

# Interrogating theoretical models of neural computation with deep inference

Sean R. Bittner, Agostina Palmigiano, Alex T. Piet, Chunyu A. Duan, Carlos D. Brody,  
Kenneth D. Miller, and John P. Cunningham.

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

<sup>2</sup> The 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 experimen-  
<sup>4</sup> tally observed phenomenon – whether behavioral or in terms of neural activity – and thus can offer  
<sup>5</sup> insights into neural computation. The operation of these circuits, like all models, critically depends  
<sup>6</sup> on the choices of model parameters. Historically, the gold standard has been to analytically derive  
<sup>7</sup> the relationship between model parameters and computational properties. However, this enterprise  
<sup>8</sup> quickly becomes infeasible as biologically realistic constraints are included into the model increas-  
<sup>9</sup> ing its complexity, often resulting in *ad hoc* approaches to understanding the relationship between  
<sup>10</sup> model and computation. We bring recent machine learning techniques – the use of deep generative  
<sup>11</sup> models for probabilistic inference – to bear on this problem, learning distributions of parameters  
<sup>12</sup> that produce the specified properties of computation. Importantly, the techniques we introduce  
<sup>13</sup> offer a principled means to understand the implications of model parameter choices on compu-  
<sup>14</sup> tational properties of interest. We motivate this methodology with a worked example analyzing  
<sup>15</sup> sensitivity in the stomatogastric ganglion. We then use it to generate insights into neuron-type  
<sup>16</sup> input-responsivity in a model of primary visual cortex, a new understanding of rapid task switch-  
<sup>17</sup> ing in superior colliculus models, and attribution of bias in recurrent neural networks solving a toy  
<sup>18</sup> mathematical problem. More generally, this work offers a quantitative grounding for theoretical  
<sup>19</sup> models going forward, pointing a way to how rigorous statistical inference can enhance theoretical  
<sup>20</sup> neuroscience at large.

## <sup>21</sup> 2 Introduction

<sup>22</sup> The fundamental practice of theoretical neuroscience is to use a mathematical model to understand  
<sup>23</sup> neural computation, whether that computation enables perception, action, or some intermediate  
<sup>24</sup> processing [1]. In this field, a neural computation is systematized with a set of equations – the  
<sup>25</sup> model – and these equations are motivated by biophysics, neurophysiology, and other conceptual  
<sup>26</sup> considerations. The function of this system is governed by the choice of model parameters, which

27 when configured appropriately, give rise to a measurable signature of a computation. The work of  
28 analyzing a model then requires solving the inverse problem: given a computation of interest, how  
29 can we reason about these suitable parameter configurations? The inverse problem is crucial for  
30 reasoning about likely parameter values, uniquenesses and degeneracies, attractor states and phase  
31 transitions, and predictions made by the model.

32 Consider the idealized practice: one carefully designs a model and analytically derives how model  
33 parameters govern the computation. Seminal examples of this gold standard include our field’s un-  
34 derstanding of memory capacity in associative neural networks [2] and chaos and autocorrelation  
35 timescales in random neural networks [3] (adopting approaches from physics), and the paradoxical  
36 effect [4] and decision making [5] in rate models. Unfortunately, as circuit models include more  
37 biological realism, theory via analytic derivation becomes intractable. This creates an unfavorable  
38 tradeoff. On the one hand, one may tractably analyze systems of equations with unrealistic assump-  
39 tions (for example symmetry or gaussianity), producing accurate inferences about parameters of a  
40 too-simple model. On the other hand, one may choose a more biologically accurate, scientifically  
41 relevant model at the cost of *ad hoc* approaches to analysis (simply examining simulated activity),  
42 potentially resulting in bad inferences and thus erroneous scientific predictions and conclusions.

43 Of course, this same tradeoff has been confronted in many scientific fields and engineering problems  
44 characterized by the need to do inference in complex models. In response, the machine learning  
45 community has made remarkable progress in recent years, via the use of deep neural networks as  
46 a powerful inference engine: a flexible function family that can map observed phenomena (in this  
47 case the measurable signal of some computation) back to probability distributions quantifying the  
48 likely parameter configurations. One celebrated example of this approach from machine learning, of  
49 which we draw key inspiration for this work, is the variational autoencoder [6, 7], which uses a deep  
50 neural network to induce an (approximate) posterior distribution on hidden variables in a latent  
51 variable model, given data. Indeed, these tools have been used to great success in neuroscience as  
52 well, in particular for interrogating parameters (sometimes treated as hidden states) in models of  
53 both cortical population activity [8, 9, 10, 11] and animal behavior [12, 13, 14]. These works have  
54 used deep neural networks to expand the expressivity and accuracy of statistical models of neural  
55 data [15].

56 However, these inference tools have not significantly influenced the study of theoretical neuroscience  
57 models, for at least three reasons. First, at a practical level, the nonlinearities and dynamics of  
58 many theoretical models are such that conventional inference tools typically produce a narrow set of

59 insights into these models. Indeed, only in the last few years has deep learning research advanced to  
60 a point of relevance to this class of problem. Second, the object of interest from a theoretical model  
61 is not typically data itself, but rather a qualitative phenomenon – inspection of model behavior, or  
62 better, a measurable signature of some computation – an *emergent property* of the model. Third,  
63 because carefully constructed biological models do not fit cleanly into the framing of a statistical  
64 model. Technically, because many such models stipulate a noisy system of differential equations  
65 that can only be sampled or realized through forward simulation, they lack the explicit likelihood  
66 and priors central to the probabilistic modeling toolkit.

67 To address these three challenges, we developed an inference methodology – ‘emergent property  
68 inference’ – which learns a distribution over parameter configurations in a theoretical model. This  
69 distribution has two critical properties: *(i)* it is chosen such that draws from the distribution (pa-  
70 rameter configurations) correspond to systems of equations that give rise to a specified emergent  
71 property (a set of constraints); and *(ii)* it is chosen to have maximum entropy given those con-  
72 straints, such that we identify all likely parameters and can use the distribution to reason about  
73 parametric sensitivity and degeneracies [16]. First, we stipulate a bijective deep neural network that  
74 induces a flexible family of probability distributions over model parameterizations with a probabil-  
75 ity density we can calculate [17, 18, 19]. Second, we quantify the notion of emergent properties as a  
76 set of moment constraints on datasets generated by the model. Thus, an emergent property is not a  
77 single data realization, but a phenomenon or a feature of the model, which is ultimately the object  
78 of interest in theoretical neuroscience. Conditioning on an emergent property requires a variant of  
79 deep probabilistic inference methods, which we have previously introduced [20]. Third, because we  
80 cannot assume the theoretical model has explicit likelihood on data or the emergent property of  
81 interest, we use stochastic gradient techniques in the spirit of likelihood free variational inference  
82 [21]. Taken together, emergent property inference (EPI) provides a methodology for inferring pa-  
83 rameter configurations consistent with a particular emergent phenomena in theoretical models. We  
84 use a classic example of parametric degeneracy in a biological system, the stomatogastric ganglion  
85 [22], to motivate and clarify the technical details of EPI.

86 Equipped with this methodology, we then investigated three models of current importance in the-  
87 oretical neuroscience. These models were chosen to demonstrate generality through ranges of bi-  
88 ological realism (from conductance-based biophysics to recurrent neural networks), neural system  
89 function (from pattern generation to abstract cognitive function), and network scale (from four to  
90 infinite neurons). First, we use EPI to produce a set of verifiable hypotheses of input-responsivity

in a four neuron-type dynamical model of primary visual cortex; we then validate these hypotheses in the model. Second, we demonstrated how the systematic application of EPI to levels of task performance can generate experimentally testable hypotheses regarding connectivity in superior colliculus. Third, we use EPI to uncover the sources of bias in a low-rank recurrent neural network executing a toy mathematical computation. The novel scientific insights offered by EPI contextualize and clarify the previous studies exploring these models [23, 24, 25, 26] and more generally, suggests a departure from realism vs tractability considerations towards the use of modern machine learning for sophisticated interrogation of biologically relevant models.

We note that, during our preparation and early presentation of this work [27, 28], another work has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural circuits [29]. We are excited by this broad problem being recognized by the community, and we emphasize that these works offer complementary neuroscientific contributions and use different technical methodologies. While we have advanced our research on deep generative modeling [20] to a point of significant relevance to statistical inference in theoretical neuroscience, they have also furthered their research on approximate Bayesian inference in such models [30]. The existence of these complementary methodologies emphasizes the increased importance and timeliness of both works.

## 3 Results

### 3.1 Motivating emergent property inference of theoretical models

Consideration of the typical workflow of theoretical modeling clarifies the need for emergent property inference. First, one designs or chooses an existing model that, it is hypothesized, captures the computation of interest. To ground this process in a well-known example, consider the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic muscle activation patterns for digestion [31]. Despite full knowledge of STG connectivity and a precise characterization of its rhythmic pattern generation, biophysical models of the STG have complicated relationships between circuit parameters and neural activity [22, 32]. A model of the STG [23] is shown schematically in Figure 1A, and note that the behavior of this model will be critically dependent on its parameterization – the choices of conductance parameters  $z = [g_{el}, g_{synA}]$ . Specifically, the two fast neurons ( $f1$  and  $f2$ ) mutually inhibit one another, and oscillate at a faster frequency than the mutually inhibiting slow neurons ( $s1$  and  $s2$ ), and the hub neuron (hub) couples



Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. A. For a choice of model (STG) and emergent property (network syncing), emergent property inference (EPI, gray box) learns a distribution of the model parameters  $z = [g_{el}, g_{synA}]$  producing network syncing. 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. Network syncing traces are colored by log probability of their generating parameters in the EPI-inferred distribution. B. An EPI distribution of STG model parameters producing network syncing. Samples are colored by log density. Distribution contours of emergent property value error are shown at levels of  $2 \times 10^{-6}$ ,  $2 \times 10^{-5}$ , and  $2 \times 10^{-4}$ . Eigenvectors of the Hessian at the mode of the inferred distribution are indicated as  $v_1$  and  $v_2$ . Simulated activity is shown for three samples (stars). (Inset) Sensitivity of the system with respect to network syncing along all dimensions of parameter space away from the mode. (see Section A.2.1). C. Deep probability distributions map a latent random variable  $w$  through a deep neural network with weights and biases  $\theta$  to parameters  $z = f_\theta(w)$  distributed as  $q_\theta(z)$ . D. EPI optimization: To learn the EPI distribution  $q_\theta(z)$  of model parameters that produce an emergent property, the emergent property statistics  $T(x)$  are set in expectation over model parameter samples  $z \sim q_\theta(z)$  and model simulations  $x \sim p(x | z)$  to emergent property values  $\mu$ . The maximum entropy distribution producing the emergent property.

121 with the fast or slow population or both.

122 Second, once the model is selected, one defines the emergent property, the measurable signal of  
 123 scientific interest. To continue our running STG example, one such emergent property is the  
 124 phenomenon of *network syncing* – in certain parameter regimes, the frequency of the hub neuron  
 125 matches that of the fast and slow populations at an intermediate frequency. This emergent property  
 126 is shown in Figure 1A at a frequency of 0.54Hz.

