can write  
 829 this KL divergence in terms of entropy of the variational approximation:

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

830

$$= -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} [\tilde{\eta}(x)^\top \tilde{T}(z) - A(\eta(x))]. \quad (34)$$

831 As far as the variational optimization is concerned, the log normalizing constant is independent of  
 832  $q_\theta(z)$ , so it can be dropped

$$\operatorname{argmin}_{q_\theta \in Q} KL(q_\theta || p(z | x)) = \operatorname{argmin}_{q_\theta \in Q} -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} [\tilde{\eta}(x)^\top \tilde{T}(z)]. \quad (35)$$

833 Further, we can write the objective in terms of the first moment of the sufficient statistics  $\mu =$   
 834  $\mathbb{E}_{z \sim p(z|x)} [T(z)]$ :

$$= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} \left[ \tilde{\eta}(x)^\top (\tilde{T}(z) - \mu) \right] + \tilde{\eta}(x)^\top \mu, \quad (36)$$

835 which simplifies to

$$= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} \left[ \tilde{\eta}(x)^\top (\tilde{T}(z) - \mu) \right]. \quad (37)$$

836 .

837 In comparison, in emergent property inference (EPI), we solve the following problem:

$$q_\theta^*(z) = \underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta(z)), \text{ s.t. } \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] = \mu. \quad (38)$$

838 The Lagrangian objective (without augmentation) is

$$q_\theta^* = \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \eta_{\text{opt}}^\top \left( \mathbb{E}_{z \sim q_\theta} [\tilde{T}(z)] - \mu \right). \quad (39)$$

839 Thus, as the optimization proceeds,  $\eta_{\text{opt}}^\top$  should converge to the natural parameter  $\tilde{\eta}(x)$  through  
 840 its adaptations in each epoch (see Section 5.1.2).

841 We have shown that there is indeed a clear relationship between Bayesian inference and EPI.  
 842 Specifically, EPI is executing variational inference in an exponential family posterior, whose suffi-  
 843 cient statistics are the emergent property statistics and mean parameterization are the emergent  
 844 property values. However, in EPI we have not specified a prior distribution, or collected data,  
 845 which can inform us about model parameters. Instead we have a mathematical specification of  
 846 an emergent property, which the model must produce, and a maximum entropy selection prin-  
 847 ciple. Accordingly, we replace the notation of  $p(z | x)$  with  $p(z | \mathcal{B})$  conceptualizing an inferred  
 848 distribution that obeys emergent property  $\mathcal{B}$  (see Section 5.1).

## 849 5.2 Theoretical models

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

### 852 5.2.1 Stomatogastric ganglion

853 We analyze how the parameters  $z = [g_{\text{el}}, g_{\text{synA}}]$  govern the emergent phenomena of network syncing  
 854 in a model of the stomatogastric ganglion (STG) [27] shown in Figure 1A with activity  $x =$   
 855  $[x_{\text{f1}}, x_{\text{f2}}, x_{\text{hub}}, x_{\text{s1}}, x_{\text{s2}}]$ , using the same hyperparameter choices as Gutierrez et al. Each neuron's

856 membrane potential  $x_\alpha(t)$  for  $\alpha \in \{\text{f1, f2, hub, s1, s2}\}$  is the solution of the following differential  
 857 equation:

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

858 The membrane potential of each neuron is affected by the leak, calcium, potassium, hyperpolariza-  
 859 tion, electrical and synaptic currents, respectively, which are functions of all membrane potentials  
 860 and the conductance parameters  $z$ . The capacitance of the cell membrane was set to  $C_m = 1nF$ .  
 861 Specifically, the currents are the difference in the neuron's membrane potential and that current  
 862 type's reversal potential multiplied by a conductance:

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

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

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

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

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

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

863 The reversal potentials were set to  $V_{\text{leak}} = -40mV$ ,  $V_{Ca} = 100mV$ ,  $V_K = -80mV$ ,  $V_{hyp} = -20mV$ ,  
 864 and  $V_{syn} = -75mV$ . The other conductance parameters were fixed to  $g_{\text{leak}} = 1 \times 10^{-4}\mu S$ .  $g_{Ca}$ ,  
 865  $g_K$ , and  $g_{hyp}$  had different values based on fast, intermediate (hub) or slow neuron. The fast  
 866 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  
 867 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  
 868 slow conductances had values  $g_{Ca} = 8.5 \times 10^{-3}$ ,  $g_K = 1.5 \times 10^{-2}$ , and  $g_{hyp} = 1.0 \times 10^{-2}$ .

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

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

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

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

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

879

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

880

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

881

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

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

Finally, there is a synaptic gating variable as well:

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

When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear dynamical system.

