

# Interrogating theoretical models of neural computation with deep learning

Sean R. Bittner, Agostina Palmigiano, Alex T. Piet, Chunyu A. Duan, Francesca Mastrogiovanni, Srdjan Ostojic, 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 of scientific importance. Such models are valuable when they give  
<sup>4</sup> rise to an experimentally observed phenomenon – whether behavioral or in terms of neural activity –  
<sup>5</sup> and thus can offer insight into neural computation. The operation of these circuits, like all models,  
<sup>6</sup> critically depends on the choices of model parameters. Historically, the gold standard has been  
<sup>7</sup> to analytically derive the relationship between model parameters and computational properties.  
<sup>8</sup> However, this enterprise quickly becomes infeasible as biologically realistic constraints are included  
<sup>9</sup> into the model, 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 offer  
<sup>13</sup> a principled means to understand the implications of model parameter choices on computational  
<sup>14</sup> properties of interest. We motivate this methodology with a worked example analyzing sensitivity in  
<sup>15</sup> the stomatogastric ganglion. We then use it to generate insights into neuron-type input-responsivity  
<sup>16</sup> in a model of primary visual cortex, a new understanding of rapid task switching in superior  
<sup>17</sup> colliculus models, and improved attribution of bias in low-rank recurrent neural networks. More  
<sup>18</sup> generally, this work suggests a departure from realism vs tractability considerations towards the  
<sup>19</sup> use of modern machine learning for sophisticated interrogation of biologically relevant models.

## <sup>20</sup> 2 Introduction

<sup>21</sup> The fundamental practice of theoretical neuroscience is to use a mathematical model to understand  
<sup>22</sup> neural computation, whether that computation enables perception, action, or some intermediate  
<sup>23</sup> processing [1]. In this field, a neural computation is systematized with a set of equations – the  
<sup>24</sup> model – and these equations are motivated by biophysics, neurophysiology, and other conceptual  
<sup>25</sup> considerations. The function of this system is governed by the choice of model parameters, which  
<sup>26</sup> when configured appropriately, give rise to a measurable signature of a computation. The work of

27 analyzing a model then becomes the inverse problem: given a computation of interest, how can we  
28 reason about these suitable parameter configurations – their likely values, their uniquenesses and  
29 degeneracies, their attractor states and phase transitions, and more?

30 Consider the idealized practice: a theorist considers a model carefully and analytically derives how  
31 model parameters govern the computation. Seminal examples of this gold standard include our  
32 field’s understanding of memory capacity in associative neural networks [2], chaos and autocorrela-  
33 tion timescales in random neural networks [3], and the paradoxical effect in excitatory/inhibitory  
34 networks [4]. Unfortunately, as circuit models include more biological realism, theory via analytic  
35 derivation becomes intractable. This fact creates an unfavorable tradeoff for the theorist. On the  
36 one hand, one may tractably analyze systems of equations with unrealistic assumptions (for ex-  
37 ample symmetry or gaussianity), producing accurate inferences about parameters of a too-simple  
38 model. On the other hand, one may choose a more biologically relevant model at the cost of *ad hoc*  
39 approaches to analysis (simply examining simulated activity), producing questionable or partial  
40 inferences about parameters of an appropriately complex, scientifically relevant model.

41 Of course, this same tradeoff has been confronted in many scientific fields and engineering problems  
42 characterized by the need to do inference in complex models. In response, the machine learning  
43 community has made remarkable progress in recent years, via the use of deep neural networks as a  
44 powerful inference engine: a flexible function family that can map observed phenomena (in this case  
45 the measurable signal of some computation) back to probability distributions quantifying the likely  
46 parameter configurations. One celebrated example of this approach from the machine learning  
47 community, from which we draw key inspiration for this work, is the variational autoencoder [5, 6],  
48 which uses a deep neural network to induce an (approximate) posterior distribution on hidden  
49 variables in a latent variable model, given data. Indeed, these tools have been used to great success  
50 in neuroscience as well, in particular for interrogating parameters (sometimes treated as hidden  
51 states) in models of both cortical population activity [7, 8, 9, 10] and animal behavior [11, 12, 13].  
52 These works have used deep neural networks to expand the expressivity and accuracy of statistical  
53 models of neural data [14].

54 However, these inference tools have not significantly influenced the study of theoretical neuroscience  
55 models, for at least three reasons. First, at a practical level, the nonlinearities and dynamics of  
56 many theoretical models are such that conventional inference tools typically produce a narrow set  
57 of insights into these models. Indeed, only in the last few years has the deep learning toolkit  
58 expanded to a point of relevance to this class of problem. Second, the object of interest from a

59 theoretical model is not typically data itself, but rather a qualitative phenomenon – inspection of  
60 model behavior, or better, a measurable signature of some computation – an *emergent property* of  
61 the model. Third, because theoreticians work carefully to construct a model that has biological  
62 relevance, such a model as a result often does not fit cleanly into the framing of a statistical model.  
63 Technically, because many such models stipulate a noisy system of differential equations that can  
64 only be sampled or realized through forward simulation, they lack the explicit likelihood and priors  
65 central to the probabilistic modeling toolkit.

66 To address these three challenges, we developed an inference methodology – ‘emergent property  
67 inference’ – which learns a distribution over parameter configurations in a theoretical model. Crit-  
68 ically, this distribution is such that draws from the distribution (parameter configurations) corre-  
69 spond to systems of equations that give rise to a specified emergent property. First, we stipulate a  
70 bijective deep neural network that induces a flexible family of probability distributions over model  
71 parameterizations with a probability density we can calculate [16, ?, ?]. Second, we quantify the  
72 notion of emergent properties as a set of moment constraints on datasets generated by the model.  
73 Thus an emergent property is not a single data realization, but a phenomenon or a feature of the  
74 model, which is the central object of interest to the theorist (unlike say the statistical neuroscien-  
75 tist). Conditioning on an emergent property requires an extension of deep probabilistic inference  
76 methods, which we have produced [17]. Third, because we can not assume the theoretical model  
77 has explicit likelihood on data or the emergent property of interest, we use stochastic gradient  
78 techniques in the spirit of likelihood free variational inference [18]. Taken together, emergent prop-  
79 erty inference (EPI) provides a methodology for inferring and then reasoning about parameter  
80 configurations that give rise to particular emergent phenomena in theoretical models.

81 Equipped with this methodology, we investigated three models of current importance in theoretical  
82 neuroscience. These models were chosen to demonstrate generality through ranges of biological  
83 realism (conductance-based biophysics to recurrent neural networks), neural system function (pat-  
84 tern generation to abstract cognitive function), and network scale (four to infinite neurons). First,  
85 to motivate the contribution of emergent property inference, we investigated network syncing in  
86 a classic model of the stomatogastric ganglion [19]. Second, we generated then evaluated a set  
87 of verifiable hypotheses of input-responsibility in a four neuron-type dynamical model of primary  
88 visual cortex. Third, we demonstrated how the systematic application of EPI to levels of task  
89 performance can generate experimentally testable hypotheses regarding connectivity in superior  
90 colliculus. Fourth, we leveraged the flexibility of EPI to uncover the sources of bias in a low-rank

91 recurrent neural network executing Bayesian inference. The novel scientific insights offered by EPI  
92 contextualize and clarify the previous studies exploring these models [19, 20, 21, 22] and more  
93 generally offer a quantitative grounding for theoretical models going forward, pointing a way to  
94 how rigorous statistical inference can enhance theoretical neuroscience at large.

95 We note that, during our preparation and early presentation of this work [23, 24], another work  
96 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural  
97 circuits [25]. We are excited by this broad problem being recognized by the community, and we  
98 emphasize that these works offer complementary neuroscientific contributions and use different  
99 technical methodologies. Scientifically, our work has focused primarily on systems-level theoretical  
100 models, while their focus has been on lower-level cellular models. Secondly, there are several key  
101 technical differences in the approaches (see Section A.1.4) perhaps most notably is our focus on  
102 the emergent property – the measurable signal of the computation in question, vs their focus  
103 on observed datasets; both certainly are worthy pursuits. The existence of these complementary  
104 methodologies emphasizes the increased importance and timeliness of both works.