127 Third, qualitative parameter analysis ensues: since precise mathematical analysis is intractable in  
 128 this model, a brute force sweep of parameters is done [23]. Subsequently, a qualitative description  
 129 is formulated to describe the different parameter configurations that lead to the emergent property.  
 130 In this last step lies the opportunity for a precise quantification of the emergent property as a  
 131 statistical feature of the model. Once we have such a methodology, we can infer a probability  
 132 distribution over parameter configurations that produce this emergent property.

133 Before presenting technical details (in the following section), let us understand emergent property  
 134 inference schematically: EPI (Fig. 1A gray box) takes, as input, the model and the specified  
 135 emergent property, and as its output, produces the parameter distribution shown in Figure 1B.  
 136 This distribution – represented for clarity as samples from the distribution – is then a scientifically  
 137 meaningful and mathematically tractable object. In the STG model, this distribution can be  
 138 specifically queried to reveal the prototypical parameter configuration for network syncing (the  
 139 mode; Figure 1B yellow star), and how network syncing decays based on changes away from the  
 140 mode. Intuitively, the probability density of the samples is in agreement with the emergent property  
 141 value error (Fig. 1B contours). Furthermore, the eigenvectors of the distribution Hessian at the  
 142 mode can be queried to quantitatively formalize the robustness of network syncing (Fig. 1B  $v_1$  and  
 143  $v_2$ ). Indeed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity  
 144 ( $v_1$ ) and degeneracy ( $v_2$ ) have diminished or preserved network syncing, respectively (Figure 1B  
 145 inset and activity traces). Further validation of EPI is available in the supplementary materials,  
 146 where we analyze a simpler model for which ground-truth statements can be made (Section A.1.1).

### 147 3.2 A deep generative modeling approach to emergent property inference

148 Emergent property inference (EPI) systematizes the three-step procedure of the previous section.  
 149 First, we consider the model as a coupled set of differential (and potentially stochastic) equations  
 150 [23]. In the running STG example, its activity  $x = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$  is the membrane potential

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

$$C_m \frac{dx}{dt} = -h(x; z) = -[h_{leak}(x; z) + h_{Ca}(x; z) + h_K(x; z) + h_{hyp}(x; z) + h_{elec}(x; z) + h_{syn}(x; z)] \quad (1)$$

152 where  $C_m=1\text{nF}$ , and  $h_{leak}$ ,  $h_{Ca}$ ,  $h_K$ ,  $h_{hyp}$ ,  $h_{elec}$ ,  $h_{syn}$  are the leak, calcium, potassium, hyperpolarization, electrical, and synaptic currents, all of which have their own complicated dependence on  $x$  and  $z = [g_{el}, g_{synA}]$  (see Section A.2.1).

155 Second, we define the emergent property, which as above is network syncing: oscillation of the  
 156 entire population at an intermediate frequency of our choosing (Figure 1A bottom). Quantifying  
 157 this phenomenon is straightforward: we define network syncing to be that each neuron’s spiking  
 158 frequency – denoted  $\omega_{f1}(x)$ ,  $\omega_{f2}(x)$ , etc. – is close to an intermediate frequency of 0.54Hz. Mathematically,  
 159 we achieve this via constraints on the mean and variance of  $\omega_i(x)$  for each neuron  
 160  $i \in \{f1, f2, \text{hub}, s1, s2\}$ , and thus:

$$\mathbb{E}[T(x)] \triangleq \mathbb{E} \begin{bmatrix} \omega_{f1}(x) \\ \vdots \\ (\omega_{f1}(x) - 0.54)^2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0.54 \\ \vdots \\ 0.025^2 \\ \vdots \end{bmatrix} \triangleq \mu, \quad (2)$$

161 which completes the quantification of the emergent property.

162 Third, we perform emergent property inference: we find a distribution over parameter configura-  
 163 tions  $z$ , and insist that samples from this distribution produce the emergent property; in other  
 164 words, they obey the constraints introduced in Equation 2. This distribution will be chosen from  
 165 a family of probability distributions  $\mathcal{Q} = \{q_\theta(z) : \theta \in \Theta\}$ , defined by a deep generative distribution  
 166 of the normalizing flow class [17, 18, 19] – neural networks which transform a simple distribution  
 167 into a suitably complicated distribution (as is needed here). This deep distribution is represented  
 168 in Figure 1C (and see Methods for more detail). Then, mathematically, we must solve the following  
 169 optimization program:

$$\begin{aligned} & \underset{q_\theta \in \mathcal{Q}}{\operatorname{argmax}} H(q_\theta(z)) \\ & \text{s.t. } \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] = \mu, \end{aligned} \quad (3)$$

170 where  $T(x), \mu$  are defined as in Equation 2, and  $p(x|z)$  is the intractable distribution of data from  
 171 the model ( $x$ ), given that model’s parameters  $z$  (we access samples from this distribution by running  
 172 the model forward). The purpose of each element in this program is detailed in Figure 1D. Finally,

we recognize that many distributions in  $\mathcal{Q}$  will respect the emergent property constraints, so we require a normative principle to select amongst them. This principle is captured in Equation 3 by the primal objective  $H$ . Here we chose Shannon entropy as a means to find parameter distributions with minimal assumptions beyond some chosen structure [33, 34, 20, 35], but we emphasize that the EPI method is unaffected by this choice (but the results of course will depend on the primal objective chosen).

EPI optimizes the weights and biases  $\theta$  of the deep neural network (which induces the probability distribution) by iteratively solving Equation 3. The optimization is complete when the sampled models with parameters  $z \sim q_\theta$  produce activity consistent with the specified emergent property. Such convergence is evaluated with a hypothesis test that the mean of each emergent property statistic is not different than its emergent property value (see Section A.1.2). Equipped with this method, now prove out the value of EPI by using it to investigate and produce novel insights about three prominent models in neuroscience.

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

Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output function have succeeded in explaining a host of experimentally documented phenomena. In a regime characterized by inhibitory stabilization of strong recurrent excitation, these models gives rise to paradoxical responses [4], selective amplification [36], surround suppression [37] and normalization [38]. Despite their strong predictive power, E-I circuit models rely on the assumption that inhibition can be studied as an indivisible unit. Instead, experimental evidence shows that inhibition is composed of distinct elements (parvalbumin (P), somatostatin(S), vip (V)) composing 80% of GABAergic interneurons in V1 [39, 40, 41] and that these inhibitory cell types follow specific connectivity patterns (Fig. 2A) [42]. Recent theoretical advances [24, 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 use EPI to go beyond linear theory by systematically generating and evaluating hypotheses of circuit model function using distributions of parameters 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 ( $\|\cdot\|_+$ ) and exponentiated sum of recurrent ( $Wx$ ) and external  $h$  inputs:



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. 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).

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

- 204 The effective connectivity weights  $W$  were obtained from experimental recordings of publicly avail-  
 205 able datasets of mouse V1 [45, 46] (see Section A.2.2). The input  $h = b + dh$  is comprised of  
 206 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  
 207 neuron-type population. Throughout subsequent analyses, the baseline input is  $b = [1, 1, 1, 1]^\top$ .  
 208 With this model, we are interested in the differential responses of each neuron-type population to  
 209 changes in input  $dh$ . Initially, we studied the linearized response of the system to input  $\frac{dx_{ss}}{dh}$  at the  
 210 steady state response  $x_{ss}$ , i.e. a fixed point. All analyses of this model consider the steady state  
 211 response, so we drop the notation  $ss$  from here on. While this linearization accurately predicts  
 212 differential responses  $dx = [dx_E, dx_P, dx_S, dx_V]$  for small differential inputs to each population  
 213  $dh = [0.1, 0.1, 0.1, 0.1]$  (Fig 2B left), the linearization is a poor predictor in this nonlinear model  
 214 more generally (Fig. 2B right). Currently available approaches to deriving the steady state response  
 215 of the system are limited.  
 216 To get a more comprehensive picture of the input-responsivity of each neuron-type beyond linear  
 217 theory, we used EPI to learn a distribution of the differential inputs to each population  $dh$  that  
 218 produce an increase of  $y \in \{0.1, 0.5\}$  in the rate of each neuron-type population  $\alpha \in \{E, P, S, V\}$ .  
 219 We want to know the differential inputs  $dh$  that result in a differential steady state  $dx_\alpha$  (the change  
 220 in  $x_\alpha$  when receiving input  $h = b + dh$  with respect to the baseline  $h = b$ ) of value  $y$  with some small,  
 221 arbitrarily chosen amount of variance 0.01<sup>2</sup>. 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 prop-  
 223 erty, which represents a different signature of computation in each application. In each column  
 224 of Figure 2C visualizes the inferred distribution, available through EPI, of  $dh$  corresponding to  
 225 an excitatory (red), parvalbumin (blue), somatostatin (red) and vip (green) neuron-type increase,  
 226 while each row corresponds to amounts of increase 0.1 and 0.5. For each pair of parameters we  
 227 show the two-dimensional marginal distribution of samples colored by  $\log q_\theta(dh \mid \mathcal{B}(\alpha, y))$ . The  
 228 inferred distributions immediately suggest four hypotheses:  
 229

- 230 H1: as is intuitive, each neuron-type's firing rate should be sensitive to that neuron-type's

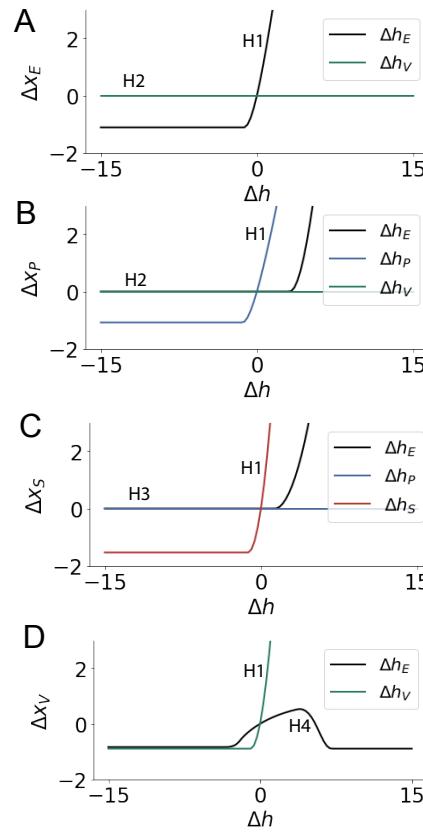


Figure 3: Confirming EPI generated hypotheses in V1. A. Differential responses by the E-population to changes in individual input  $\Delta h_\alpha u_\alpha$  away from the mode of the EPI distribution  $dh^*$ . B-D Same plots for the P-, S-, and V-populations. Labels H1, H2, H3, and H4 indicate which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

231 direct input (e.g. Fig. 2C H1 gray box indicates low variance in  $dh_E$  when  $\alpha = E$ . Same  
 232 observation in all inferred distributions);

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

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

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

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

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

242  $\Delta x_\alpha$  is the change in steady state response to the system with input  $h = b + dh^* + \Delta h_\alpha u_\alpha$  compared  
 243 to  $h = b + dh^*$ , where  $u_\alpha$  is a unit vector in the dimension of  $\alpha$ . The EPI-generated hypotheses are  
 244 confirmed.