In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for  $T = 200$  time steps of  $dt = 25ms$ . In EPI, since gradients are taken through the simulation process, the number of time steps are kept modest if possible. The chosen  $dt$  and  $T$  were the most computationally convenient choices yielding accurate frequency measurement. Poor resolution afforded by the discrete Fourier transform motivated the use of an alternative basis of complex exponentials to measure spiking frequency. Instead, we used a basis of complex exponentials with frequencies from 0.0-1.0 Hz at 0.01Hz resolution,  $\Phi = [0.0, 0.01, \dots, 1.0]^\top$

Another consideration was that the frequency spectra of the neuron membrane potentials had several peaks. High-frequency sub-threshold activity obscured the maximum frequency measurement in the complex exponential basis. Accordingly, subthreshold activity was set to zero, and the whole signal was low-pass filtered with a moving average window of length 20. The signal was subsequently mean centered. After this preprocessing, the maximum frequency in the filter bank accurately reflected the firing frequency.

Finally, to differentiate through the maximum frequency identification, we used a sum-of-powers normalization. Let  $\mathcal{X}_\alpha \in \mathcal{C}^{|\Phi|}$  be the complex exponential filter bank dot products with the signal  $x_\alpha \in \mathbb{R}^N$ , where  $\alpha \in \{f1, f2, \text{hub}, s1, s2\}$ . The “frequency identification” vector is

$$v_\alpha = \frac{|\mathcal{X}_\alpha|^\beta}{\sum_{k=1}^N |\mathcal{X}_\alpha(k)|^\beta}. \quad (55)$$

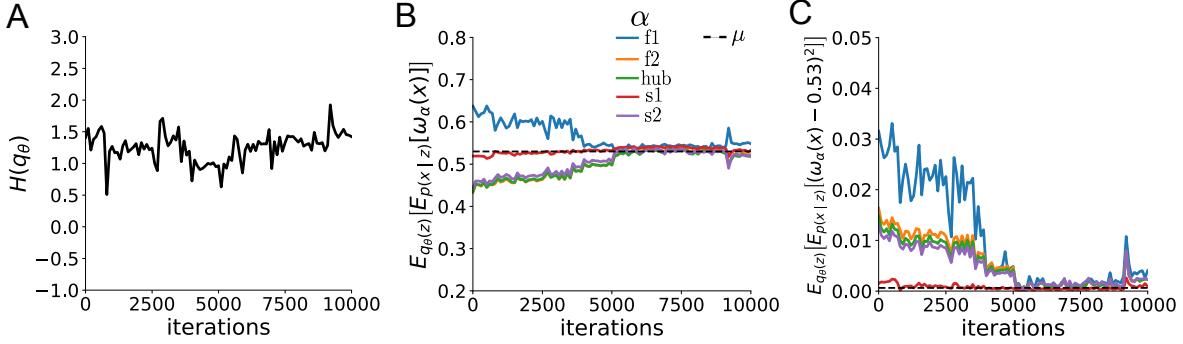


Fig. S4: EPI optimization of the STG model producing network syncing. A. Entropy throughout optimization. B. The first moment emergent property statistics converge to the emergent property values at 10,000 iterations, following the fourth augmented Lagrangian epoch of 2,500 iterations. Since  $q_\theta(z)$  failed to produce enough samples yielding  $\omega_{f1}(x)$  less than 0.53Hz, the convergence criteria were not satisfied after the third epoch at 7,500 iterations. C. The second moment emergent property statistics converge to the emergent property values.

- 903 The frequency is then calculated as  $\omega_\alpha = v_\alpha^\top \Phi$  with  $\beta = 100$ .  
 904 Network syncing, like all other emergent properties in this work, are defined by the emergent  
 905 property statistics and values. The emergent property statistics are the first and second moments  
 906 of the firing frequencies. The first moments were set to 0.53Hz, and the second moments were set  
 907 to 0.025Hz<sup>2</sup>:

$$E \begin{bmatrix} \omega_{f1} \\ \omega_{f2} \\ \omega_{\text{hub}} \\ \omega_{s1} \\ \omega_{s2} \\ (\omega_{f1} - 0.53)^2 \\ (\omega_{f2} - 0.53)^2 \\ (\omega_{\text{hub}} - 0.53)^2 \\ (\omega_{s1} - 0.53)^2 \\ (\omega_{s2} - 0.53)^2 \end{bmatrix} = \begin{bmatrix} 0.53 \\ 0.53 \\ 0.53 \\ 0.53 \\ 0.53 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \end{bmatrix} \quad (56)$$

- 908 for the EPI distribution shown in Fig. 1B. Throughout optimization, the augmented Lagrangian  
 909 parameters  $\eta$  and  $c$ , were updated after each epoch of 2,500 iterations (see Section 5.1.2). The  
 910 optimization converged after four epochs (Fig. S4).

911 For EPI in Fig 2C, we used a real NVP architecture with four masks and two layers of 10 units  
 912 per mask, and batch normalization momentum of 0.99 mapped onto a support of  $z = [g_{\text{el}}, g_{\text{synA}}] \in$   
 913  $[4, 8] \times [0, 4]$ . We used an augmented Lagrangian coefficient of  $c_0 = 10^2$ , a batch size  $n = 300$ ,  
 914 set  $\nu = 0.1$ , and initialized  $q_\theta(z)$  to produce an isotropic Gaussian with mean  $\mu_{\text{init}} = [6, 2]$  with  
 915 standard deviation  $\sigma_{\text{init}} = 0.5$ .