### 105 3 Results

#### 106 3.1 Motivating emergent property inference of theoretical models

107 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-  
108 erty inference. First, the theorist designs or chooses an existing model that, it is hypothesized,  
109 captures the computation of interest. To ground this process in a well-known example, consider  
110 the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple  
111 rhythmic muscle activation patterns for digestion [26]. A model of the STG [19] is shown schemat-  
112 ically in Figure 1A, and note that the behavior of this model will be critically dependent on its  
113 parameterization – the choices of conductance parameters  $z = [g_{el}, g_{synA}]$ . Specifically, the two  
114 fast neurons ( $f_1$  and  $f_2$ ) mutually inhibit one another, and oscillate at a faster frequency than  
115 the mutually inhibiting slow neurons ( $s_1$  and  $s_2$ ), and the hub neuron (hub) couples with the fast  
116 or slow population or both. Second, once the model is selected, the theorist defines the emergent  
117 property, the measurable signal of scientific interest. To continue our running STG example, one  
118 such emergent property is the phenomenon of *network syncing* – in certain parameter regimes,  
119 the frequency of the hub neuron matches that of the fast and slow populations at an intermediate  
120 frequency. This emergent property is shown in Figure 1A at a frequency of 0.55Hz. Third, qualiti-

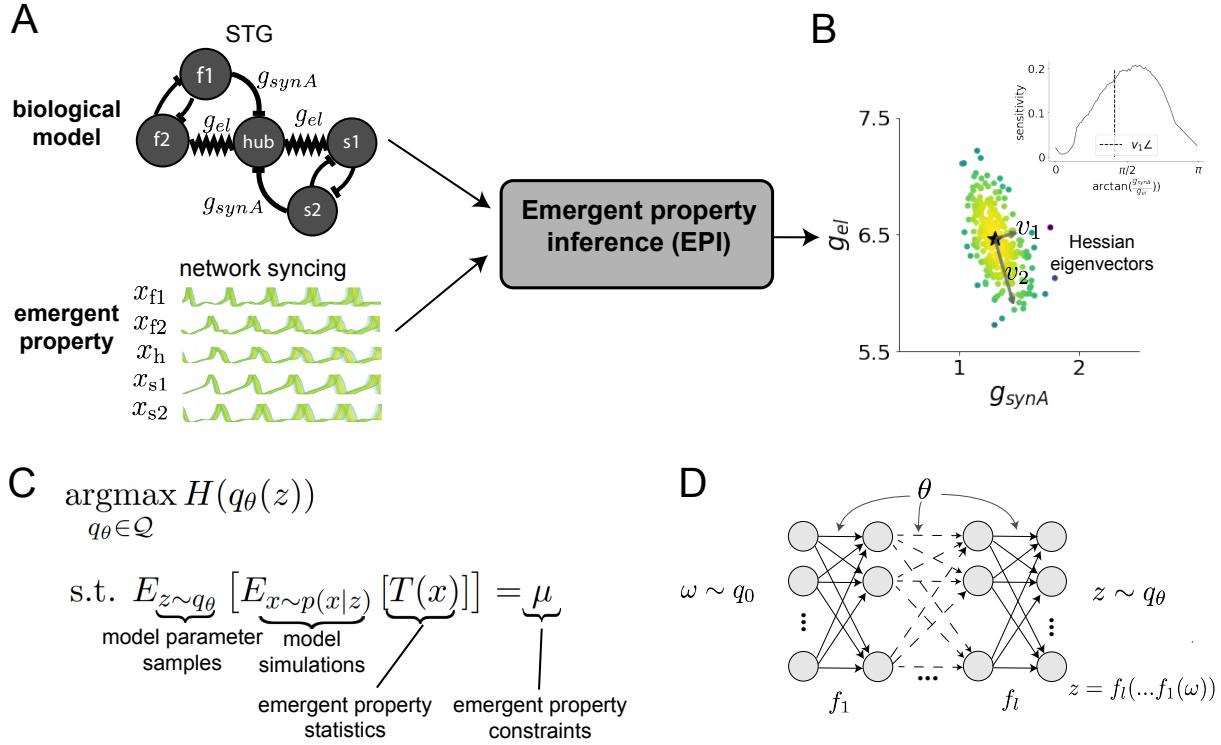


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. A. For a choice of model (STG) and emergent property (network syncing), emergent property inference (EPI) learns a posterior distribution of the model parameters  $z = [g_{el}, g_{synA}]^\top$  conditioned on network syncing. B. An EPI distribution of STG model parameters producing network syncing. The eigenvectors of the Hessian at the mode of the inferred distribution are indicated as  $v_1$  and  $v_2$ . (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. EPI learns a distribution  $q_\theta(z)$  of model parameters that produce an emergent property: the emergent property statistics  $T(x)$  are fixed in expectation over parameter distribution samples  $z \sim q_\theta(z)$  to particular values  $\mu$ . EPI distributions maximize randomness via entropy, although other measures are sensible. D. Deep probability distributions map a latent random variable  $\omega \sim q_0$ , where  $q_0$  is chosen to be simple distribution such as an isotropic gaussian, through a highly expressive function family  $f_\theta(\omega) = f_l(\dots f_1(\omega))$  parameterized by the neural network weights and biases  $\theta \in \Theta$ . This mapping induces an implicit probability model  $q(g_\theta(\omega)) \in \mathcal{Q}$

tative parameter analysis ensues: since precise mathematical analysis is intractable in this model, a brute force sweep of parameters is done. Subsequently, a qualitative description is formulated to describe of the different parameter configurations that lead to the emergent property. In this last step lies the opportunity for a precise quantification of the emergent property as a statistical feature of the model. Equipped with this methodology, we can infer a probability distribution over parameter configurations that produce this quantified emergent property.

Before presenting technical details (in the following section), let us understand emergent property inference schematically: the black box in Figure 1A takes, as input, the model and the specified emergent property, and produces as output the parameter distribution shown in Figure 1B. This distribution – represented for clarity as samples from the distribution – is then a scientifically meaningful and mathematically tractable object. It conveys parameter regions critical to the emergent property, directions in parameter space that will be invariant (or not) to that property. In the STG model, this distribution can be specifically queried to determine the prototypical parameter configuration for network syncing (the mode; Figure 1B star), and then how quickly network syncing will decay based on changes away that mode (Figure 1B, inset). While it is impossible to determine whether we have converged to the maximum entropy distribution, the emergent property statistics have converged to the emergent property values. For further validation, we apply EPI to condition a two-dimensional linear dynamical system model on a band of oscillations around 1Hz, from which we can analytically derive the contours of the inferred distribution (see Section A.1.1). Taken together, bringing careful inference to theoretical models offers deeper insight into the behavior of these models, and the opportunity to make rigorous this last step in the practice of theoretical neuroscience.

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

Emergent property inference (EPI) systematizes the three-step procedure of the previous section. First, we consider the model as a coupled set of differential (and potentially stochastic) equations [19]. In the running STG example the dynamical state  $x = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$  is the membrane potential for each neuron, which evolves according to the biophysical conductance-based equation:

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

where  $C_m=1nF$ , 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  $z = [g_{\text{el}}, g_{\text{synA}}]$  (see Section A.2.1). Second, we define the emergent property, which as above is network syncing: the phase locking of the population and its oscillation at an intermediate frequency (Figure 1A bottom). Quantifying this phenomenon is straightforward: we define network syncing to be that the spiking frequency of each neuron is close to an intermediate frequency of 0.55Hz. Thus, our measurable signature of computation – the firing frequencies of each neuron  $\Omega_{\text{f}1}(x), \Omega_{\text{f}2}(x)$ , etc.– are statistics of the membrane potential activity  $x$  which we insist be near a particular value 0.55Hz. This notion of an emergent property is then naturally embodied as a set of values for the emergent property statistics

$$E[T(x)] = \mu \quad (2)$$

where the first and second moments of these frequencies are chosen such that each neuron is firing near the network syncing frequency:  $E[\Omega_i] = 0.55$  and  $E[(\Omega_i - 0.55)^2] = 0.025^2$  for  $i \in \{\text{f}1, \text{f}2, \text{hub}, \text{s}1, \text{s}2\}$ . Third, having mathematically rationalized the above components, we can introduce deep generative modeling for performing emergent property inference. We seek a distribution over parameter configurations  $z$ , and insist that samples from this distribution produce the emergent property; in other words, they obey the constraints introduced in Equation 2. This results in the following optimization program:

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

It is worth emphasizing that  $T(x)$  is a vector of statistics, and  $\mu$  is a vector of their constraining values. The purpose of each element in this program is detailed in Figure 1D. Finally, we recognize that many distributions will produce the emergent property, 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 to model parameter distributions with minimal assumptions beyond some chosen structure [27, 28, 17, 29], but the EPI methods (not the results) offered here are largely unaffected by this choice. Stating such a problem is easy enough; finding a tractable and suitably flexible family of probability distributions ( $\mathcal{Q}$ ) is hard. EPI employs ‘normalizing flows’ [16, ?, ?], which are neural networks, which induce a flexible class of deep probability distributions (Fig. 1E). With normalizing flows, we leverage the tractable calculation of log sample probability  $\log q_\theta(z)$  to optimize entropy [17].