H1: the neuron-type responses are sensitive to their direct inputs (Fig. 3A black, 3B blue, 3C red, 3D green);  
H2: the E- and P-populations are not affected by  $dh_V$  (Fig. 3A green, 3B green);  
H3: the S-population is not affected by  $dh_P$  (Fig. 3C blue);  
H4: the V-population exhibits a nonmonotonic response to  $dh_E$  (Fig. 3D black), and is in fact the on population to do so (Fig. 3A-C black).

These hypotheses were in stark contrast to what was available to us via traditional analytical linear prediction (Fig. 2C, magenta). To this point, we have shown the utility of EPI on relatively low-level emergent properties like network syncing and differential neuron-type population responses. In the remainder of the study, we focus on using EPI to understand models of more abstract cognitive function.

### 3.4 Identifying neural mechanisms of behavioral learning.

Identifying measurable biological changes that result in improved behavior is important for neuroscience, since they may indicate how the learning brain adapts. In a rapid task switching experiment [47], rats were explicitly cued on each trial to either orient towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A) task (Fig. 4a). Neural recordings in the midbrain superior colliculus (SC) exhibited two populations of neurons that simultaneously represented both task context (Pro or Anti) and motor response (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons [25]. Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is a four-population dynamical system. Here, the neuron-type populations are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right (R)). The Pro- or Anti-populations receive an input determined by the cue, and then the left and right populations receive an input based on the side of the light stimulus. Activities were bounded between 0 and 1, so that a high output of the Pro population in a given hemisphere corresponds to the contralateral response. An additional stipulation is that when one Pro population responds with a high-output, the opposite Pro population must respond with a low output. Finally, this circuit operates in the presence of Gaussian noise resulting in trial-to-trial variability (see Section A.2.3). The connectivity matrix is parameterized by the geometry of the population arrangement (Fig. 4B).

Here, we used EPI to learn distributions of the SC weight matrix parameters  $z = W$  conditioned on various levels of rapid task switching accuracy  $\mathcal{B}(p)$  for  $p \in \{50\%, 60\%, 70\%, 80\%, 90\%\}$  (see

276 Section A.2.3). Following the approach in Duan et al., we decomposed the connectivity matrix  
 277  $W = QAQ^{-1}$  in such a way (the Schur decomposition) that the basis vectors  $q_i$  are the same for all  
 278  $W$  (Fig. 4C). These basis vectors have intuitive roles in processing for this task, and are accordingly  
 279 named the *all* mode - all neurons co-fluctuate, *side* mode - one side dominates the other, *task* mode  
 280 - the Pro or Anti populations dominate the other, and *diag* mode - Pro- and Anti-populations of  
 281 opposite hemispheres dominate the opposite pair. The corresponding eigenvalues (e.g.  $a_{\text{task}}$ , which  
 282 change according to  $W$ ) indicate the degree to which activity along that mode is increased or  
 283 decreased by  $W$ .

284 EPI demonstrates that, for greater task accuracies, the task mode eigenvalue increases, indicating  
 285 the importance of  $W$  to the task representation (Fig. 4D, purple). Stepping from random chance  
 286 (50%) networks to marginally task-performing (60%) networks, there is a marked decrease of the  
 287 side mode eigenvalues (Fig. 4D, orange). Such side mode suppression remains in the models  
 288 achieving greater accuracy, revealing its importance towards task performance. There were no  
 289 interesting trends with learning in the all or diag mode (hence not shown in Fig. 4). Importantly,  
 290 we can conclude from our methodology that side mode suppression in  $W$  allows rapid task switching,  
 291 and that greater task-mode representations in  $W$  increase accuracy. These hypotheses are confirmed  
 292 by forward simulation of the SC model (Fig. 4E). Thus, EPI produces novel, experimentally testable  
 293 predictions: effective connectivity between these populations changes throughout learning, in a way  
 294 that increases its task mode and decreases its side mode eigenvalues.

### 295 3.5 Linking RNN connectivity to computational error

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

$$\frac{dx(t)}{dt} = -x(t) + W\phi(x(t)) + h(t), \quad (7)$$

301 where  $x(t)$  is the network activity,  $W$  is the network connectivity,  $\phi(\cdot) = \tanh(\cdot)$ , and  $h(t)$  is the  
 302 input to the system. Such RNNs are trained to do a task from a systems neuroscience experiment,  
 303 and then the unit activations of the trained RNN are compared to recorded neural activity. Fully-  
 304 connected RNNs with tens of thousands of parameters are challenging to characterize [49], especially  
 305 making statistical inferences about their parameterization. Alternatively, we consider a rank-1,  $N$ -

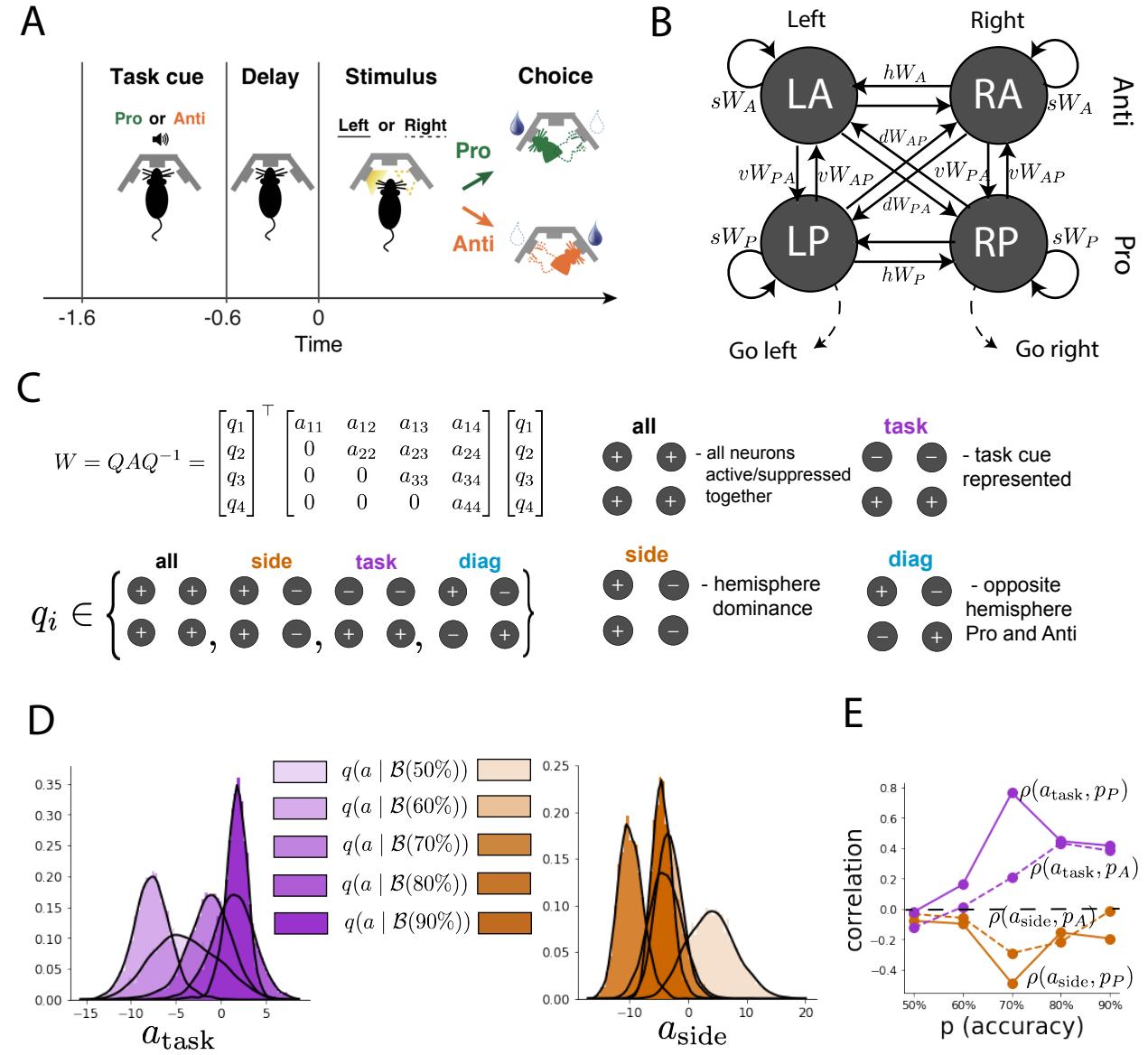


Figure 4: EPI reveals changes in SC [25] connectivity that control task accuracy. A. Rapid task switching behavioral paradigm (see text). B. Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters:  $sW$  - self,  $hW$  - horizontal,  $vW$  - vertical,  $dW$  - diagonal weights. C. The Schur decomposition of the weight matrix  $W = QAQ^{-1}$  is a unique decomposition with orthogonal  $Q$  and upper triangular  $A$ . Schur modes:  $q_{\text{all}}$ ,  $q_{\text{task}}$ ,  $q_{\text{side}}$ , and  $q_{\text{diag}}$ . 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.

306 neuron RNN with connectivity

$$W = g\chi + \frac{1}{N}mn^\top, \quad (8)$$

307 where  $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ ,  $g$  is the random strength, and the entries of  $m$  and  $n$  are drawn from Gaussian  
 308 distributions  $m_i \sim \mathcal{N}(M_m, 1)$  and  $n_i \sim \mathcal{N}(M_n, 1)$ . We use EPI to infer the parameterizations of  
 309 rank-1 RNNs solving an example task, enabling discovery of properties of connectivity that result  
 310 in different types of computational errors.

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

$$\mu_{post} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y}{\sigma_y^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_y^2}} \quad \sigma_{post}^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_y^2}}. \quad (9)$$

315 The RNN is trained to solve this task by producing readout activity that is on average the posterior  
 316 mean  $\mu_{post}$ , and activity whose variability is the posterior variance  $\sigma_{post}^2$  (a setup inspired by  
 317 [50]). To solve this Gaussian posterior conditioning task, the RNN response to a constant input  
 318  $h(t) = yw + (n - M_n)$  must equal the posterior mean along readout vector  $r$ , where

$$\kappa_r = \frac{1}{N} \sum_{j=1}^N r_j \phi(x_j) \quad (10)$$

319 Additionally, the amount of chaotic variance  $\Delta_T$  must equal the posterior variance.  $\kappa_r$  and  $\Delta_T$  can  
 320 be expressed in terms of each other through a solvable system of nonlinear equations (see Section  
 321 A.2.4) [26]. This theory allows us to mathematically formalize the execution of this task into an  
 322 emergent property, where the emergent property statistics of the RNN activity are  $k_r$  and  $\Delta_T$  and  
 323 the emergent property values are the ground truth posterior mean  $\mu_{post}$  and variance  $\sigma_{post}^2$ :

$$E \begin{bmatrix} \kappa_r \\ \Delta_T \\ (\kappa_r - \mu_{post})^2 \\ (\Delta_T^2 - \sigma_{post}^2)^2 \end{bmatrix} = \begin{bmatrix} \mu_{post} \\ \sigma_{post}^2 \\ 0.1 \\ 0.1 \end{bmatrix} \quad (11)$$

324 We specify a substantial amount of variability in the variance constraints so that the inferred  
 325 distribution results in RNNs with a variety biases in their solutions to the gaussian posterior  
 326 conditioning problem.