916 We calculated the Hessian at the mode of the inferred EPI distribution. The Hessian of a probability  
 917 model is the second order gradient of the log probability density  $\log q_\theta(z)$  with respect to the  
 918 parameters  $z$ :  $\frac{\partial^2 \log q_\theta(z)}{\partial z \partial z^\top}$ . With EPI, we can examine the Hessian, which is analytically available  
 919 throughout distribution, to indicate the dimensions of parameter space that are sensitive (high  
 920 magnitude eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the  
 921 emergent property produced. In Figure 1B, the eigenvectors of the Hessian  $v_1$  and  $v_2$  are shown  
 922 evaluated at the mode of the distribution. The length of the arrows is inversely proportional to the  
 923 square root of absolute value of their eigenvalues  $\lambda_1 = -10.8$  and  $\lambda_2 = -2.27$ . We quantitatively  
 924 measured the sensitivity of the model with respect to network syncing along the eigenvectors of the  
 925 Hessian (Fig. 1B, inset). Sensitivity was measured as the slope coefficient of linear regression fit  
 926 to network syncing error (the sum of squared differences of each neuron's frequency from 0.53Hz)  
 927 as a function of parametric perturbation magnitude (maximum 0.25) away from the mode along  
 928 both orientations indicated by the eigenvector with 100 equally spaced samples. The sensitivity  
 929 slope coefficient of eigenvector  $v_1$  with respect to network syncing was significant ( $\beta = 4.82 \times 10^{-2}$ ,  
 930  $p < 10^{-4}$ ). In contrast, eigenvector  $v_2$  did not identify a dimension of parameter space significantly  
 931 sensitive to network syncing ( $\beta = 8.65 \times 10^{-4}$  with  $p = .67$ ). These sensitivities were compared to  
 932 all other dimensions of parameter space (100 equally spaced angles from 0 to  $\pi$ ), revealing that the  
 933 Hessian eigenvectors indeed identified the directions of greatest sensitivity and degeneracy (Fig.  
 934 1B, inset). The contours of Figure 1 were calculated as error in  $T(x)$  from  $\mu$  in both the first and  
 935 second moment emergent property statistics.

### 936 5.2.2 Primary visual cortex

937 The dynamics of each neural populations average rate  $x = [x_E, x_P, x_S, x_V]^\top$  are given by:

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n. \quad (57)$$

938 By consolidating information from many experimental datasets, Billeh et al. [46] produce estimates

<sup>939</sup> of the synaptic strength (in mV)

$$M = \begin{bmatrix} 0.36 & 0.48 & 0.31 & 0.28 \\ 1.49 & 0.68 & 0.50 & 0.18 \\ 0.86 & 0.42 & 0.15 & 0.32 \\ 1.31 & 0.41 & 0.52 & 0.37 \end{bmatrix} \quad (58)$$

<sup>940</sup> and connection probability

$$C = \begin{bmatrix} 0.16 & 0.411 & 0.424 & 0.087 \\ 0.395 & .451 & 0.857 & 0.02 \\ 0.182 & 0.03 & 0.082 & 0.625 \\ 0.105 & 0.22 & 0.77 & 0.028 \end{bmatrix}. \quad (59)$$

<sup>941</sup> Multiplying these connection probabilities and synaptic efficacies gives us an effective connectivity

<sup>942</sup> matrix:

$$W_{\text{full}} = C \odot M = \begin{bmatrix} 0.16 & 0.411 & 0.424 & 0.087 \\ 0.395 & .451 & 0.857 & 0.02 \\ 0.182 & 0.03 & 0.082 & 0.625 \\ 0.105 & 0.22 & 0.77 & 0.028 \end{bmatrix}. \quad (60)$$

<sup>943</sup> Theoretical work on these systems considers a subset of the effective connectivities [28, 43, 44]

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & 0 \\ W_{PE} & W_{PP} & W_{PS} & 0 \\ W_{SE} & 0 & 0 & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & 0 \end{bmatrix}. \quad (61)$$

<sup>944</sup> In coherence with this work, we only keep the entries of  $W_{\text{full}}$  corresponding to parameters in  
<sup>945</sup> Equation 61.

<sup>946</sup> We look at how this four-dimensional nonlinear dynamical model of V1 responds to different inputs,  
<sup>947</sup> and compare the predictions of the linear response to the approximate posteriors obtained through  
<sup>948</sup> EPI. The input to the system is the sum of a baseline input  $b = [1, 1, 1, 1]^\top$  and a differential input  
<sup>949</sup>  $dh$ :

$$h = b + dh. \quad (62)$$

<sup>950</sup> All simulations of this system had  $T = 100$  time points, a time step  $dt = 5\text{ms}$ , and time constant  
<sup>951</sup>  $\tau = 20\text{ms}$ . The system was initialized to a random draw  $x(0)_i \sim \mathcal{N}(1, 0.01)$ .