In EPI, the weights and biases  $\theta$  of the deep probability distribution are optimized by the objective in Equation 3. The optimization is complete when the sampled models with parameters  $z \sim q_\theta$

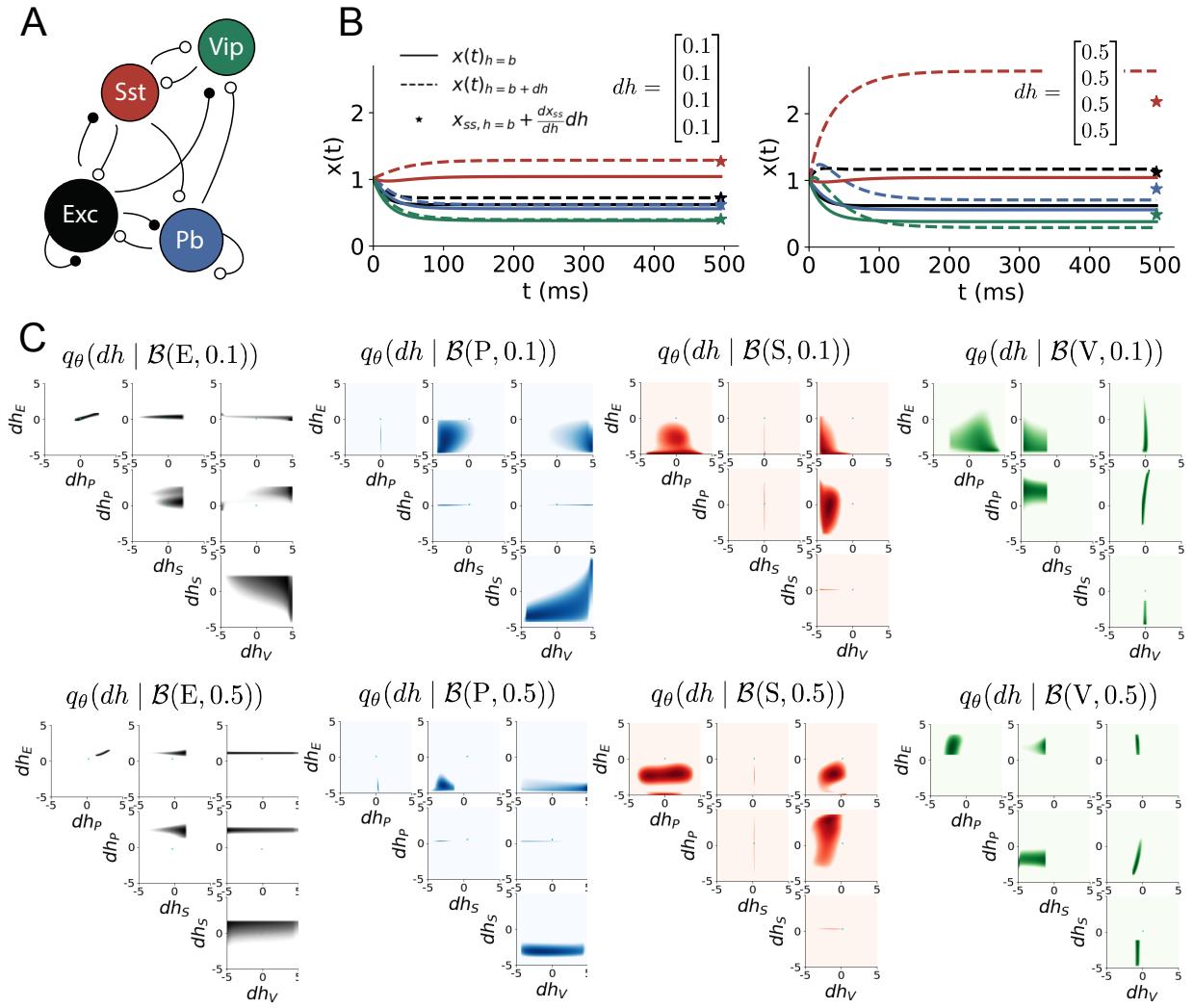


Figure 2: Exploring neuron-type responsivity in V1. A. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and vip (green) neurons. Some neuron-types largely do not form synaptic projections to others (excitatory and inhibitory projections filled and unfilled, respectively). B. Linear response predictions become inaccurate with greater input strength. V1 model simulations for input (solid)  $h = b$  and (dashed)  $h = b + dh$  with  $b = [1, 1, 1, 1]^T$  and (left)  $dh = [0.1, 0.1, 0.1, 0.1]^T$  (right)  $dh = [0.5, 0.5, 0.5, 0.5]^T$ . Stars indicate the linear response prediction. C. EPI distributions on differential input  $dh$  conditioned on differential response  $\mathcal{B}(\alpha, y)$  (see text). The linear prediction from two standard deviations away from  $y$  (from negative to positive) is overlaid in cyan (very small, near origin).

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 no different than its emergent property value. (see Section A.1.2). Armed with this method, we now prove out the value of this technology by investigating a range of models and using EPI to produce novel scientific insights.

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

The first model we focus on is a nonlinear dynamical model of sensory processing in primary visual cortex (V1). Dynamical models with two populations (excitatory and inhibitory) have already been used to reproduce a host of experimentally documented phenomena in primary visual cortex. In particular regimes of excitation and inhibition, these E/I models exhibit the paradoxical effect [4], selective amplification [30], surround suppression [31], and sensory integrative properties [32]. Extending this using experimental evidence three genetically-defined classes of inhibitory neurons [33, 34], recent work [20] has investigated a four-population model – excitatory (E), parvalbumin (P), somatostatin (S), and vasointestinal peptide (V) neurons – as shown in Fig. 2A. The dynamical state of this model is the firing rate of each neuron-type population  $x = [x_E, x_P, x_S, x_V]^\top$ , which evolves according to rectified and exponentiated dynamics:

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

with effective connectivity weights  $W$  and input  $h$ . In our analysis, we set the time constant  $\tau = 20\text{ms}$  and dynamics coefficient  $n = 2$ . Also, as is fairly standard, we obtain an informative estimate of the effective connectivities between these neuron-types  $W$  in mice by multiplying their probability of connection with their average synaptic strength [?] (citation on way) (see Section A.2.2). Given these fixed choices of  $W$ ,  $n$ , and  $\tau$ , we studied the system’s response to input

$$h = b + dh \quad (5)$$

where the input  $h$  is comprised of a baseline input  $b = [b_E, b_P, b_S, b_V]$  and a differential input  $dh = [dh_E, dh_P, dh_S, dh_V]$  to each neuron-type population. Throughout subsequent analyses, the baseline input is  $b = [1, 1, 1, 1]$ .

Having established our model, we now turn to defining the emergent property. We derived the linearized response of the system  $\frac{dx_{ss}}{dh}$  at fixed points  $x_{ss}$ . While this linearization accurately predicts differential responses  $dx_{ss} = [dx_{E,ss} \ dx_{P,ss} \ dx_{S,ss} \ dx_{V,ss}]$  for small differential inputs

205 to each population  $dh = [0.1 \ 0.1 \ 0.1 \ 0.1]$  (Fig. 2B, left), it can be misleading in such a nonlinear  
 206 model for a greater differential input strength  $dh = [0.1 \ 0.1 \ 0.1 \ 0.1]$  (Fig. 3B, right). In fact,  
 207 the linearly predicted response for the V-population to  $dh = [0.5 \ 0.5 \ 0.5 \ 0.5]$  was actually  
 208 in the opposite direction of the true response (Fig. 2B, right, green). This shows that currently  
 209 available approaches to deriving the steady state response of this system are limited.