327 We used EPI to learn distributions of RNN connectivity properties  $z = [g \ M_m \ M_n]$  executing  
 328 Gaussian posterior conditioning given an input of  $y = 2$ . (see Section A.2.4) (Fig. 5B). The true

329 Gaussian conditioning posterior for an input of  $y = 2$  is  $\mu_{\text{post}} = 3$  and  $\sigma_{\text{post}} = 0.5$ . We examined  
 330 the nature of the over- and under-estimation of the posterior means (Fig. 5B, left) and variances  
 331 (Fig. 5B, right) in the inferred distributions. There is rough symmetry in the  $M_m$ - $M_n$  plane,  
 332 suggesting a degeneracy in the product of  $M_m$  and  $M_n$  (Fig. 5B). The product of  $M_m$  and  $M_n$   
 333 almost completely determines the posterior mean (Fig. 5B, left), and the random strength  $g$  is the  
 334 most influential variable on the temporal variance (Fig. 5B, right). Neither of these observations  
 335 were obvious from what mathematical analysis is available in networks of this type (see Section  
 336 A.2.4). They lead to the following hypotheses:

- 337 H1: The posterior mean of the RNN increases with the product of  $M_m$  and  $M_n$ ;  
 338 H2: The posterior variance increases with  $g$ ;

339

340 Testing these now in finite-size networks. Will write end of this later.

341 This novel procedure of doing inference in interpretable parameterizations of RNNs conditioned on  
 342 the emergent property of task execution is straightforwardly generalizable to other tasks like noisy  
 343 integration and context-dependent decision making (Fig. S1).

## 344 4 Discussion

### 345 4.1 EPI is a general tool for theoretical neuroscience

346 Models of biological systems are often comprised of complex nonlinear differential equations, mak-  
 347 ing traditional theoretical analysis and statistical inference intractable. In contrast, EPI is capable  
 348 of learning distributions of parameters in such models producing measurable signatures of compu-  
 349 tation. We have demonstrated its utility on biological models (STG), intermediate-level models of  
 350 interacting genetically- and functionally-defined neuron-types (V1, SC), and the most abstract of  
 351 models (RNNs). We are able to condition both deterministic and stochastic models on low-level  
 352 emergent properties like firing rates of membrane potentials, as well as high-level cognitive func-  
 353 tion like Gaussian posterior conditioning. Technically, EPI is tractable when the emergent property  
 354 statistics are continuously differentiable with respect to the model parameters, which is very often  
 355 the case; this emphasizes the general utility of EPI.

356 In this study, we have focused on applying EPI to low dimensional parameter spaces of models  
 357 with low dimensional dynamical state. These choices were made to present the reader with a series

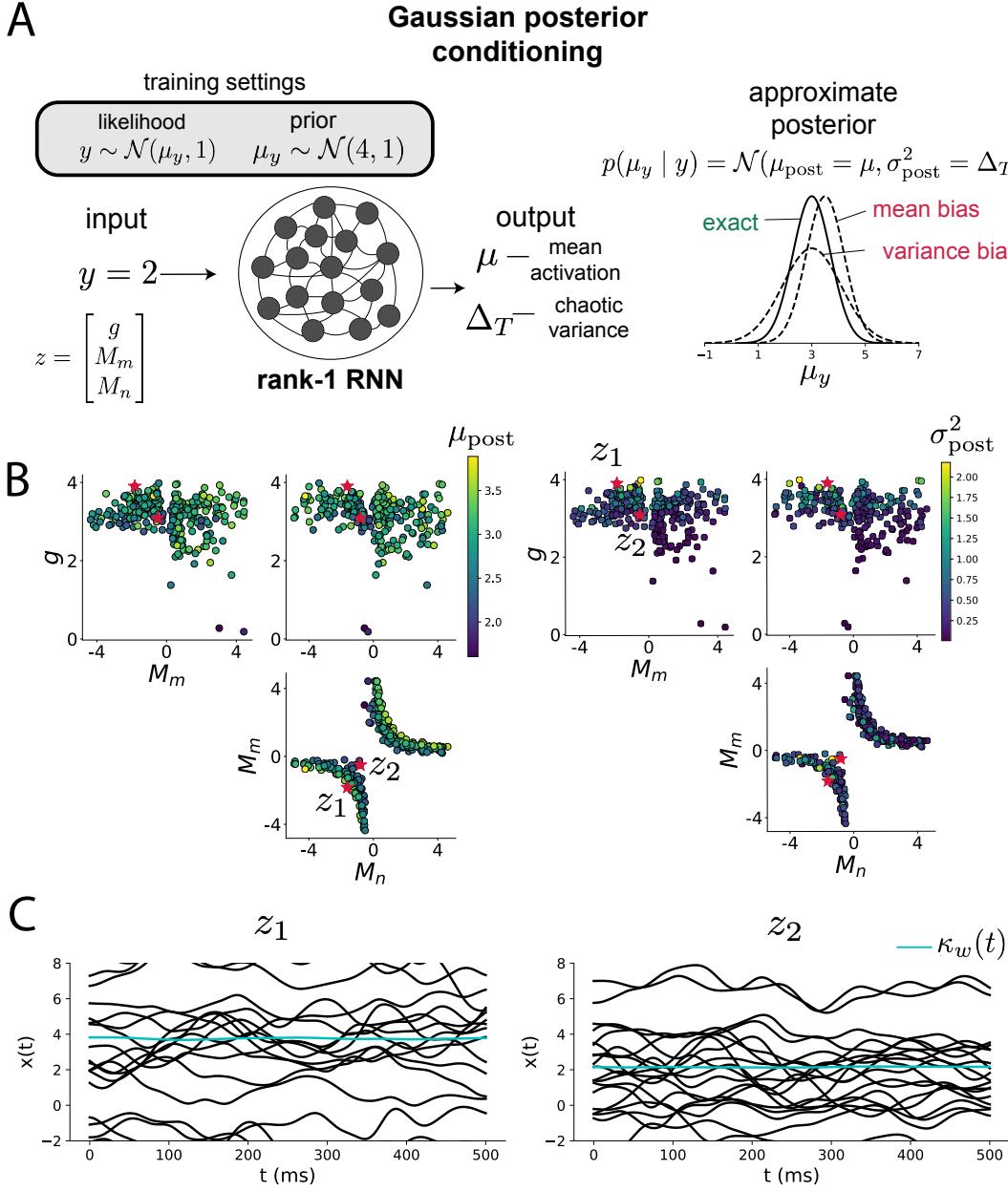


Figure 5: Sources of solution bias in an RNN computation. A. (left) A rank-1 RNN executing a Gaussian posterior conditioning computation on  $\mu_y$ . (right) Bias in this computation can come from over- or under-estimating the posterior mean or variance. B. EPI distribution of rank-1 RNNs executing 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 networks sampled from the distribution perform the calculation and have the computational biases expected from their parameter values. Activity along readout  $\kappa_r$  (cyan).

358 of interpretable conclusions, which is more challenging in high dimensional spaces. In fact, EPI  
 359 should scale reasonably to high dimensional parameter spaces, as the underlying technology has  
 360 produced state-of-the-art performance on high-dimensional tasks such as texture generation [20].  
 361 Of course, increasing the dimensionality of the dynamical state of the model makes optimization  
 362 more expensive, and there is a practical limit there as with any machine learning approach. For  
 363 systems with high dimensional state, we recommend using theoretical approaches (e.g. [26]) to  
 364 reason about reduced parameterizations of such high-dimensional systems.

365 There are additional technical considerations when assessing the suitability of EPI for a particu-  
 366 lar modeling question. First and foremost, as in any optimization problem, the defined emergent  
 367 property should always be appropriately conditioned (constraints should not have wildly different  
 368 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution  
 369 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-  
 370 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,  
 371 one should consider the computational cost of the gradient calculations. In the best circumstance,  
 372 there is a simple, closed form expression (e.g. Section A.1.1) for the emergent property statistic  
 373 given the model parameters. On the other end of the spectrum, many forward simulation iterations  
 374 may be required before a high quality measurement of the emergent property statistic is available  
 375 (e.g. Section A.2.1). In such cases, optimization will be expensive.

## 376 4.2 Novel hypotheses from EPI

377 Machine learning has played an effective, multifaceted role in neuroscientific progress. Primarily,  
 378 it has revealed structure in large-scale neural datasets [51, 52, 53, 54, 55, 56] (see review, [15]).  
 379 Secondarily, trained algorithms of varying degrees of biological relevance are beginning to be viewed  
 380 as fully-observable computational systems comparable to the brain [49, 57].

381 For example, consider the fact that we do not fully understand the four-dimensional models of V1  
 382 [24]. Because analytical approaches to studying nonlinear dynamical systems become increasingly  
 383 complicated when stepping from two-dimensional to three- or four-dimensional systems in the  
 384 absence of restrictive simplifying assumptions [58], it is unsurprising that this model has been a  
 385 challenge. In Section 3.3, we showed that EPI was far more informative about neuron-type input  
 386 responsibility than the predictions afforded through analysis. By flexibly conditioning this V1 model  
 387 on different emergent properties, we performed an exploratory analysis of a *model* rather than a  
 388 dataset, which generated and proved out a set of testable predictions.

389 Of course, exploratory analyses can also be directed. For example, when interested in model  
390 changes during learning, one can use EPI to condition as we did in Section 3.4. This analysis  
391 identified experimentally testable predictions (proved out *in-silico*) of changes in connectivity in  
392 SC throughout learning. Precisely, we predict that an initial reduction in side mode eigenvalue,  
393 and a steady increase in task mode eigenvalue will take place, during learning, in the effective  
394 connectivity matrices of learning rats.

395 In our final analysis, we present a novel procedure for doing statistical inference on interpretable  
396 parameterizations of RNNs executing simple tasks . This methodology relies on recently extended  
397 theory of responses in random neural networks with minimal structure [26]. With this methodology,  
398 we can finally open the probabilistic model selection toolkit reasoning about the connectivity of  
399 RNNs solving tasks.

## 400 References

- 401 [1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 402 [2] John J Hopfield. Neural networks and physical systems with emergent collective computational  
403 abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 404 [3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural  
405 networks. *Physical review letters*, 61(3):259, 1988.
- 406 [4] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-  
407 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,  
408 17(11):4382–4388, 1997.
- 409 [5] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in  
410 perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 411 [6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Confer-  
412 ence on Learning Representations*, 2014.
- 413 [7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation  
414 and variational inference in deep latent gaussian models. *International Conference on Machine  
415 Learning*, 2014.

- [8] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical neural population models through nonlinear embeddings. In *Advances in neural information processing systems*, pages 163–171, 2016.
- [9] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- [10] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders. *bioRxiv*, page 399246, 2018.
- [11] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, page 1, 2018.
- [12] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- [13] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- [14] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall, Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet: nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural Information Processing Systems*, 2019.
- [15] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241, 2018.
- [16] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, 143(1):07B201\_1, 2015.
- [17] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 2015.