952 We can describe the dynamics of this system more generally by

$$\dot{x}_i = -x_i + f(u_i) \quad (63)$$

953 where the input to each neuron is

$$u_i = \sum_j W_{ij} x_j + h_i. \quad (64)$$

954 Let  $F_{ij} = \gamma_i \delta(i, j)$ , where  $\gamma_i = f'(u_i)$ . Then, the linear response is

$$\frac{dx_{ss}}{dh} = F(W \frac{dx_{ss}}{dh} + I) \quad (65)$$

955 which is calculable by

$$\frac{dx_{ss}}{dh} = (F^{-1} - W)^{-1}. \quad (66)$$

956 This calculation is used to produce the magenta lines in Figure 2C, which show the linearly predicted  
957 inputs that generate a response from two standard deviations (of  $\mathcal{B}$ ) below and above  $y$ .

958 The emergent property we considered was the first and second moments of the change in steady  
959 state rate  $dx_{ss}$  between the baseline input  $h = b$  and  $h = b + dh$ . We use the following notation to  
960 indicate that the emergent property statistics were set to the following values:

$$\mathcal{B}(\alpha, y) \triangleq \mathbb{E} \begin{bmatrix} dx_{\alpha,ss} \\ (dx_{\alpha,ss} - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix}. \quad (67)$$

961 In the final analysis for this model, we sweep the input one neuron at a time away from the mode  
962 of each inferred distributions  $dh^* = z^* = \text{argmax}_z \log q_\theta(z | \mathcal{B}(\alpha, 0.1))$ . The differential responses  
963  $\delta x_{\alpha,ss}$  are examined at perturbed inputs  $h = b + dh^* + \delta h_\alpha \hat{u}_\alpha$  where  $\hat{u}_\alpha$  is a unit vector in the  
964 dimension of  $\alpha$  and  $\delta x$  is evaluated at 101 equally spaced samples of  $\delta h_\alpha$  from -15 to 15.

965 We measured the linear regression slope between neuron-types of  $\delta x$  and  $\delta h$  to confirm the hy-  
966 potheses H1-H3 (H4 is simply observing the nonmonotonicity) and report the p values for tests of  
967 non-zero slope.

968 H1: the neuron-type responses are sensitive to their direct inputs. E-population:  $\beta = 1.62$ ,  
969  $p < 10^{-4}$  (Fig. 3A black), P-population:  $\beta = 1.06$ ,  $p < 10^{-4}$  (Fig. 3B blue), S-population:  
970  $\beta = 6.80$ ,  $p < 10^{-4}$  (Fig. 3C red), V-population:  $\beta = 6.41$ ,  $p < 10^{-4}$  (Fig. 3D green).

971 H2: the E-population ( $\beta = 0$ ,  $p = 1$ ) and P-populations ( $\beta = 0$ ,  $p = 1$ ) are not affected by  
972  $\delta h_V$  (Fig. 3A green, 3B green);

973 H3: the S-population is not affected by  $\delta h_P$  ( $\beta = 0$ ,  $p = 1$ ) (Fig. 3C blue);

974

975 For each  $\mathcal{B}(\alpha, y)$  with  $\alpha \in \{E, P, S, V\}$  and  $y \in \{0.1, 0.5\}$ , we ran EPI using a real NVP architecture  
 976 of four masks layers with two hidden layers of 10 units, mapped to a support of  $z_i \in [-5, 5]$  with  
 977 no batch normalization. We used an augmented Lagrangian coefficient of  $c_0 = 10^5$ , a batch size  
 978  $n = 1000$ , set  $\nu = 0.5$ . The EPI distributions shown in Fig. 2 are the converged distributions with  
 979 maximum entropy across random seeds.

980 We set the parameters of the Gaussian initialization  $\mu_{\text{init}}$  and  $\Sigma_{\text{init}}$  to the mean and covariance of  
 981 random samples  $z^{(i)} \sim \mathcal{U}(-5, 5)$  that produced emergent property statistic  $dx_{\alpha,ss}$  within a bound  
 982  $\epsilon$  of the emergent property value  $y$ .  $\epsilon = 0.01$  was set to be one standard deviation of the emergent  
 983 property value according to the emergent property value  $0.01^2$  of the variance emergent property  
 984 statistic.

### 985 5.2.3 Superior colliculus

986 In the model of Duan et al [29], there are four total units: two in each hemisphere corresponding to  
 987 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right  
 988 Pro (RP) and right Anti (RA). Each unit has an activity ( $x_\alpha$ ) and internal variable ( $u_\alpha$ ) related  
 989 by

$$x_\alpha = \left( \frac{1}{2} \tanh \left( \frac{u_\alpha - \epsilon}{\zeta} \right) + \frac{1}{2} \right) \quad (68)$$

990 where  $\alpha \in \{LP, LA, RA, RP\}$   $\epsilon = 0.05$  and  $\zeta = 0.5$  control the position and shape of the nonlin-  
 991 earity, respectively.