210 To get a more comprehensive picture of the input-responsivity of each neuron-type, we used EPI  
 211 to learn a distribution of differential inputs  $dh$  that cause the rate of each neuron-type population  
 212  $\alpha \in \{E, P, S, V\}$  to increase by a value  $y \in 0.1, 0.5$  denoted by the emergent property

$$\mathcal{B}(\alpha, y) \leftrightarrow E \begin{bmatrix} dx_{\alpha,ss} \\ (dx_{\alpha,ss} - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix} \quad (6)$$

213 Note that we restrict the variance of the emergent property statistic  $dx_{\alpha,ss}$  by setting its second  
 214 moment to a small value. In Fig. 2C, each column visualizes the inferred distribution of  $dh$   
 215 corresponding to a specific neuron-type increase, while each row corresponds to amounts of increase  
 216 0.1 and 0.5. For visualization of this four-dimensional distribution, we show the two-dimensional  
 217 marginal densities. The inferred distributions suggest a slate of testable hypotheses. 1. As expected,  
 218 each neuron-type's rate is sensitive to its direct input. 2. The E- and P-populations are largely  
 219 unaffected by  $dh_V$ . 3. Similarly, The S-population is largely unaffected by  $dh_P$ . 4. Since EPI  
 220 showed that negative  $dh_E$  results in small  $dx_{V,ss}$ , but positive  $dh_E$  elicited a larger  $dx_{V,ss}$  we predict  
 221 that there is a nonmonotonic response of  $dx_{V,ss}$  with  $dh_E$ .

222 We evaluate these hypotheses by taking steps in individual neuron-type input  $\Delta h_\alpha$  away from the  
 223 modes of the inferred distributions

$$dh^* = z^* = \underset{z}{\operatorname{argmax}} \log q_\theta(z \mid \mathcal{B}(\alpha, 0.1)) \quad (7)$$

224 Now,  $dx_{\alpha,ss}$  is the steady state response to the system with input  $h = b + dh^* + \Delta h_\alpha u_\alpha$  where  
 225  $u_\alpha$  is a unit vector in the dimension of  $\alpha$ . Our hypotheses suggested by EPI are confirmed. 1.  
 226 the neuron-type responses are sensitive to their direct inputs (Fig. 3A black, 3B blue, 3C red, 3D  
 227 green), 2. the E- and P-populations are not affected by  $dh_V$  (Fig. 3A green, 3B green), 3. the  
 228 S-population is not affected by  $dh_P$  (Fig. 3C blue), and 4. the V-population has a nonmonotonic  
 229 response to  $dh_E$  (Fig. 3D black). All of this validated insight gained beyond what the analytic  
 230 linear prediction told us (Fig. 2C, cyan).

231 To this point, we have shown the utility of EPI on relatively low-level emergent properties like  
 232 network syncing and differential neuron-type population responses. In the remainder of the study,

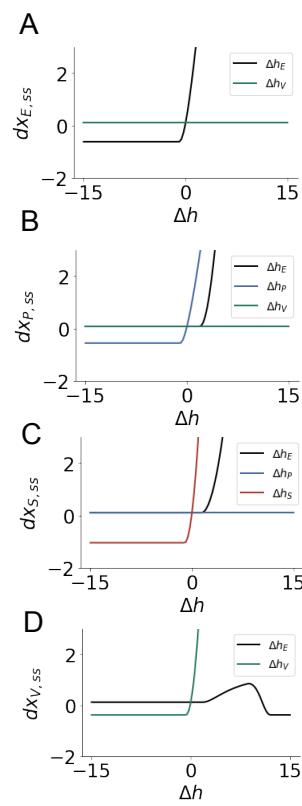


Figure 3: Confirming EPI generated hypotheses. A. Differential responses by the E-population to changes in individual input  $\Delta h_\alpha u_\alpha$  away from the mode of the EPI distribution  $dh^*$ . B-D Same plots for the P-, S-, and V-populations for the inputs for which hypotheses were formulated.

<sup>233</sup> we focus on using EPI to understand models of more abstract cognitive function.

### <sup>234</sup> 3.4 Identifying neural mechanisms of behavioral learning.

<sup>235</sup> Identifying measurable biological changes that result in improved behavior is important for neuro-  
<sup>236</sup> science, since they may indicate how the learning brain adapts. In a rapid task switching exper-  
<sup>237</sup> iment [?], where rats were to respond right (R) or left (L) to the side of a light stimulus in the  
<sup>238</sup> pro (P) task, and oppositely in the anti (A) task predicated by an auditory cue (Fig. 3A), neural  
<sup>239</sup> recordings exhibited two population of neurons in each hemisphere of superior colliculus (SC) that  
<sup>240</sup> simultaneously represented both task condition and motor response: the Pro/contralateral and  
<sup>241</sup> Anti/ipsilateral neurons [21]. Duan et al. proposed a model of SC that, like the V1 model analyzed  
<sup>242</sup> in the previous section, is a four-population dynamical system. Here, the neuron-type populations  
<sup>243</sup> are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right  
<sup>244</sup> (R)). The Pro- or Anti-populations receive an input determined by the cue, and then the left and  
<sup>245</sup> right populations receive an input based on the side of the light stimulus. Activities were bounded  
<sup>246</sup> from 0-1, so that a high output (1) of the Pro population in a given hemisphere corresponds to  
<sup>247</sup> 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 (0). 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. 3B).

Here, we used EPI to learn connectivity distributions consistent with various levels of accuracy in the rapid task switching behavioral paradigm. EPI was used to learn distributions of the SC weight matrix parameters  $z = W$  conditioned on of various levels of rapid task switching accuracy  $\mathcal{B}(p)$  for  $p \in \{50\%, 60\%, 70\%, 80\%, 90\%\}$  (see Section A.2.3). There is a decomposition for of the connectivity matrix  $W = QAQ^{-1}$ , in which the eigenvectors  $q_i$  are the same for all  $W$  (Fig. 3C). These consistent eigenvectors have intuitive roles in processing for this task, and are accordingly named the *all* - all neurons co-fluctuate, *side* - one side dominates the other, *task* - the Pro or Anti populations dominate the other, and *diag* - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair. The corresponding eigenvalues (e.g.  $a_{\text{task}}$ , which change according to  $W$ ) indicate the degree to which activity along that mode is increased or decreased by  $W$ .

For greater task accuracies, the task mode eigenvalue increases, indicating the criticality of supporting the task representation in the connectivity of  $W$ , (Fig. 4D, purple). Stepping from random chance (50%) networks to marginally task-performing (60%) networks, there is a marked decrease of the side mode eigenvalues (Fig. 3D, orange). Such side mode suppression remains in the models achieving greater accuracy, revealing its importance towards task performance. There were no interesting trends with learning in the all or diag mode. Significantly, we can conclude from our methodology optimized to find all connectivities consistent with a level of accuracy, that side mode suppression in  $W$  allows rapid task switching, and that greater task-mode representations in  $W$  increase accuracy. These hypotheses are proved out in the model (Fig. 3E). Thus, our EPI-enabled analyses produce novel, experimentally testable predictions that effective connectvity between these populations changes throughout learning in a way that increases its task mode and decreses its side mode eigenvalues.