- [445] [18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
[446] *arXiv preprint arXiv:1605.08803*, 2016.
- [447] [19] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density  
[448] estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.
- [449] [20] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow  
[450] networks. *International Conference on Learning Representations*, 2017.
- [451] [21] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-  
[452] free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–  
[453] 5533, 2017.
- [454] [22] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness,  
[455] and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.
- [456] [23] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an  
[457] electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.  
[458] *Neuron*, 77(5):845–858, 2013.
- [459] [24] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and vi-  
[460] sual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,  
[461] 115(3):1399–1409, 2016.
- [462] [25] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J  
[463] Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor  
[464] routing. *bioRxiv*, page 245613, 2018.
- [465] [26] Francesca Mastrogiovanni and Srdjan Ostojic. Linking connectivity, dynamics, and computa-  
[466] tions in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- [467] [27] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-  
[468] ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience  
[469] Meeting (COSYNE), Lisbon, Portugal*, 2019.
- [470] [28] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,  
[471] Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with  
[472] degenerate solution networks. *Bernstein Conference*, 2019.

- 473 [29] Jan-Matthis Lueckmann, Pedro Goncalves, Chaitanya Chintaluri, William F Podlaski, Giacomo Bassetto, Tim P Vogels, and Jakob H Macke. Amortised inference for mechanistic models  
474 of neural dynamics. In *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 477 [30] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural  
478 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 480 [31] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodulation. *Neural Networks*, 15(4-6):479–493, 2002.
- 482 [32] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit  
483 parameters. *Nature neuroscience*, 7(12):1345, 2004.
- 484 [33] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,  
485 1957.
- 486 [34] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:  
487 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 488 [35] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise  
489 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 490 [36] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of  
491 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 492 [37] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory  
493 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–  
494 592, 2009.
- 495 [38] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear  
496 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,  
497 85(2):402–417, 2015.
- 498 [39] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and  
499 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,  
500 5(10):793, 2004.

- 501 [40] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of  
502 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-*  
503 *biology*, 71(1):45–61, 2011.
- 504 [41] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex:  
505 From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 506 [42] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-  
507 bition of inhibition in visual cortex: the logic of connections between molecularly distinct  
508 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 509 [43] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing Wang.  
510 Paradoxical response reversal of top- down modulation in cortical circuits with three interneu-  
511 ron types. *Elife*, 6:1–15, 2017.
- 512 [44] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying  
513 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 514 [45] (2018) Allen Institute for Brain Science. Layer 4 model of v1. available from:  
515 <https://portal.brain-map.org/explore/models/l4-mv1>.
- 516 [46] Yayan N Billeh, Binghuang Cai, Sergey L Gratiy, Kael Dai, Ramakrishnan Iyer, Nathan W  
517 Gouwens, Reza Abbasi-Asl, Xiaoxuan Jia, Joshua H Siegle, Shawn R Olsen, et al. Systematic  
518 integration of structural and functional data into multi-scale models of mouse primary visual  
519 cortex. *bioRxiv*, page 662189, 2019.
- 520 [47] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain  
521 regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.
- 522 [48] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*  
523 *opinion in neurobiology*, 46:1–6, 2017.
- 524 [49] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-  
525 dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.
- 526 [50] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like  
527 dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*,  
528 page 696088, 2019.

- 529 [51] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,  
530 13(8):1713–1720, 2001.
- 531 [52] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.  
532 A statistical paradigm for neural spike train decoding applied to position prediction from  
533 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–  
534 7425, 1998.
- 535 [53] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding  
536 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 537 [54] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and  
538 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis  
539 of neural population activity. In *Advances in neural information processing systems*, pages  
540 1881–1888, 2009.
- 541 [55] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W  
542 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.  
543 *Science*, 349(6244):184–187, 2015.
- 544 [56] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable  
545 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th Inter-*  
546 *national Conference on Machine Learning*, 2019.
- 547 [57] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,  
548 2019.
- 549 [58] Steven H Strogatz. Nonlinear dynamics and chaos: with applications to physics. *Biology,*  
550 *Chemistry, and Engineering (Studies in Nonlinearity)*, Perseus, Cambridge, UK, 1994.
- 551 [59] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial*  
552 *Intelligence and Statistics*, pages 814–822, 2014.
- 553 [60] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and  
554 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 555 [61] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
556 *Proceedings of the 5th International Conference on Learning Representations*, 2017.

557 **A Methods**

558 **A.1 Emergent property inference (EPI)**

559 Emergent property inference (EPI) learns distributions of theoretical model parameters that pro-  
 560 duce emergent properties of interest. EPI combines ideas from likelihood-free variational inference  
 561 [21] and maximum entropy flow networks [20]. A maximum entropy flow network is used as a deep  
 562 probability distribution for the parameters, while these samples often parameterize a differentiable  
 563 model simulator, which may lack a tractable likelihood function.

564 Consider model parameterization  $z$  and data  $x$  generated from some theoretical model simulator  
 565 represented as  $p(x | z)$ , which may be deterministic or stochastic. Theoretical models usually have  
 566 known sampling procedures for simulating activity given a circuit parameterization, yet often lack  
 567 an explicit likelihood function due to the nonlinearities and dynamics. With EPI, a distribution  
 568 on parameters  $z$  is learned, that yields an emergent property of interest  $\mathcal{B}$ ,

$$\mathcal{B} \triangleq \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] = \mu \quad (12)$$

569 by making an approximation  $q_\theta(z)$  to  $p(z | \mathcal{B})$  (see Section A.1.5). So, over the DSN distribution  
 570  $q_\theta(z)$  of model  $p(x | z)$  for behavior  $\mathcal{B}$ , the emergent properties  $T(x)$  are constrained in expectation  
 571 to  $\mu$ .

572 In deep probability distributions, a simple random variable  $w \sim q_0$  is mapped deterministically via  
 573 a function  $f_\theta$  parameterized by a neural network to the support of the distribution of interest where  
 574  $z = f_\theta(w) = f_l(\dots f_1(w))$ . Given a theoretical model  $p(x | z)$  and some behavior of interest  $\mathcal{B}$ , the  
 575 deep probability distributions are trained by optimizing the neural network parameters  $\theta$  to find a  
 576 good approximation  $q_\theta^*$  within the deep variational family  $\mathcal{Q}$  to  $p(z | \mathcal{B})$ .

577 In most settings (especially those relevant to theoretical neuroscience) the likelihood of the behavior  
 578 with respect to the model parameters  $p(T(x) | z)$  is unknown or intractable, requiring an alternative  
 579 to stochastic gradient variational Bayes [6] or black box variational inference[59]. These types  
 580 of methods called likelihood-free variational inference (LFVI, [21]) skate around the intractable  
 581 likelihood function in situations where there is a differentiable simulator. Akin to LFVI, DSNs are  
 582 optimized with the following objective for a given theoretical model, emergent property statistics  
 583  $T(x)$ , and emergent property constraints  $\mu$ :

$$\begin{aligned} q_\theta^*(z) &= \operatorname{argmax}_{q_\theta \in Q} H(q_\theta(z)) \\ \text{s.t. } \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] &= \mu \end{aligned} \tag{13}$$

584 Optimizing this objective is a technological accomplishment in its own right, the details of which  
 585 we elaborate in Section A.1.2. Before going through those details, we ground this optimization in  
 586 a toy example.

587 **A.1.1 Example: 2D LDS**

588 To gain intuition for EPI, consider two-dimensional linear dynamical systems,  $\tau \dot{x} = Ax$  with

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

589 that produce a band of oscillations. To do EPI with the dynamics matrix elements as the free  
 590 parameters  $z = [a_1, a_2, a_3, a_4]$ , and fixing  $\tau = 1$ , such that the posterior yields a band of oscillations,  
 591 the emergent property statistics  $T(x)$  are chosen to contain the first- and second-moments of the  
 592 oscillatory frequency  $\omega$  and the growth/decay factor  $d$  of the oscillating system. To learn the  
 593 distribution of real entries of  $A$  that yield a distribution of  $d$  with mean zero with variance  $0.25^2$ ,  
 594 and oscillation frequency  $\omega$  with mean 1 Hz with variance  $(0.1\text{Hz})^2$ , then we would select the real  
 595 part of the complex conjugate eigenvalues  $\operatorname{real}(\lambda_1) = d$  (via an arbitrary choice of eigenvalue of the  
 596 dynamics matrix  $\lambda_1$ ) and the positive imaginary component of one of the eigenvalues  $\operatorname{imag}(\lambda_1) =$   
 597  $2\pi\omega$  as the emergent property statistics. Those emergent property statistics are then constrained  
 598 to

$$\mu = \mathbb{E} \begin{bmatrix} \operatorname{real}(\lambda_1) \\ \operatorname{imag}(\lambda_1) \\ (\operatorname{real}(\lambda_1) - 0)^2 \\ (\operatorname{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \tag{14}$$

599 where  $\omega = 1\text{Hz}$ . Unlike the models we study in the paper which calculate  $\mathbb{E}_{x \sim p(x|z)} [T(x)]$  via  
 600 forward simulation, we have a closed form for the eigenvalues of the dynamics matrix.  $\lambda$  can be  
 601 calculated using the quadratic formula:

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2} \tag{15}$$



Fig. S2: A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix  $A$  are the parameters. B. The DSN distribution for a 2D LDS with  $\tau = 1$  that produces an average of 1Hz oscillations with some small amount of variance. C. Entropy throughout the optimization. At the beginning of each augmented Lagrangian epoch (5,000 iterations), the entropy dips 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 move closer to their constraints.

602 where  $\lambda_1$  is the eigenvalue of  $\frac{1}{\tau}A$  with greatest real part. Even though  $\mathbb{E}_{x \sim p(x|z)}[T(x)]$  is calculable  
 603 directly via a closed form function and does not require simulation, we cannot derive the distribution  
 604  $q_\theta^*$  directly. This is due to the formally hard problem of the backward mapping: finding the natural  
 605 parameters  $\eta$  from the mean parameters  $\mu$  of an exponential family distribution [60]. Instead, we  
 606 can use EPI to learn the linear system parameters producing such a band of oscillations (Fig. S2B).  
 607 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the  
 608 parameter distribution. To validate our method (further than that of the underlying technology  
 609 on a ground truth solution [20]) we can analytically derive the contours of the probability density  
 610 from the emergent property statistics and values (Fig. S3). In the  $a_1 - a_4$  plane, is a black line  
 611 at  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ , a dotted black line at the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 1$ , and a  
 612 grey line at twice the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 2$  (Fig. S3A). Here the lines denote the  
 613 set of solutions at fixed behaviors, which overlay the posterior obtained through EPI. The learned  
 614 DSN distribution precisely reflects the desired statistical constraints and model degeneracy in the  
 615 sum of  $a_1$  and  $a_4$ . Intuitively, the parameters equivalent with respect to emergent property statistic  
 616  $\text{real}(\lambda_1)$  have similar log densities.

617 To explain the structure in the bimodality of the DSN posterior, we can look at the imaginary  
 618 component of  $\lambda_1$ . When  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ , we have

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4-a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

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

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \quad (17)$$

621 Since we constrained  $\mathbb{E}_{z \sim q_\theta}[\text{imag}(\lambda)] = 2\pi$  (with  $\omega = 1$ ), we can plot contours of the equation  
 622  $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$  for various  $a_1a_4$  (Fig. S3A). If  $\sigma_{1,4} = \mathbb{E}_{z \sim q_\theta}(|a_1a_4 - E_{q_\theta}[a_1a_4]|)$ ,  
 623 then we plot the contours as  $a_1a_4 = 0$  (black),  $a_1a_4 = -\sigma_{1,4}$  (black dotted), and  $a_1a_4 = -2\sigma_{1,4}$   
 624 (grey dotted) (Fig. S3B). This validates the curved structure of the inferred distribution learned  
 625 through EPI. We take steps in negative standard deviation of  $a_1a_4$  (dotted and gray lines), since  
 626 there are few positive values  $a_1a_4$  in the posterior. Subtler model-behavior combinations will have  
 627 even more complexity, further motivating the use of EPI for understanding these systems. Indeed,  
 628 we sample a distribution of systems oscillating near 1Hz (Fig. S4).