992 We order the elements of  $x$  and  $u$  in the following manner

$$x = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad u = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}. \quad (69)$$

993 The internal variables follow dynamics:

$$\tau \frac{du}{dt} = -u + Wx + h + \sigma dB \quad (70)$$

994 with time constant  $\tau = 0.09s$  and Gaussian noise  $\sigma dB$  controlled by the magnitude of  $\sigma = 1.0$ . The  
 995 weight matrix has 8 parameters  $sW_P$ ,  $sW_A$ ,  $vW_{PA}$ ,  $vW_{AP}$ ,  $hW_P$ ,  $hW_A$ ,  $dW_{PA}$ , and  $dW_{AP}$  (Fig.

996 4B):

$$W = \begin{bmatrix} sW_P & vW_{PA} & hW_P & dW_{PA} \\ vW_{AP} & sW_A & dW_{AP} & hW_A \\ hW_P & dW_{PA} & sW_P & vW_{PA} \\ dW_{AP} & hW_A & vW_{AP} & sW_A \end{bmatrix}. \quad (71)$$

997 The system receives five inputs throughout each trial, which has a total length of 1.8s.

$$h = h_{\text{rule}} + h_{\text{choice-period}} + h_{\text{light}}. \quad (72)$$

998 There are rule-based inputs depending on the condition,

$$h_{P,\text{rule}}(t) = \begin{cases} I_{P,\text{rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (73)$$

999

$$h_{A,\text{rule}}(t) = \begin{cases} I_{A,\text{rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (74)$$

1000 a choice-period input,

$$h_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (75)$$

1001 and an input to the right or left-side depending on where the light stimulus is delivered.

$$h_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } t > 1.2s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } t > 1.2s \text{ and Right} \\ 0, & t \leq 1.2s \end{cases}. \quad (76)$$

1002 The input parameterization was fixed to  $I_{P,\text{rule}} = 10$ ,  $I_{A,\text{rule}} = 10$ ,  $I_{\text{choice}} = 2$ , and  $I_{\text{light}} = 1$ .

1003 To produce an accuracy rate of  $p_{LP}$  in the Left, Pro condition, let  $\hat{p}_i$  be the empirical average

1004 steady state response (final  $x_{LP}$  at end of task) over M=500 Gaussian noise draws for a given SC

1005 model parameterization  $z_i$ :

$$\hat{p}_i = \mathbb{E}_{\sigma dB} [x_{LP} | s = L, c = P, z = z_i] = \frac{1}{M} \sum_{j=1}^M x_{LP}(s = L, c = P, z = z_i, \sigma dB_j) \quad (77)$$

1006 where stimulus  $s \in \{L, R\}$ , cue  $c \in \{P, A\}$ , and  $\sigma dB_j$  is the Gaussian noise on trial  $j$ . As with the

1007 V1 model, we only consider steady state responses of  $x$ , so  $x_\alpha$  is used from here on to denote the

1008 steady state activity at the end of the trial. For the first emergent property statistic, the average  
 1009 over EPI samples (from  $q_\theta(z)$ ) is set to the desired value  $p_{LP}$ :

$$\mathbb{E}_{z_i \sim q_\phi} [\mathbb{E}_{\sigma dB} [x_{LP,ss} \mid s = L, c = P, z = z_i]] = \mathbb{E}_{z_i \sim q_\phi} [\hat{p}_i] = p_{LP}. \quad (78)$$

1010 For the next emergent property statistic, we ask that the variance of the steady state responses  
 1011 across Gaussian draws, is the Bernoulli variance for the empirical rate  $\hat{p}_i$ :

$$\mathbb{E}_{z \sim q_\phi} [\sigma_{err}^2] = 0 \quad (79)$$

1012 where the Bernoulli variance error  $\sigma_{err}^2$  for the Pro task, left condition is

$$\sigma_{err}^2 = Var_{\sigma dB} [x_{LP} \mid s = L, c = P, z = z_i] - \hat{p}_i(1 - \hat{p}_i). \quad (80)$$

1013 We have an additional constraint that the Pro neuron on the opposite hemisphere should have the  
 1014 opposite value (0 and 1). We can enforce this with another constraint:

$$\mathbb{E}_{z \sim q_\phi} [d_P] = 1, \quad (81)$$

1015 where the distance between Pro neuron steady states  $d_P$  in the Pro condition is

$$d_P = \mathbb{E}_{\sigma dB} [(x_{LP} - x_{RP})^2 \mid s = L, c = P, z = z_i] \quad (82)$$