### 3.5 Characterizing the sources of bias during approximate inference in RNNs

So far, each biologically realistic model we have studied was designed from fundamental biophysical principles, genetically- or functionally-defined neuron types. At a more abstract level of modeling, recurrent neural networks (RNNs) are high-dimensional models of computation, which have become increasingly popular in systems neuroscience research [35]. Typically, RNNs are trained to do a

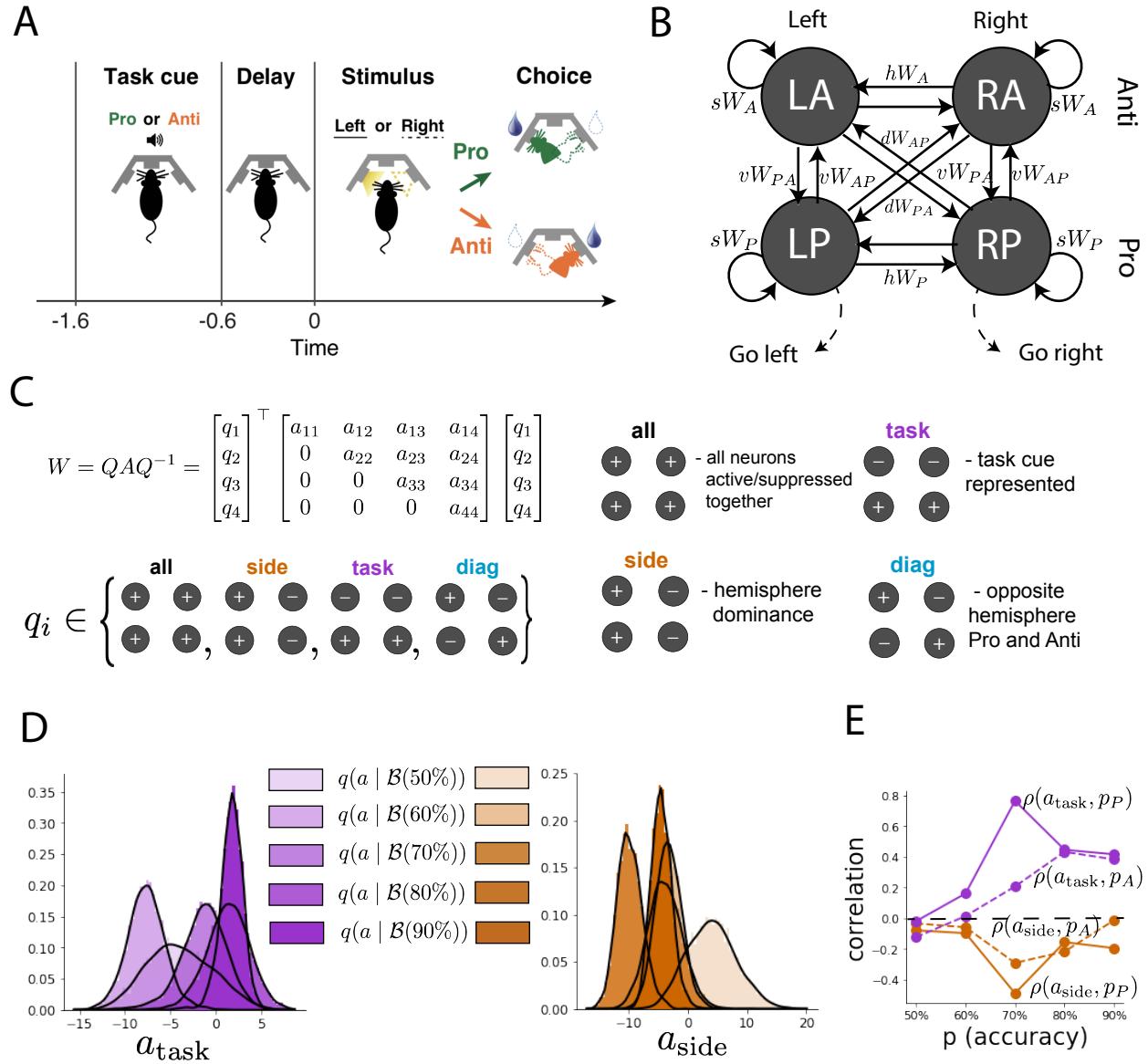


Figure 4: EPI reveals changes in SC [21] connectivity that result in greater task accuracy. A. Rapid task switching behavioral paradigm. In the Pro (Anti) condition indicated by an auditory cue, rats respond by poking into a side port to the same (opposite) side as the light stimulus that is provided after a delay to receive a reward. 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  $A$ . The invariant Schur eigenmodes (symmetry of  $W$ ) are labeled by their hypothesized role in computation:  $q_{\text{all}}$ ,  $q_{\text{task}}$ ,  $q_{\text{side}}$ , and  $q_{\text{diag}}$ . The values of  $A$  are what change for different realizations of  $W$ . 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.

279 task from a systems neuroscience experiment, and then the unit activations of the trained RNN  
 280 are compared to recorded neural activity. A monumental challenge for this line of work is to link  
 281 findings at this level of abstraction with interpretable biophysical mechanisms in the brain.

282 Importantly, recent work establishes such a link between RNN connectivity weights and the re-  
 283 sulting dynamical responses of the network using dynamic mean field theory (DMFT) for neural  
 284 networks [3]. Specifically, DMFT describes the properties of activity in infinite-size neural networks  
 285 given a distribution on the connectivity weights. This theory has been extended from random neu-  
 286 ral networks to low rank RNNs, which have low-dimensional parameterizations of RNN connectivity  
 287 via the pairwise correlations of the low-rank vectors (i.e. the low-rank “geometry”) [22]. For ex-  
 288 ample, connectivity of a rank-1 RNN is the sum of a random component with strength determined  
 289 by  $g$  and a structured component determined by the outer product of vectors  $m$  and  $n$ :

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

290 where  $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ . The entries of  $m$  and  $n$  are drawn from gaussian distributions  $m_i \sim \mathcal{N}(M_m, 1)$   
 291 and  $n_i \sim \mathcal{N}(M_n, 1)$ , whose parameters  $M_m$  and  $M_n$  determine their degree of correlation.

292 Mastrogiuseppe et al. use this theory to design low-rank connectivities via the pairwise correlations  
 293 of such low-rank vectors that solve tasks from behavioral neuroscience. If we consider this DMFT as  
 294 a black-box that takes in a low-rank parameterization and outputs task-relevant response variables  
 295 (e.g. average network activity  $\mu$ , the temporal variability in the network  $\Delta_T$ , or network activity  
 296 along a given dimension  $\kappa$ ), we can combine it with EPI to learn distributions on such connectivity  
 297 parameters that solve tasks via an emergent property defined on the DMFT-produced task-relevant  
 298 responses.

299 We combined this theory with EPI to characterize the mechanistic sources of bias in a rank-1 RNN  
 300 executing approximate Bayesian inference.

301 The rank-1 RNNs were to produce the posterior mean  $\mu_{\text{post}}$  and variance  $\sigma_{\text{post}}^2$  in their mean activity  
 302  $\mu$  and temporal variance  $\Delta_T$  (see Section A.2.4). The Bayesian inference problem was to produce  
 303 the gaussian posterior of the mean  $\mu_y$  of a gaussian likelihood of observations  $y \sim \mathcal{N}(\mu_y, 1)$  given  
 304 a single observation of  $y = 2$  and a prior of  $\mu_y \sim \mathcal{N}(4, 1)$  (Fig. 4A). The true posterior to this  
 305 problem is  $\mu_y \sim \mathcal{N}(\mu_{\text{post}} = 3, \sigma_{\text{post}}^2 = 0.5)$ , although different parameterizations of the connectivity  
 306  $z = [g \ M_m \ M_n]^\top$  result in approximate inference procedures of varying biases in  $\mu_{\text{post}}$  and  $\sigma_{\text{post}}^2$ .

307 We ran EPI on rank-1 RNNs solving this Bayesian inference problem, while allowing a substantial  
 308 amount of variability in the second moment constraints of the network mean  $\mu$  and temporal