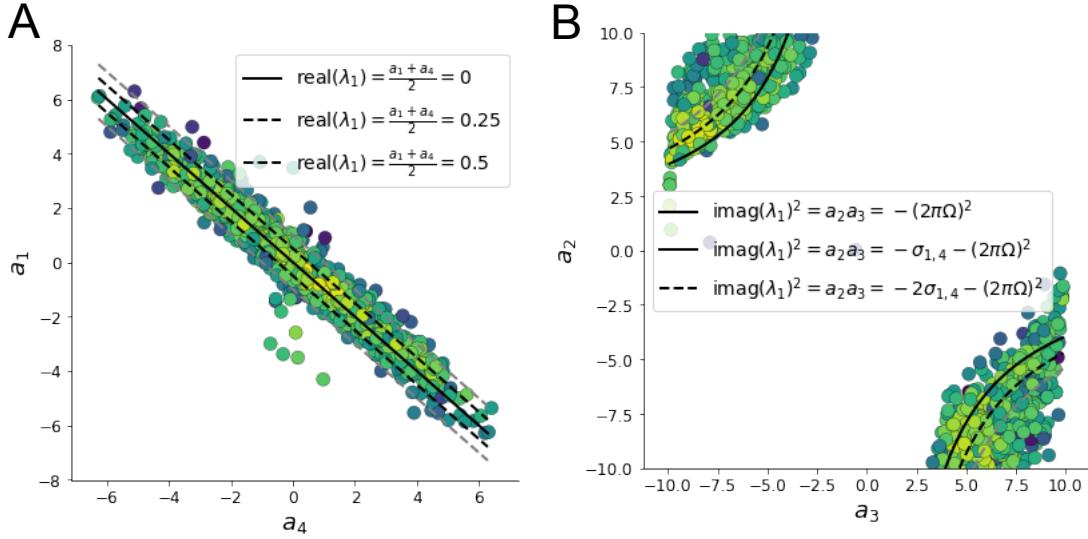


Fig. S3: A. Probability contours in the  $a_1 - a_4$  plane can be derived from the relationship to emergent property statistic of growth/decay factor. B. Probability contours in the  $a_2 - a_3$  plane can be derived from relationship to the emergent property statistic of oscillation frequency.

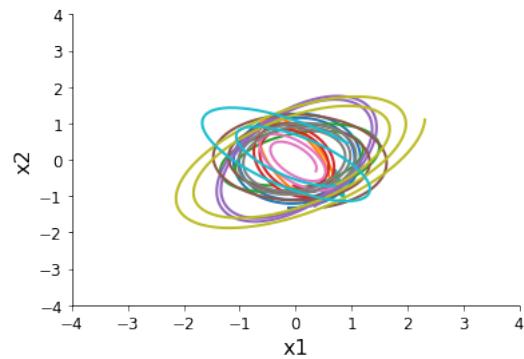


Fig. S4: Sampled dynamical system trajectories from the EPI distribution. Each trajectory is initialized at  $x(0) = \left[ \frac{\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2} \right]$ .

629 **A.1.2 Augmented Lagrangian optimization**

630 To optimize  $q_\theta(z)$  in Equation 13, the constrained optimization is performed using the augmented  
 631 Lagrangian method. The following objective is minimized:

$$L(\theta; \eta, c) = -H(q_\theta) + \eta^\top R(\theta) + \frac{c}{2} \|R(\theta)\|^2 \quad (18)$$

632 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 and  $c$  is the penalty  
 633 coefficient. For a fixed  $(\eta, c)$ ,  $\theta$  is optimized with stochastic gradient descent. A low value of  $c$  is  
 634 used initially, and increased during each augmented Lagrangian epoch – a period of optimization  
 635 with fixed  $\eta$  and  $c$  for a given number of stochastic optimization iterations. Similarly,  $\eta$  is tuned  
 636 each epoch based on the constraint violations. For the linear 2-dimensional system (Fig. S2C)  
 637 optimization hyperparameters are initialized to  $c_1 = 10^{-4}$  and  $\eta_1 = \mathbf{0}$ . The penalty coefficient  
 638 is updated based on a hypothesis test regarding the reduction in constraint violation. The p-  
 639 value of  $E[\|R(\theta_{k+1})\|] > \gamma \mathbb{E}[\|R(\theta_k)\|]$  is computed, and  $c_{k+1}$  is updated to  $\beta c_k$  with probability  
 640  $1 - p$ . Throughout the project,  $\beta = 4.0$  and  $\gamma = 0.25$  is used. The other update rule is  $\eta_{k+1} =$   
 641  $\eta_k + c_k \frac{1}{n} \sum_{i=1}^n (T(x^{(i)}) - \mu)$ . In this example, each augmented Lagrangian epoch ran for 2,000  
 642 iterations. We consider the optimization to have converged when a null hypothesis test of constraint  
 643 violations being zero is accepted for all constraints at a significance threshold 0.05. This is the dotted  
 644 line on the plots below depicting the optimization cutoff of EPI optimization for the 2-dimensional  
 645 linear system. If the optimization is left to continue running, entropy usually decreases, and  
 646 structural pathologies in the distribution may be introduced.

647 The intention is that  $c$  and  $\eta$  start at values encouraging entropic growth early in optimization.  
 648 Then, as they increase in magnitude with each training epoch, the constraint satisfaction terms are  
 649 increasingly weighted, resulting in a decrease in entropy. Rather than using a naive initialization,  
 650 before EPI, we optimize the deep probability distribution parameters to generate samples of an  
 651 isotropic Gaussian of a selected variance, such as 1.0 for the 2D LDS example. This provides a  
 652 convenient starting point, whose level of entropy is controlled by the user.

653 **A.1.3 Normalizing flows**

654 Since we are optimizing parameters  $\theta$  of our deep probability distribution with respect to the  
 655 entropy, we will need to take gradients with respect to the log-density of samples from the deep  
 656 probability distribution.

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

$$\nabla_\theta H(q_\theta(z)) = \mathbb{E}_{w \sim q_0} [-\nabla_\theta \log(q_\theta(f_\theta(w)))] \quad (20)$$

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

$$q(z') = q(f^{-1}(z')) \left| \det \frac{\partial f^{-1}(z')}{\partial z'} \right| = q(z) \left| \det \frac{\partial f(z)}{\partial z} \right|^{-1} \quad (21)$$

However, this computation has cubic complexity in dimensionality for fully connected layers. By restricting our layers to normalizing flows [17] – bijective functions with fast log determinant jacobian computations, we can tractably optimize deep generative models with objectives that are a function of sample density, like entropy. Most of our analyses use real NVP [61], which have proven effective in our architecture searches, and have the advantageous features of fast sampling and fast density evaluation.

#### A.1.4 Related work

(To come)

669

#### A.1.5 Emergent property inference as variational inference in an exponential family

(To come)

672

## A.2 Theoretical models

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

676 **A.2.1 Stomatogastric ganglion**

677 Each neuron's membrane potential  $x_m(t)$  is the solution of the following differential equation.

$$C_m \frac{dx_m}{dt} = -[h_{leak}(x; z) + h_{Ca}(x; z) + h_K(x; z) + h_{hyp}(x; z) + h_{elec}(x; z) + h_{syn}(x; z)] \quad (22)$$

678 The membrane potential of each neuron is affected by the leak, calcium, potassium, hyperpolariza-  
 679 tion, electrical and synaptic currents, respectively. The capacitance of the cell membrane was set to  
 680  $C_m = 1nF$ . Each current is a function of the neuron's membrane potential  $x_m$  and the parameters  
 681 of the circuit such as  $g_{el}$  and  $g_{syn}$ , whose effect on the circuit is considered in the motivational  
 682 example of EPI in Fig. 1. Specifically, the currents are the difference in the neuron's membrane  
 683 potential and that current type's reversal potential multiplied by a conductance:

$$h_{leak}(x; z) = g_{leak}(x_m - V_{leak}) \quad (23)$$

$$h_{elec}(x; z) = g_{el}(x_m^{post} - x_m^{pre}) \quad (24)$$

$$h_{syn}(x; z) = g_{syn}S_\infty^{pre}(x_m^{post} - V_{syn}) \quad (25)$$

$$h_{Ca}(x; z) = g_{Ca}M_\infty(x_m - V_{Ca}) \quad (26)$$

$$h_K(x; z) = g_KN(x_m - V_K) \quad (27)$$

$$h_{hyp}(x; z) = g_hH(x_m - V_{hyp}) \quad (28)$$

684 The reversal potentials were set to  $V_{leak} = -40mV$ ,  $V_{Ca} = 100mV$ ,  $V_K = -80mV$ ,  $V_{hyp} = -20mV$ ,  
 685 and  $V_{syn} = -75mV$ . The other conductance parameters were fixed to  $g_{leak} = 1 \times 10^{-4}\mu S$ .  $g_{Ca}$ ,  
 686  $g_K$ , and  $g_{hyp}$  had different values based on fast, intermediate (hub) or slow neuron. Fast:  $g_{Ca} =$   
 687  $1.9 \times 10^{-2}$ ,  $g_K = 3.9 \times 10^{-2}$ , and  $g_{hyp} = 2.5 \times 10^{-2}$ . Intermediate:  $g_{Ca} = 1.7 \times 10^{-2}$ ,  $g_K = 1.9 \times 10^{-2}$ ,  
 688 and  $g_{hyp} = 8.0 \times 10^{-3}$ . Intermediate:  $g_{Ca} = 8.5 \times 10^{-3}$ ,  $g_K = 1.5 \times 10^{-2}$ , and  $g_{hyp} = 1.0 \times 10^{-2}$ .

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

$$M_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_m - v_1}{v_2} \right) \right) \quad (29)$$

$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (30)$$

$$N_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_m - v_3}{v_4} \right) \right) \quad (31)$$

$$\lambda_N = \phi_N \cosh \left( \frac{x_m - v_3}{2v_4} \right) \quad (32)$$

699

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (33)$$

700

$$H_\infty = \frac{1}{1 + \exp\left(\frac{x_m + v_5}{v_6}\right)} \quad (34)$$

701

$$\tau_h = 272 - \left( \frac{-1499}{1 + \exp\left(\frac{-x_m + v_7}{v_8}\right)} \right) \quad (35)$$

702 where we set  $v_1 = 0mV$ ,  $v_2 = 20mV$ ,  $v_3 = 0mV$ ,  $v_4 = 15mV$ ,  $v_5 = 78.3mV$ ,  $v_6 = 10.5mV$ ,  
 703  $v_7 = -42.2mV$ ,  $v_8 = 87.3mV$ ,  $v_9 = 5mV$ , and  $v_{th} = -25mV$ . These are the same parameter  
 704 values used in [23].