1016 The emergent property statistics only need to be measured during the Left stimulus condition of  
 1017 the Pro and Anti tasks, since the network is symmetrically parameterized. In total, the emergent  
 1018 property of rapid task switching at accuracy level  $p$  was defined as

$$\mathcal{B}(p) \triangleq \mathbb{E} \begin{bmatrix} \hat{p}_P \\ \hat{p}_A \\ (\hat{p}_P - p)^2 \\ (\hat{p}_A - p)^2 \\ \sigma_{P,err}^2 \\ \sigma_{A,err}^2 \\ d_P \\ d_A \end{bmatrix} = \begin{bmatrix} p \\ p \\ 0.15^2 \\ 0.15^2 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}. \quad (83)$$

1019 Since the maximum variance of a random variable bounded from 0 to 1 is the Bernoulli variance  
 1020  $\hat{p}(1 - \hat{p})$ , and the maximum squared difference between two variables bounded from 0 to 1 is 1, we  
 1021 do not need to control the second moment of these test statistics. These variables are dynamical

1022 system states and can only exponentially decay (or saturate) to 0 (or 1), so the Bernoulli variance  
 1023 error and squared difference constraints cannot be satisfied exactly in simulation. This is important  
 1024 to be mindful of when evaluating the convergence criteria. Instead of using our usual hypothesis  
 1025 testing criteria for convergence to the emergent property, we set a slack variable threshold only for  
 1026 these technically infeasible emergent property values to 0.05.

1027 Using EPI to learn distributions of dynamical systems producing Bernoulli responses at a given rate  
 1028 (with small variance around that rate) was more challenging than expected. There is a pathology in  
 1029 this optimization setup, where the learned distribution of weights is bimodal attributing a fraction  
 1030  $p$  of the samples to an expansive mode (which always sends  $x_{LP}$  to 1), and a fraction  $1 - p$  to a  
 1031 decaying mode (which always sends  $x_{LP}$  to 0). This pathology was avoided using an inequality  
 1032 constraint prohibiting parameter samples that resulted in low variance of responses across noise.

$\lambda$	$\hat{p}$	$q_\theta(z)$	$r$	p-value
$\lambda_{\text{task}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(60\%))$	$1.24 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{task}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(70\%))$	$7.56 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{task}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(80\%))$	$4.59 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{task}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(90\%))$	$3.76 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{task}}$	$\hat{p}_A$	$q(z \mid \mathcal{B}(60\%))$	$4.80 \times 10^{-02}$	$p < .01$
$\lambda_{\text{task}}$	$\hat{p}_A$	$q(z \mid \mathcal{B}(70\%))$	$2.08 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{task}}$	$\hat{p}_A$	$q(z \mid \mathcal{B}(80\%))$	$4.84 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{task}}$	$\hat{p}_A$	$q(z \mid \mathcal{B}(90\%))$	$4.25 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(50\%))$	$-7.57 \times 10^{-02}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(60\%))$	$-6.73 \times 10^{-02}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(70\%))$	$-4.86 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(80\%))$	$-1.43 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_P$	$q(z \mid \mathcal{B}(90\%))$	$-1.93 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_A$	$q(z \mid \mathcal{B}(60\%))$	$-7.60 \times 10^{-02}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_A$	$q(z \mid \mathcal{B}(70\%))$	$-2.73 \times 10^{-01}$	$p < 10^{-4}$
$\lambda_{\text{side}}$	$\hat{p}_A$	$q(z \mid \mathcal{B}(80\%))$	$-2.74 \times 10^{-01}$	$p < 10^{-4}$

Table 1: Table of significant correlation values from Fig. 4E.

1033 For each accuracy level  $p$ , we ran EPI for 10 different random seeds using an architecture of 10  
 1034 planar flows with a support of  $z \in \mathbb{R}^8$ . We used an augmented Lagrangian coefficient of  $c_0 = 10^2$ , a

batch size  $n = 300$ , and set  $\nu = 0.5$ , and initialized  $q_\theta(z)$  to produce an isotropic Gaussian of zero mean with standard deviation  $\sigma_{\text{init}} = 1$ . The EPI distributions shown in Fig. 4 are the converged distributions with maximum entropy across random seeds.

We report significant correlations  $r$  and their p-values from Figure 4E in Table 1. Correlations were measured from 5,000 samples of  $q_\theta(z | \mathcal{B}(p))$  and p-values are reported for one-tailed tests, since we hypothesized a positive correlation between task accuracies  $p_P$  or  $p_A$  and  $\lambda_{\text{task}}$ , and a negative correlation between task accuracies  $p_P$  and  $p_A$  and  $\lambda_{\text{side}}$ .

#### 5.2.4 Rank-1 RNN

Extensive research on random fully-connected recurrent neural networks has resulted in foundational theories of their activity [3, 77]. Furthermore, independent research on training these models to perform computations suggests that learning occurs through low-rank perturbations to the connectivity (e.g. [78, 79]). Recent theoretical work extends theory for random neural networks [3] to those with added low-rank structure [30]. In Section 3.5, we used this theory to enable EPI on RNN parameters conditioned on the emergent property of task execution.