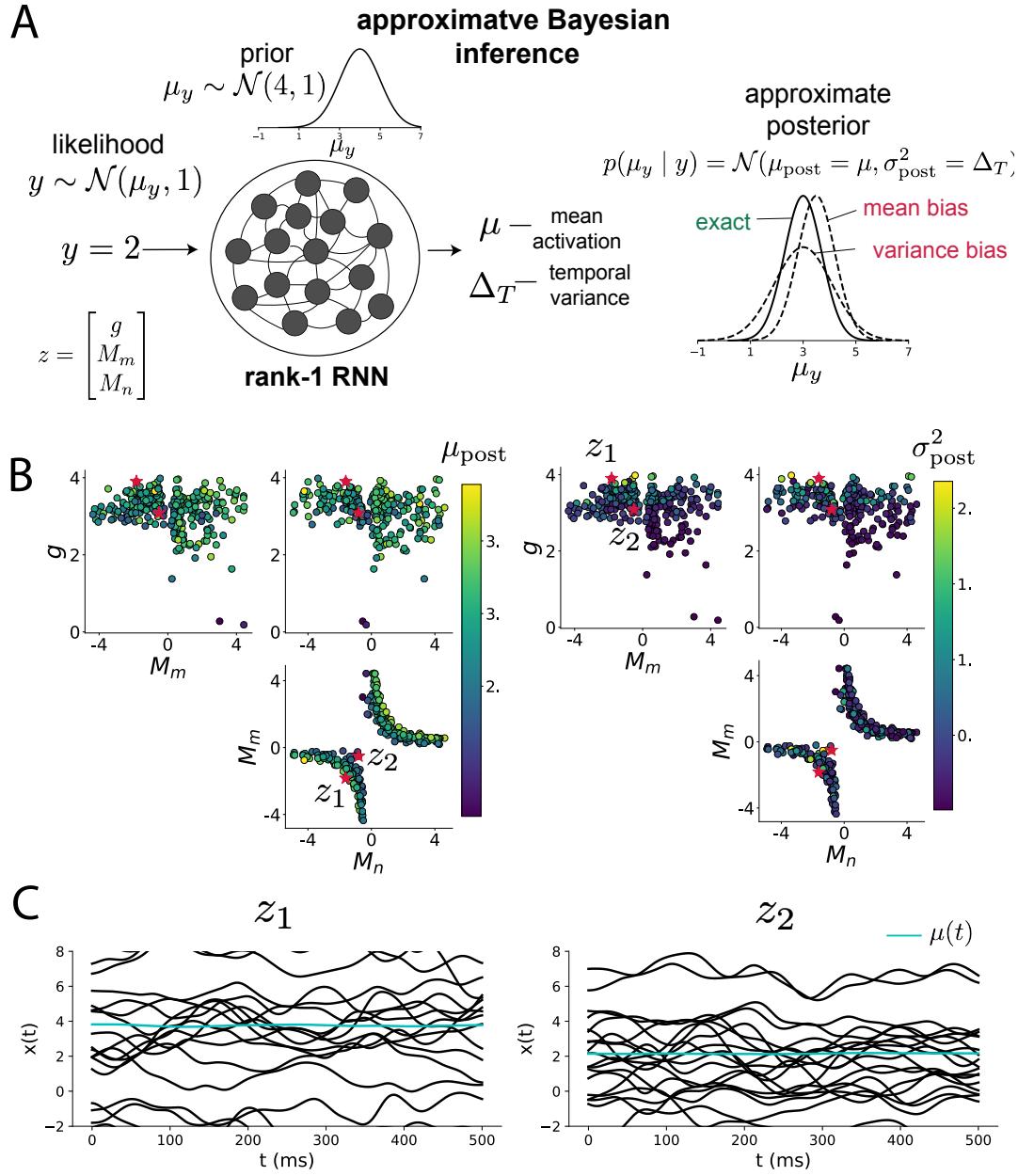


Figure 5: Mechanistic sources of bias in Bayesian inference in rank-1 RNNs. A. (left) A rank-1 RNN running approximate Bayesian inference on  $\mu_y$  assuming a gaussian likelihood variance of 1 and a prior of  $\mathcal{N}(4, 1)$ . (center) The rank-1 RNN represents the computed gaussian posterior mean  $\mu_{\text{post}}$  and variance  $\sigma_{\text{post}}^2$  in its mean activity  $\mu$  and its temporal variance  $\Delta_T$ . (right) Bias in this computation can come from over- or under-estimating the posterior mean or variance. B. Distribution of rank-1 RNNs executing approximate Bayesian inference. Samples are colored by (left) posterior mean  $\mu_{\text{post}} = \mu$  and (right) posterior variance  $\sigma_{\text{post}}^2 = \Delta_T$ . C. Finite size realizations agree with the DMFT theory.

309 variance  $\Delta_T$ . This allowed us to study the mechanistic sources of bias in the sampled rank-1 RNNs  
 310 executing the Bayesian inference computation inexactly. The posterior distribution was roughly  
 311 symmetric in the  $M_m$ - $M_n$  plane with structure suggesting there is a degeneracy in the product of  
 312  $M_m$  and  $M_n$  (Fig. 5B). The product of  $M_m$  and  $M_n$  almost completely determines the posterior  
 313 mean (Fig. 4B, left), and the random strength  $g$  is the most influential variable on the temporal  
 314 variance (Fig. 4B, right). Neither of these observations were obvious from the consistency equations  
 315 afforded by DMFT, the solvers of which we took gradients through to run EPI (see Section A.2.4).

316 Due diligence when working with DMFT requires checking that finite-size realizations of these  
 317 infinite-size networks match the predictions from DMFT. A 2,000-neuron network with param-  
 318 eters  $z_1$  produced an overestimate of the posterior mean and variance (Fig. 5C, left), while a  
 319 2,000-neuron network with parameter  $z_2$  produced underestimates (Fig. 4C, right) as we'd expect  
 320 from the conclusions of the previous paragraph. This novel procedure of doing inference in inter-  
 321 pretable parameterizations of RNNs conditioned on abstract cognitive tasks can be generally used  
 322 for modeling other tasks like noisy integration and context-dependent decision making (Fig. S1).

## 323 4 Discussion

### 324 4.1 EPI is a general tool for theoretical neuroscience.

325 Models of biological systems often have complex nonlinear differential equations, making traditional  
 326 statistical inference intractable. In contrast, EPI is capable of learning distributions of parameters  
 327 in such models producing measurable signatures of computation. We have demonstrated its utility  
 328 on biological models (STG), intermediate-level models of interacting genetically- and functionally-  
 329 defined neuron-types (V1, SC), and the most abstract of models (RNNs). We are able to condi-  
 330 tion both deterministic and stochastic models on low-level emergent properties like firing rates of  
 331 membrane potentials, as well as high-level cognitive function like approximate Bayesian inference.  
 332 Technically, EPI is tractable when the emergent property statistics are continuously differentiable  
 333 with respect to the model parameters, which is very often the case; this emphasizes the general  
 334 utility of EPI.

335 In this study, we have focused on applying EPI to low dimensional parameter spaces of models  
 336 with low dimensional dynamical state. These choices were made to present the reader with a series  
 337 of interpretable conclusions, which is more challenging in high dimensional spaces. In fact, EPI  
 338 should scale reasonably to high dimensional parameter spaces, as the underlying technology has

339 produced state-of-the-art performance on high-dimensional tasks such as texture generation [17].  
340 Of course, increasing the dimensionality of the dynamical state of the model makes optimization  
341 more expensive, and there is a practical limit there as with any machine learning approach. For  
342 systems with high dimensional state, we recommend using theoretical approaches (e.g. [22]) to  
343 reason about reduced parameterizations of such high-dimensional systems.

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

## 355 4.2 Novel hypotheses from EPI

356 Machine learning has played an effective, multifaceted role in neuroscientific progress. Primarily,  
357 it has revealed structure in large-scale neural datasets [37, 38, 39, 40, 41, 42] (see review, [14]).  
358 Secondarily, trained algorithms of varying degrees of biological relevance are beginning to be viewed  
359 as fully-observable computational systems comparable to the brain [43, 44]. Theorists can use deep  
360 learning for probabilistic inference to understand their models and their behavior.

361 For example, consider the fact that we do not yet understand just a four-dimensional, deterministic  
362 model of V1 [20]. This should not be surprising, since analytic approaches to studying nonlinear  
363 dynamical systems become increasingly complicated when stepping from two-dimensional to three-  
364 or four-dimensional systems in the absence of restrictive simplifying assumptions [45]. We promote  
365 the recognition of analytic difficulty, and alternatively the use of EPI to gain the desired model  
366 insights. In Section 3.3, we showed that EPI was far more informative about neuron-type input  
367 responsibility than the predictions afforded through analysis. By flexibly conditioning this V1 model  
368 on different emergent properties, we performed an exploratory analysis of a *model* rather than a  
369 dataset, which generated and proved out a set of testable predictions.

370 Exploratory analyses can be directed. For example, when interested in model changes during learning,  
371 one can use EPI to condition on various levels of an emergent property statistic indicative of  
372 performance like task accuracy in a behavioral paradigm (see Section 3.4). This analysis iden-  
373 tified experimentally testable predictions (proved out *in-silico*) of changes in connectivity in SC  
374 throughout learning of a rapid task switching behavior. Precisely, we predict an initial reduction  
375 in side mode eigenvalue, and a steady increase in task mode eigenvalue in the effective connectivity  
376 matrices of learning rats.