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

$$S_\infty = \frac{1}{1 + \exp\left(\frac{v_{th} - x_m}{v_9}\right)} \quad (36)$$

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

708 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated  
 709 for  $T = 500$  time steps of  $dt = 25ms$ . In EPI, since gradients are taken through the simulation  
 710 process, the number of time steps are kept as modest if possible. The chosen  $dt$  and  $T$  were the  
 711 most computationally convenient choices yielding accurate frequency measurement.

712 Our original approach to measuring frequency was to take the max of the fast Fourier transform  
 713 (FFT) of the simulated time series. There are a few key considerations here. One is resolution  
 714 in frequency space. Each FFT entry will correspond to a signal frequency of  $\frac{F_s k}{N}$ , where  $N$  is  
 715 the number of samples used for the FFT,  $F_s = \frac{1}{dt}$ , and  $k \in [0, 1, \dots, N - 1]$ . Our resolution is  
 716 improved by increasing  $N$  and decreasing  $dt$ . Increasing  $N = T - b$ , where  $b$  is some fixed number  
 717 of buffer burn-in initialization samples, necessitates an increase in simulation time steps  $T$ , which  
 718 directly increases computational cost. Increasing  $F_s$  (decreasing  $dt$ ) increases system approximation  
 719 accuracy, but requires more time steps before a full cycle is observed. At the level of  $dt = 0.025$ ,  
 720 thousands of temporal samples were required for resolution of .01Hz. These challenges in frequency  
 721 resolution with the discrete Fourier transform motivated the use of an alternative basis of complex  
 722 exponentials. Instead, we used a basis of complex exponentials with frequencies from 0.0-1.0 Hz at  
 723 0.01Hz resolution,  $\Phi = [0.0, 0.01, \dots, 1.0]^\top$

724 Another consideration was that the frequency spectra of the hub neuron has several peaks. This  
 725 was due to high-frequency sub-threshold activity. The maximum frequency was often not the firing

frequency. 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 pre-processing, the maximum frequency in the filter bank accurately reflected the firing frequency.

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

$$u_i = \frac{|\mathcal{X}_i|^\alpha}{\sum_{k=1}^N |\mathcal{X}_i(k)|^\alpha} \quad (37)$$

The frequency is then calculated as  $\omega = u_i^\top \Phi$  with  $\alpha = 100$ .

Network syncing, like all other emergent properties in this work, are defined by the emergent property statistics and values. The emergent property statistics are the first- and second-moments of the firing frequencies. The first moments are set to 0.542Hz, while the second moments are set to 0.025Hz<sup>2</sup>.

$$E \begin{bmatrix} \omega_{\text{f1}} \\ \omega_{\text{f2}} \\ \omega_{\text{hub}} \\ \omega_{\text{s1}} \\ \omega_{\text{s2}} \\ (\omega_{\text{f1}} - 0.542)^2 \\ (\omega_{\text{f2}} - 0.542)^2 \\ (\omega_{\text{hub}} - 0.542)^2 \\ (\omega_{\text{s1}} - 0.542)^2 \\ (\omega_{\text{s2}} - 0.542)^2 \end{bmatrix} = \begin{bmatrix} 0.542 \\ 0.542 \\ 0.542 \\ 0.542 \\ 0.542 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \end{bmatrix} \quad (38)$$

For EPI in Fig 2C, we used a real NVP architecture with two coupling layers. Each coupling layer had two hidden layers of 10 units each, and we mapped onto a support of  $z \in \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 \\ 8 \end{bmatrix} \right]$ . We have shown the EPI optimization that converged with maximum entropy across 2 random seeds and augmented Lagrangian coefficient initializations of  $c_0=0$ , 2, and 5.

742 **A.2.2 Primary visual cortex**743 The dynamics of each neural populations average rate  $x = \begin{bmatrix} x_E \\ x_P \\ x_S \\ x_V \end{bmatrix}$  are given by:

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n \quad (39)$$

744 Some neuron-types largely lack synaptic projections to other neuron-types [42], and it is popular

745 to only consider a subset of the effective connectivities [24].

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & 0 \\ W_{PE} & W_{PP} & W_{PS} & 0 \\ W_{SE} & 0 & 0 & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & 0 \end{bmatrix} \quad (40)$$

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

747 of the synaptic strength (in mV)

$$M = \begin{bmatrix} 0.36 & 0.48 & 0.31 & 0.28 \\ 1.49 & 0.68 & 0.50 & 0.18 \\ 0.86 & 0.42 & 0.15 & 0.32 \\ 1.31 & 0.41 & 0.52 & 0.37 \end{bmatrix} \quad (41)$$

748 and connection probability

$$C = \begin{bmatrix} 0.16 & 0.411 & 0.424 & 0.087 \\ 0.395 & .451 & 0.857 & 0.02 \\ 0.182 & 0.03 & 0.082 & 0.625 \\ 0.105 & 0.22 & 0.77 & 0.028 \end{bmatrix} \quad (42)$$

749 Multiplying these connection probabilities and synaptic efficacies gives us an effective connectivity

750 matrix:

$$W_{\text{full}} = C \odot M = \begin{bmatrix} 0.16 & 0.411 & 0.424 & 0.087 \\ 0.395 & .451 & 0.857 & 0.02 \\ 0.182 & 0.03 & 0.082 & 0.625 \\ 0.105 & 0.22 & 0.77 & 0.028 \end{bmatrix} \quad (43)$$

751 From use the entries of this full effective connectivity matrix that are not considered to be ineffectual.

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

$$h = b + dh \quad (44)$$

757 All simulations of this system had  $T = 100$  time points, a time step  $dt = 5\text{ms}$ , and time constant  
 758  $\tau = 20\text{ms}$ . And the system was initialized to a random draw  $x(0)_i \sim \mathcal{N}(1, 0.01)$ .

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

$$\dot{x}_i = -x_i + f(u_i) \quad (45)$$

760 where the input to each neuron is

$$u_i = \sum_j W_{ij}x_j + h_i \quad (46)$$

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

$$\frac{dx_{ss}}{dh} = F(W \frac{dx_{ss}}{dh} + I) \quad (47)$$

762 which is calculable by

$$\frac{dx_{ss}}{dh} = (F^{-1} - W)^{-1} \quad (48)$$

763 The emergent property we considered was the first and second moments of the change in rate  $dx$   
 764 between the baseline input  $h = b$  and  $h = b + dh$ . We use the following notation to indicate that  
 765 the emergent property statistics were set to the following values:

$$\mathcal{B}(\alpha, y) \triangleq \mathbb{E} \begin{bmatrix} dx_{\alpha,ss} \\ (dx_{\alpha,ss} - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix} \quad (49)$$

766 In the final analysis for this model, we sweep the input one neuron at a time away from the mode  
 767 of each inferred distributions  $dh^* = z^* = \text{argmax}_z \log q_\theta(z \mid \mathcal{B}(\alpha, 0.1))$ . The differential responses  
 768  $dx_{\alpha,ss}$  are examined at perturbed inputs  $h = b + dh^* + \Delta h_\alpha u_\alpha$  where  $u_\alpha$  is a unit vector in the  
 769 dimension of  $\alpha$  and  $\Delta h_\alpha \in [-15, 15]$ .

770 For each  $\mathcal{B}(\alpha, y)$  with  $\alpha \in \{E, P, S, V\}$  and  $y \in \{0.1, 0.5\}$ , we ran EPI with five different random  
 771 initial seeds using an architecture of four coupling layers, each with two hidden layers of 10 units.

772 We set  $c_0 = 10^5$ . The support of the learned distribution was restricted to  $z_i \in [-5, 5]$ .

<sup>773</sup> **A.2.3 Superior colliculus**

<sup>774</sup> There are four total units: two in each hemisphere corresponding to the Pro/Contra and Anti/Ipsi  
<sup>775</sup> populations. Each unit has an activity ( $x_i$ ) and internal variable ( $u_i$ ) related by

$$x_i(t) = \left( \frac{1}{2} \tanh \left( \frac{v_i(t) - \epsilon}{\zeta} \right) + \frac{1}{2} \right) \quad (50)$$

<sup>776</sup>  $\epsilon = 0.05$  and  $\zeta = 0.5$  control the position and shape of the nonlinearity, respectively.

<sup>777</sup> We can order the elements of  $x_i$  and  $v_i$  into vectors  $x$  and  $v$  with elements

$$x = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad v = \begin{bmatrix} v_{LP} \\ v_{LA} \\ v_{RP} \\ v_{RA} \end{bmatrix} \quad (51)$$

<sup>778</sup> The internal variables follow dynamics:

$$\tau \frac{dv}{dt} = -v + Wx + h + \sigma dB \quad (52)$$

<sup>779</sup> with time constant  $\tau = 0.09s$  and Gaussian noise  $\sigma dB$  controlled by the magnitude of  $\sigma = 1.0$ . The  
<sup>780</sup> weight matrix has 8 parameters  $sW_P$ ,  $sW_A$ ,  $vW_{PA}$ ,  $vW_{AP}$ ,  $hW_P$ ,  $hW_A$ ,  $dW_{PA}$ , and  $dW_{AP}$  (Fig.  
<sup>781</sup> 4B).

$$W = \begin{bmatrix} sW_P & vW_{PA} & hW_P & dW_{PA} \\ vW_{AP} & sW_A & dW_{AP} & hW_A \\ hW_P & dW_{PA} & sW_P & vW_{PA} \\ dW_{AP} & hW_A & vW_{AP} & sW_A \end{bmatrix} \quad (53)$$

<sup>782</sup> The system receives five inputs throughout each trial, which has a total length of 1.8s.

$$h = h_{\text{rule}} + h_{\text{choice-period}} + h_{\text{light}} \quad (54)$$

<sup>783</sup> There are rule-based inputs depending on the condition,

$$h_{P,\text{rule}}(t) = \begin{cases} I_{P,\text{rule}} \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (55)$$

<sup>784</sup>

$$h_{A,\text{rule}}(t) = \begin{cases} I_{A,\text{rule}} \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (56)$$

785 a choice-period input,

$$h_{\text{choice}}(t) = \begin{cases} I_{\text{choice}} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (57)$$

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

$$h_{\text{light}}(t) = \begin{cases} I_{\text{light}} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Left} \\ I_{\text{light}} \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Right} \\ 0, & t \leq 1.2s \end{cases} \quad (58)$$

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

788 To produce a Bernoulli rate of  $p_{LP}$  in the Left, Pro condition (we can generalize this to either cue,  
789 or stimulus condition), let  $\hat{p}_i$  be the empirical average steady state (ss) response (final  $x_{LP}$  at end  
790 of task) over M=500 Gaussian noise draws for a given SC model parameterization  $z_i$ :

$$\hat{p}_i = \mathbb{E}_{\sigma dB} [x_{LP,ss} | s = L, c = P, z_i] = \frac{1}{M} \sum_{j=1}^M x_{LP,ss}(s = L, c = P, z_i, \sigma dB_j) \quad (59)$$

791 For the first constraint, the average over posterior samples (from  $q_\theta(z)$ ) to be  $p_{LP}$ :

$$\mathbb{E}_{z_i \sim q_\phi} [\mathbb{E}_{\sigma dB} [x_{LP,ss} | s = L, c = P, z_i]] = \mathbb{E}_{z_i \sim q_\phi} [\hat{p}_i] = p_{LP} \quad (60)$$

792 We can then ask that the variance of the steady state responses across Gaussian draws, is the  
793 Bernoulli variance for the empirical rate  $\hat{p}_i$ .

$$\mathbb{E}_{z \sim q_\phi} [\sigma_{err}^2] = 0 \quad (61)$$

794

$$\sigma_{err}^2 = Var_{\sigma dB} [x_{LP,ss} | s = L, c = P, z_i] - \hat{p}_i(1 - \hat{p}_i) \quad (62)$$

795 We have an additional constraint that the Pro neuron on the opposite hemisphere should have the  
796 opposite value. We can enforce this with a final constraint:

$$\mathbb{E}_{z \sim q_\phi} [d_P] = 1 \quad (63)$$

797

$$\mathbb{E}_{\sigma dB} [(x_{LP,ss} - x_{RP,ss})^2 | s = L, c = P, z_i] \quad (64)$$

798 We refer to networks obeying these constraints as Bernoulli, winner-take-all networks. Since the  
799 maximum variance of a random variable bounded from 0 to 1 is the Bernoulli variance ( $\hat{p}(1 - \hat{p})$ ),

and the maximum squared difference between two variables bounded from 0 to 1 is 1, we do not need to control the second moment of these test statistics. In reality, these variables are dynamical system states and can only exponentially decay (or saturate) to 0 (or 1), so the Bernoulli variance error and squared difference constraints can only be undershot. This is important to be mindful of when evaluating the convergence criteria. Instead of using our usual hypothesis testing criteria for convergence to the emergent property, we set a slack variable threshold for these technically infeasible constraints to 0.05.