Such RNNs have the following dynamics:

$$\frac{dx}{dt} = -x + W\phi(x) + h, \quad (84)$$

where  $x$  is network activity,  $W$  is the connectivity weight matrix,  $\phi(\cdot) = \tanh(\cdot)$  is the input-output function, and  $h$  is the input to the system. In a rank-1 RNN (which was sufficiently complex for the Gaussian posterior conditioning task),  $W$  is the sum of a random component with strength  $g$  and a structured component determined by the outer product of vectors  $m$  and  $n$ :

$$W = g\chi + \frac{1}{N}mn^\top, \quad (85)$$

where  $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ , and the entries of  $m$  and  $n$  are distributed as  $m_i \sim \mathcal{N}(M_m, 1)$  and  $n_i \sim \mathcal{N}(M_n, 1)$ . For EPI, we consider  $z = [g, M_m, M_n]$ , which are the parameters governing the connectivity properties of the RNN.

From such a parameterization  $z$ , the theory of Mastrogiovisepp et al. produces solutions for variables describing the low dimensional response properties of the RNN. These “dynamic mean field” (DMF) variables (e.g. the activity along a vector  $\kappa_v$ , the total variance  $\Delta_0$ , structured variance  $\Delta_\infty$ , and the chaotic variance  $\Delta_T$ ) are derived to be functions of one another and connectivity parameters  $z$ . The collection of these derived functions results in a system of equations, whose solution must

1062 be obtained through a nonlinear system of equations solver. The iterative steps of this system  
 1063 of equations solver are differentiable, so we take gradients through this solve process. The DMF  
 1064 variables provide task-relevant information about the RNN's response to task inputs.

1065 In the Gaussian posterior conditioning example,  $\kappa_r$  and  $\Delta_T$  are DMF variables used as task-relevant  
 1066 emergent property statistics  $\mu_{\text{post}}$  and  $\sigma_{\text{post}}^2$ . Specifically, we solve for the DMF variables  $\kappa_r$ ,  $\kappa_n$ ,  
 1067  $\Delta_0$  and  $\Delta_\infty$ , where the readout is nominally chosen to point in the unit orthant  $r = [1, \dots, 1]^\top$ . The  
 1068 consistency equations for these variables in the presence of a constant input  $h = yr - (n - M_n)$  can  
 1069 be derived following [30]:

$$\begin{aligned} \kappa_r &= G_1(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = M_m \kappa_n + y \\ \kappa_n &= G_2(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = M_n \langle [\phi_i] \rangle + \langle [\phi'_i] \rangle \\ \frac{\Delta_0^2 - \Delta_\infty^2}{2} &= G_3(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = g^2 \left( \int \mathcal{D}z \Phi^2(\kappa_r + \sqrt{\Delta_0} z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\kappa_r + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right) \\ &\quad + (\kappa_n^2 + 1)(\Delta_0 - \Delta_\infty) \\ \Delta_\infty &= G_4(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = g^2 \int \mathcal{D}z \left[ \int \mathcal{D}x \phi(\kappa_r + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 + \kappa_n^2 + 1 \end{aligned} \quad (86)$$

1070 where here  $z$  is a gaussian integration variable. We can solve these equations by simulating the  
 1071 following Langevin dynamical system to a steady state:

$$\begin{aligned} l(t) &= \frac{\Delta_0(t)^2 - \Delta_\infty(t)^2}{2} \\ \Delta_0(t) &= \sqrt{2l(t) + \Delta_\infty(t)^2} \\ \frac{d\kappa_r(t)}{dt} &= -\kappa_r(t) + G_1(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\ \frac{d\kappa_n(t)}{dt} &= -\kappa_n(t) + G_2(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\ \frac{dl(t)}{dt} &= -l(t) + G_3(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\ \frac{d\Delta_\infty(t)}{dt} &= -\Delta_\infty(t) + G_4(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \end{aligned} \quad (87)$$

1072 Then, the chaotic variance, which is necessary for the Gaussian posterior conditioning example, is  
 1073 simply calculated via  $\Delta_T = \Delta_0 - \Delta_\infty$ .

1074 We ran EPI using a real NVP architecture of two masks and two layers per mask with 10 units  
 1075 mapped to a support of  $z = [g, M_m, M_n] \in [0, 5] \times [-5, 5] \times [-5, 5]$  with no batch normalization.  
 1076 We used an augmented Lagrangian coefficient of  $c_0 = 1$ , a batch size  $n = 300$ , set  $\nu = 0.15$ ,  
 1077 and initialized  $q_\theta(z)$  to produce an isotropic Gaussian with mean  $\mu_{\text{init}} = [2.5, 0, 0]$  with standard

1078 deviation  $\sigma_{\text{init}} = 2.0$ . The EPI distribution shown in Fig. 5 is the converged distributions with  
1079 maximum entropy across five random seeds.