377 In our final analysis, we present a novel procedure for doing statistical inference on interpretable  
378 parameterizations of RNNs executing tasks from behavioral neuroscience. This methodology relies  
379 on recently extended theory of responses in random neural networks with minimal structure [22].  
380 With this methodology, we can finally open the probabilistic model selection toolkit reasoning  
381 about the connectivity of RNNs solving tasks.

## 382 References

- 383 [1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 384 [2] John J Hopfield. Neurons with graded response have collective computational properties like  
385 those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092,  
386 1984.
- 387 [3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural  
388 networks. *Physical review letters*, 61(3):259, 1988.
- 389 [4] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-  
390 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,  
391 17(11):4382–4388, 1997.
- 392 [5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Confer-  
393 ence on Learning Representations*, 2014.
- 394 [6] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation  
395 and variational inference in deep latent gaussian models. *International Conference on Machine  
396 Learning*, 2014.

- [7] 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.
- [8] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- [9] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders. *bioRxiv*, page 399246, 2018.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- [16] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 2015.

- 425 [17] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow  
426 networks. *International Conference on Learning Representations*, 2017.
- 427 [18] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-  
428 free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–  
429 5533, 2017.
- 430 [19] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an  
431 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.  
432 *Neuron*, 77(5):845–858, 2013.
- 433 [20] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and vi-  
434 sual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,  
435 115(3):1399–1409, 2016.
- 436 [21] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J  
437 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor  
438 routing. *bioRxiv*, page 245613, 2018.
- 439 [22] Francesca Mastrogiovanni and Srdjan Ostojic. Linking connectivity, dynamics, and computa-  
440 tions in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- 441 [23] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-  
442 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience  
443 Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 444 [24] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,  
445 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with  
446 degenerate solution networks. *Bernstein Conference*, 2019.
- 447 [25] Jan-Matthis Lueckmann, Pedro Goncalves, Chaitanya Chintaluri, William F Podlaski, Gia-  
448 como Bassetto, Tim P Vogels, and Jakob H Macke. Amortised inference for mechanistic models  
449 of neural dynamics. In *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon,  
450 Portugal*, 2019.
- 451 [26] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-  
452 tion. *Neural Networks*, 15(4-6):479–493, 2002.

- 453 [27] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,  
454 1957.
- 455 [28] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:  
456 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 457 [29] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise  
458 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 459 [30] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of  
460 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 461 [31] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory  
462 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–  
463 592, 2009.
- 464 [32] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear  
465 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,  
466 85(2):402–417, 2015.
- 467 [33] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and  
468 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,  
469 5(10):793, 2004.
- 470 [34] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of  
471 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-*  
472 *biology*, 71(1):45–61, 2011.
- 473 [35] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*  
474 *opinion in neurobiology*, 46:1–6, 2017.
- 475 [36] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean,  
476 Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A  
477 system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems*  
478 *Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.
- 479 [37] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,  
480 13(8):1713–1720, 2001.

- [38] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–7425, 1998.
- [39] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- [40] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.
- [41] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.
- [42] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- [43] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.
- [44] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 2019.
- [45] Steven H Strogatz. Nonlinear dynamics and chaos: with applications to physics, *Biology, Chemistry, and Engineering (Studies in Nonlinearity)*, Perseus, Cambridge, UK, 1994.
- [46] Niru Maheswaranathan, Alex H Williams, Matthew D Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *arXiv preprint arXiv:1907.08549*, 2019.
- [47] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.
- [48] Kenji Doya. Universality of fully connected recurrent neural networks. *Dept. of Biology, UCSD, Tech. Rep*, 1993.

- 510 [49] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial  
511 Intelligence and Statistics*, pages 814–822, 2014.
- 512 [50] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and  
513 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 514 [51] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
515 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 516 [52] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-  
517 bition of inhibition in visual cortex: the logic of connections between molecularly distinct  
518 interneurons. *Nature neuroscience*, 16(8):1068, 2013.

519 **A Methods**

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

521 **Draft in progress:**

522 Emergent property inference (EPI) learn distributions of theoretical model parameters that produce  
523 emergent properties of interest. EPI combines ideas from likelihood-free variational inference [18]  
524 and maximum entropy flow networks [17]. A maximum entropy flow network is used as a deep  
525 probability distribution for the parameters, while these samples often parameterize a differentiable  
526 model simulator, which may lack a tractable likelihood function.

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

$$\mathcal{B} : E_{z \sim q_\theta} [E_{x \sim p(x|z)} [T(x)]] = \mu \quad (9)$$

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

535 In deep probability distributions, a simple random variable  $w \sim p_0$  is mapped deterministically  
536 via a function  $f_\theta$  parameterized by a neural network to the support of the distribution of interest

537 where  $z = f_\theta(\omega) = f_l(\dots f_1(\omega))$ . Given a theoretical model  $p(x | z)$  and some behavior of interest  
 538  $\mathcal{B}$ , the deep probability distributions are trained by optimizing the neural network parameters  $\theta$  to  
 539 find a good approximation  $q_\theta^*$  within the deep variational family  $Q$  to  $p(z | \mathcal{B})$ .

540 In most settings (especially those relevant to theoretical neuroscience) the likelihood of the behavior  
 541 with respect to the model parameters  $p(T(x) | z)$  is unknown or intractable, requiring an alternative  
 542 to stochastic gradient variational bayes [5] or black box variational inference[49]. These types  
 543 of methods called likelihood-free variational inference (LFVI, [18]) skate around the intractable  
 544 likelihood function in situations where there is a differentiable simulator. Akin to LFVI, DSNs are  
 545 optimized with the following objective for a given theoretical model, emergent property statistics  
 546  $T(x)$ , and emergent property constraints  $\mu$ :

$$\begin{aligned} q_\theta^*(z) &= \underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta(z)) \\ \text{s.t. } E_{z \sim q_\theta} [E_{x \sim p(x|z)} [T(x)]] &= \mu \end{aligned} \tag{10}$$

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

548 **Draft in progress:**

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

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

550 that produce a band of oscillations. To do EPI with the dynamics matrix elements as the free  
 551 parameters  $z = [a_1, a_2, a_3, a_4]$ , and fixing  $\tau = 1$ , such that the posterior yields a band of oscillations,  
 552 the emergent property statistics  $T(x)$  are chosen to contain the first- and second-moments of the  
 553 oscillatory frequency  $\Omega$  and the growth/decay factor  $d$  of the oscillating system (the real part of the  
 554 complex conjugate pairs of eigenvalues). To learn the distribution of real entries of  $A$  that yield a  
 555 distribution of  $d$  with mean zero with variance 1, and oscillation frequency  $\Omega$  with mean 1 Hz with  
 556 variance 1, the emergent property values would be set to:

$$\mu = E \begin{bmatrix} d \\ \Omega \\ (d - 0)^2 \\ (\Omega - 1)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 1.0 \\ 1.0 \\ 1.025 \end{bmatrix} \tag{11}$$

557 To obtain a differentiable estimate of the oscillation frequency with respect to the dynamics matrices,  
 558 we could simulate system activity  $x$  from  $z = A$  for some finite number of time steps, and estimate  
 559  $\Omega$  by e.g. taking the peak of the discrete Fourier transform. Instead, the emergent property  
 560 statistics for this oscillating behavior are computed through a closed form function  $g(z)$  by taking  
 561 the eigendecomposition of the dynamics matrix

$$g(z) = E_{x \sim p(x|z)} [T(x)] = \begin{bmatrix} \text{real}(\lambda_1) \\ \frac{\text{imag}(\lambda_1)}{2\pi} \\ \text{real}(\lambda_1)^2 \\ (\frac{\text{imag}(\lambda_1)}{2\pi})^2 \end{bmatrix} \quad (12)$$

562

$$\lambda = \frac{(\frac{a_1+a_4}{\tau}) \pm \sqrt{(\frac{a_1+a_4}{\tau})^2 + 4(\frac{a_2a_3-a_1a_4}{\tau})}}{2} \quad (13)$$

563 where  $\lambda_1$  is the eigenvalue of  $\frac{1}{\tau}A$  with greatest real part. Even though  $E_{x \sim p(x|z)} [T(x)]$  is calculable  
 564 directly via  $g(z)$ , we cannot derive the distribution  $q_\theta^*$ , since the backward mapping from the mean  
 565 parameters  $\mu$  to the natural parameters  $\eta$  of his exponential family is unknown [50]. Instead, we  
 566 can use EPI to learn the linear system parameters producing such a band of oscillations (Fig. S2B).