Training DSNs to learn distributions of dynamical system parameterizations that produce Bernoulli responses at a given rate (with small variance around that rate) was harder to do than expected. There is a pathology in this optimization setup, where the learned distribution of weights is bimodal attributing a fraction  $p$  of the samples to an expansive mode (which always sends  $x_{LP}$  to 1), and a fraction  $1 - p$  to a decaying mode (which always sends  $x_{LP}$  to 0). This pathology was avoided using an inequality constraint prohibiting parameter samples that resulted in low variance of responses across noise.

In total, the emergent property of rapid task switching accuracy at level  $p$  was defined as

$$\mathcal{B}(p) \triangleq \begin{bmatrix} \hat{p}_P \\ \hat{p}_A \\ (\hat{p}_P - p)^2 \\ (\hat{p}_A - p)^2 \\ \sigma_{P,err}^2 \\ \sigma_{A,err}^2 \\ d_P \\ d_A \end{bmatrix} = \begin{bmatrix} p \\ p \\ 0.15^2 \\ 0.15^2 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad (65)$$

For each accuracy level  $p$ , we ran EPI for 10 different random seeds and selected the maximum entropy solution using an architecture of 10 planar flows with  $c_0 = 2$ . The support of  $z$  was  $\mathbb{R}^8$ .

#### 817 A.2.4 Rank-1 RNN

Recent work establishes a link between RNN connectivity weights and the resulting dynamical responses of the network, using dynamic mean field theory (DMFT) [26]. Specifically, DMFT describes the properties of activity in infinite-size neural networks given a distribution on the connectivity weights. In such a model, the connectivity of a rank-1 RNN (which was sufficient for

822 our task), has weight matrix  $W$ , whis is the sum of a random component with strength determined  
 823 by  $g$  and a structured component determined by the outer product of vectors  $m$  and  $n$ :

$$W = g\chi + \frac{1}{N}mn^\top, \quad (66)$$

824 where the activity  $x$  evolves as and  $I(t)$  is some input,  $\phi$  is the tanh nonlinearity, and  $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ .  
 825 The entries of  $m$  and  $n$  are drawn from Gaussian distributions  $m_i \sim \mathcal{N}(M_m, 1)$  and  $n_i \sim \mathcal{N}(M_n, 1)$ .  
 826 From such a parameterization, this theory produces consistency equations for the dynamic mean  
 827 field variables in terms of parameters like  $g$ ,  $M_m$ , and  $M_n$ , which we study in Section 3.5. That  
 828 is the dynamic mean field variables (e.g. the activity along along a vector  $\kappa_v$ , the total variance  
 829  $\Delta_0$ , structured variance  $\Delta_\infty$ , and the chaotic variance  $\Delta_T$ ) are written as functions of one another  
 830 in terms of connectivity parameters. The values of these variables can be used obtained using a  
 831 nonlinear system of equations solver. These dynamic mean field variables are then cast as task-  
 832 relevant variables with respect to the context of the provided inputs. Mastrogiuseppe et al. designed  
 833 low-rank RNN connectivities via minimalist connectivity parameters to solve canonical tasks from  
 834 behavioral neuroscience.

835 We consider the DMFT equation solver as a black box that takes in a low-rank parameterization  
 836  $z$  (e.g.  $z = [g \ M_m \ M_n]$ ) and outputs the values of the dynamic mean field variables, of which  
 837 we cast  $\kappa_r$  and  $\Delta_T$  as task-relevant variables  $\mu_{\text{post}}$  and  $\sigma_{\text{post}}^2$  in the Gaussian posterior conditioning  
 838 toy example. Importantly, the solution produced by the solver is differentiable with respect to the  
 839 input parameters, allowing us to use DMFT to calculate the emergent property statistics in EPI  
 840 to learn distributions on such connectivity parameters of RNNs that execute tasks.

841 Specifically, we solve for the mean field variables  $\kappa_r$ ,  $\kappa_n$ ,  $\Delta_0$  and  $\Delta_\infty$ , where the readout is nominally  
 842 chosen to point in the unit orthant  $r = [1 \ \dots \ 1]^\top$ . The consistency equations for these variables  
 843 in the presence of an constant input  $h(t) = y - (n - M_n)$  can be derived following [26] are

$$\begin{aligned} \kappa_r &= F_1(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = M_m \kappa_n + y \\ \kappa_n &= F_2(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = M_n \langle [\phi_i] \rangle + \langle [\phi'_i] \rangle \\ \frac{\Delta_0^2 - \Delta_\infty^2}{2} &= F_3(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = g^2 \left( \int \mathcal{D}z \Phi^2(\kappa_r + \sqrt{\Delta_0} z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\kappa_r + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right) \\ &\quad + (\kappa_n^2 + 1)(\Delta_0 - \Delta_\infty) \\ \Delta_\infty &= F_4(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = g^2 \int \mathcal{D}z \left[ \int \mathcal{D}x \phi(\kappa_r + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 + \kappa_n^2 + 1 \end{aligned} \quad (67)$$

844 where  $z$  here is a gaussian integration variable. We can solve these equations by simulating the

845 following Langevin dynamical system to a steady state.

$$\begin{aligned}
 l(t) &= \frac{\Delta_0(t)^2 - \Delta_\infty(t)^2}{2} \\
 \Delta_0(t) &= \sqrt{2x(t) + \Delta_\infty(t)^2} \\
 \frac{d\kappa_r(t)}{dt} &= -\kappa_r(t) + F(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\
 \frac{d\kappa_n(t)}{dt} &= -\kappa_n + G(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\
 \frac{dI(t)}{dt} &= -l(t) + H(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\
 \frac{d\Delta_\infty(t)}{dt} &= -\Delta_\infty(t) + L(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t))
 \end{aligned} \tag{68}$$

846 Then, the chaotic variance, which is necessary for the Gaussian posterior conditioning example, is  
 847 simply calculated via

$$\Delta_T = \Delta_0 - \Delta_\infty \tag{69}$$

848 **A.3 Supplementary Figures**

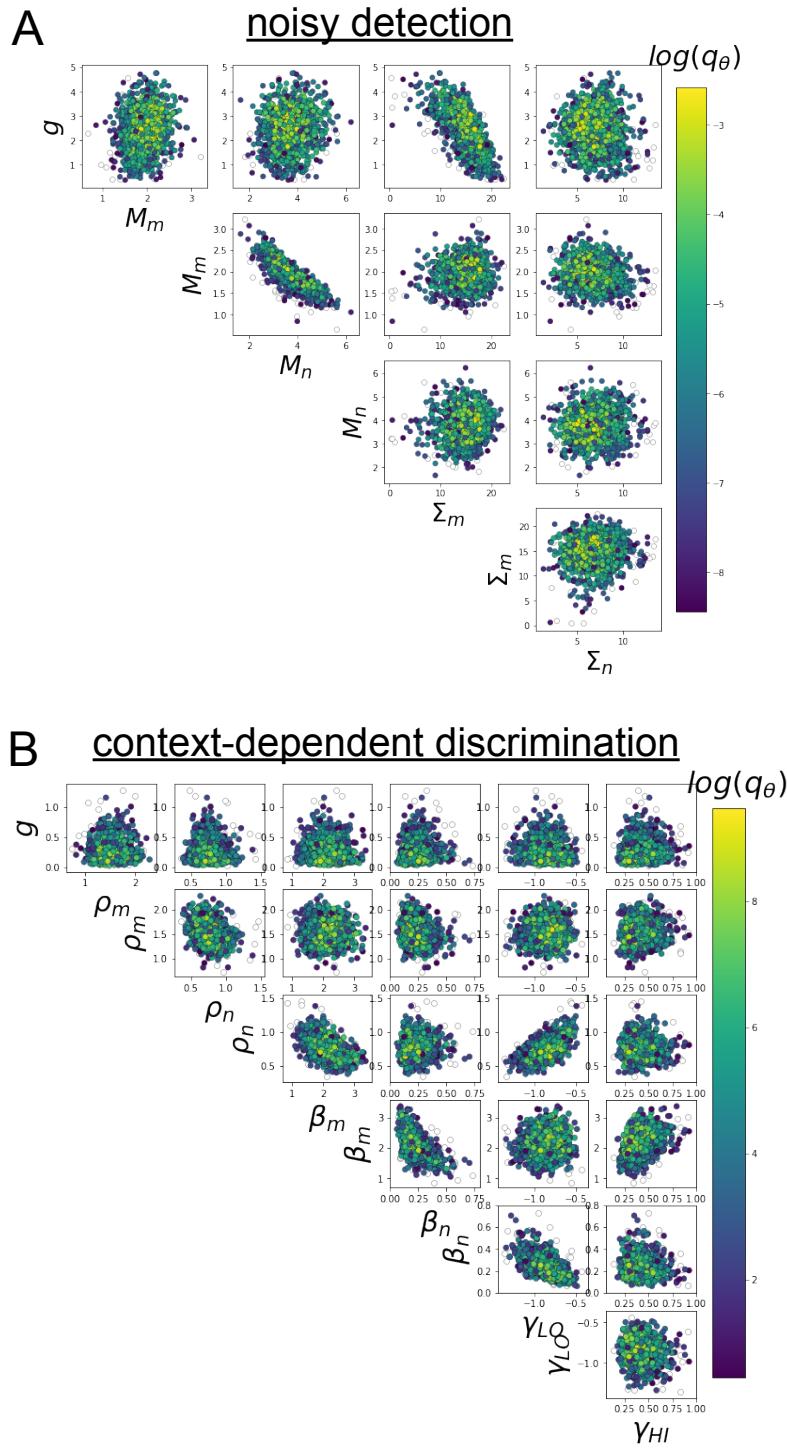


Fig. S1: A. EPI for rank-1 networks doing discrimination. B. EPI for rank-2 networks doing context-dependent discrimination.