1080 To examine the effect of product  $M_m M_n$  on the posterior mean,  $\mu_{\text{post}}$  we took perturbations in  
1081  $M_m M_n$  away from two representative parameters  $z_1$  and  $z_2$  in 21 equally space increments from  
1082 -1 to 1. For each perturbation, we sampled 10 2,000-neuron RNNs and measure the calculated  
1083 posterior means. In Fig. 5D, we plot the product of  $M_m M_n$  in the perturbation versus the average  
1084 posterior mean across 10 network realizations with standard error bars. The correlation between  
1085 perturbation product  $M_m M_n$  and  $\mu_{\text{post}}$  was measured over all simulations. For perturbations away  
1086 from  $z_1$  the correlation was 0.995 with  $p < 10^{-4}$ , and for perturbations away from  $z_2$  the correlation  
1087 was 0.983 with  $p < 10^{-4}$ .

1088 In addition to the Gaussian posterior conditioning example in Section 3.5, we modeled two tasks  
1089 from Mastrogiuseppe et al.: noisy detection and context-dependent discrimination. We used the  
1090 same theoretical equations and task setups described in their study.

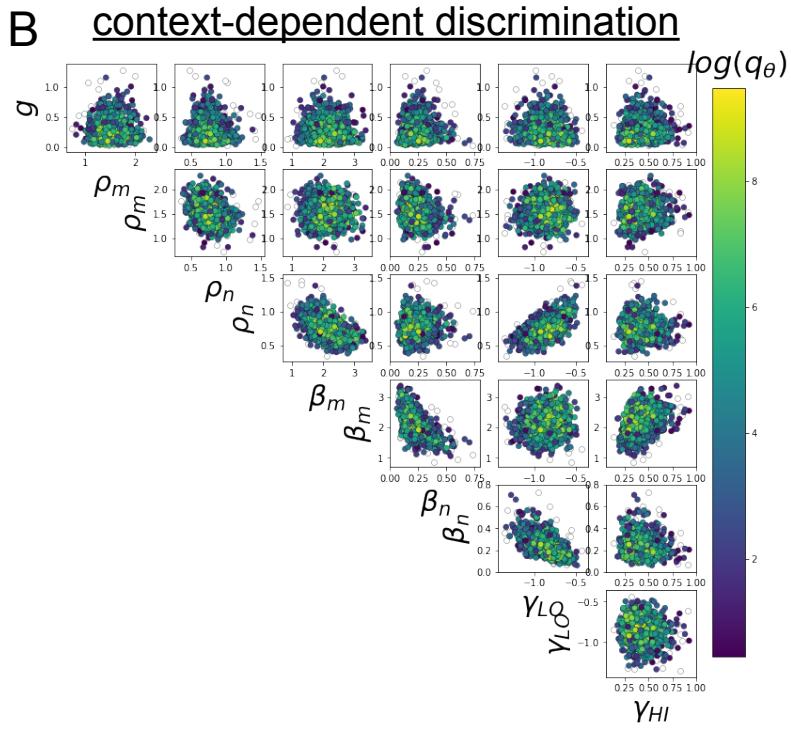
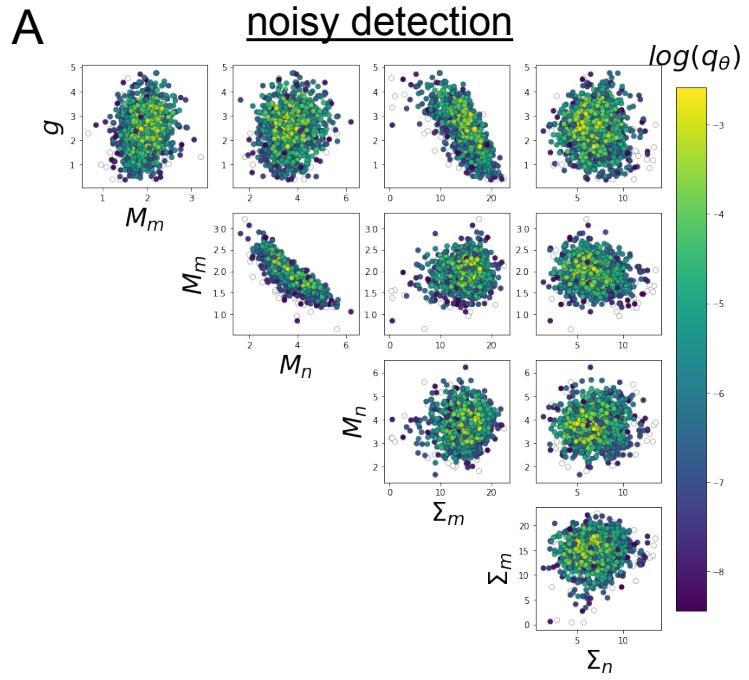


Fig. S5: A. EPI for rank-1 networks doing noisy discrimination. B. EPI for rank-2 networks doing context-dependent discrimination. See [30] for theoretical equations and task description.