567 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the  
 568 parameter distribution. The contours of the probability density can be derived from the emergent  
 569 property statistics and values (Fig. S3). In the  $a_1 - a_4$  plane, is a black line at  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ ,  
 570 a dotted black line at the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 1$ , and a grey line at twice the  
 571 standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 2$  (Fig. S3A). Here the lines denote the set of solutions at  
 572 fixed behaviors, which overlay the posterior obtained through EPI. The learned DSN distribution  
 573 precisely reflects the desired statistical constraints and model degeneracy in the sum of  $a_1$  and  
 574  $a_4$ . Intuitively, the parameters equivalent with respect to emergent property statistic  $\text{real}(\lambda_1)$  have  
 575 similar log densities.

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4-a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

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

$$\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 \quad (15)$$

580 Since we constrained  $E_{q_\theta}[\text{imag}(\lambda)] = 2\pi$  (with  $\omega = 1$ ), we can plot contours of the equation  
 581  $\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 = (2\pi)^2$  for various  $a_1 a_4$  (Fig. S3A). If  $\sigma_{1,4} = E_{q_\theta}(|a_1 a_4 - E_{q_\theta}[a_1 a_4]|)$ ,  
 582 then we plot the contours as  $a_1 a_4 = 0$  (black),  $a_1 a_4 = -\sigma_{1,4}$  (black dotted), and  $a_1 a_4 = -2\sigma_{1,4}$  (grey  
 583 dotted) (Fig. S3B). We take steps in negative standard deviation of  $a_1 a_4$  (dotted and gray lines),  
 584 since there are few positive values  $a_1 a_4$  in the posterior. More subtle model-behavior combinations  
 585 will have even more complexity, further motivating the use of EPI for understanding these systems.  
 586 For futher validation of the underlying technology, see recovery of ground truth distributions with  
 587 maximum entorpy flow networks [17].

588 **A.1.2 Augmented Lagrangian optimization**

589 **Draft in progress:**

590 To optimize  $q_\theta(z)$  in equation 1, the constrained optimization is performed using the augmented  
 591 Lagrangian method. The following objective is minimized:

$$L(\theta; \alpha, c) = -H(q_\theta) + \alpha^\top \delta(\theta) + \frac{c}{2} \|\delta(\theta)\|^2 \quad (16)$$

592 where  $\delta(\theta) = E_{z \sim q_\theta} [E_{x \sim p(x|z)} [T(x) - \mu]]$ ,  $\alpha \in \mathcal{R}^m$  are the Lagrange multipliers and  $c$  is the  
 593 penalty coefficient. For a fixed  $(\alpha, c)$ ,  $\theta$  is optimized with stochastic gradient descent. A low  
 594 value of  $c$  is used initially, and increased during each augmented Lagrangian epoch – a period  
 595 of optimization with fixed *alpha* and  $c$  for a given number of stochastic optimziation iterations.  
 596 Similarly,  $\alpha$  is tuned each epoch based on the constraint violations. For the linear 2-dimensional  
 597 system (Fig. S4C) optimization hyperparameters are initialized to  $c_1 = 10^{-4}$  and  $\alpha_1 = 0$ . The  
 598 penalty coefficient is updated based on a hypothesis test regarding the reduction in constraint  
 599 violation. The p-value of  $E[\|\delta(\theta_{k+1})\|] > \gamma E[\|\delta(\theta_k)\|]$  is computed, and  $c_{k+1}$  is updated to  $\beta c_k$   
 600 with probability  $1 - p$ . Throughout the project,  $\beta = 4.0$  and  $\gamma = 0.25$  is used. The other update  
 601 rule is  $\alpha_{k+1} = \alpha_k + c_k \frac{1}{n} \sum_{i=1}^n (T(x^{(i)}) - \mu)$ . In this example, each augmented Lagrangian epoch ran  
 602 for 5,000 iterations. We consider the optimization to have converged when a null hypothesis test  
 603 of constraint violations being zero is accepted for all constraints at a significance threshold 0.05.  
 604 This is the dotted line on the plots below depicting the optimization cutoff of EPI optimization for  
 605 the 2-dimensional linear system. If the optimization is left to continue running, entropy usually  
 606 decreases, and structural pathologies in the distribution may be introduced.

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

613 **A.1.3 Normalizing flows**

614 **Draft in progress:**

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

$$H(q_\theta(z)) = \int -q_\theta(z) \log(q_\theta(z)) dz = E_{z \sim q_\theta} [-\log(q_\theta(z))] = E_{\omega \sim q_0} [-\log(q_\theta(f_\theta(\omega)))] \quad (17)$$

$$\nabla_\theta H(q_\theta(z)) = E_{\omega \sim q_0} [-\nabla_\theta \log(q_\theta(f_\theta(\omega)))] \quad (18)$$

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

$$q(z') = q(f^{-1}(z')) \left| \det \frac{\partial f^{-1}(z')}{\partial z'} \right| = q(z) \left| \det \frac{\partial f(z)}{\partial z} \right|^{-1} \quad (19)$$

622 However, this computation has cubic complexity in dimensionality for fully connected layers. By  
 623 restricting our layers to normalizing flows [16] – bijective functions with fast log determinant  
 624 jacobian computations, we can tractably optimize deep generative models with objectives that are  
 625 a function of sample density, like entropy. Most of our analyses use real NVP [51], which have  
 626 proven effective in our architecture searches, and have the advantageous features of fast sampling  
 627 and fast density evaluation.

628 **A.1.4 Related work**

629 **Draft in progress:**

630

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

632 **Draft in progress:**

633 Consider the goal of doing variational inference (VI) in with an exponential family posterior dis-  
 634 tribution  $p(z | x)$ . We'll use the following abbreviated notation to collect the base measure and  
 635 sufficient statistics into  $\tilde{T}(z)$  and likewise concatenate a 1 onto the end of the natural parameter  
 636  $\tilde{\eta}(x)$ . The log normalizing constant  $A(\eta(x))$  will remain unchanged.

$$\begin{aligned} p(z | x) &= b(z) \exp \left( \eta(x)^\top T(z) - A(\eta(x)) \right) = \exp \left( \begin{bmatrix} \eta(x) \\ 1 \end{bmatrix}^\top \begin{bmatrix} T(z) \\ b(z) \end{bmatrix} - A(\eta(x)) \right) \\ &= \exp \left( \tilde{\eta}(x)^\top \tilde{T}(z) - A(\eta(x)) \right) \end{aligned} \quad (20)$$

637 VI looks with an exponential family posterior distribution uses optimization to minimize the fol-  
 638 lowing divergence [15]:

$$q_\theta^* = \underset{q_\theta \in Q}{\operatorname{argmin}} KL(q_\theta || p(z | x)) \quad (21)$$

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

$$KL(q_\theta || p(z | x)) = E_{z \sim q_\theta} [\log(q_\theta(z))] - E_{z \sim q_\theta} [\log(p(z | x))] \quad (22)$$

$$\begin{aligned} &= -H(q_\theta) - E_{z \sim q_\theta} [\tilde{\eta}(x)^\top \tilde{T}(z) - A(\eta(x))] \end{aligned} \quad (23)$$

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

$$\underset{q_\theta \in Q}{\operatorname{argmin}} KL(q_\theta || p(z | x)) = \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - E_{z \sim q_\theta} [\tilde{\eta}(x)^\top \tilde{T}(z)] \quad (24)$$

644 Further, we can write the objective in terms of the first moment of the sufficient statistics  $\mu =$   
 645  $E_{z \sim p(z|x)} [T(z)]$ .

$$= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - E_{z \sim q_\theta} [\tilde{\eta}(x)^\top (\tilde{T}(z) - \mu)] + \tilde{\eta}(x)^\top \mu \quad (25)$$

$$= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - E_{z \sim q_\theta} [\tilde{\eta}(x)^\top (\tilde{T}(z) - \mu)] \quad (26)$$

647 In emergent property inference (EPI), we're solving the following problem.

$$q_\theta^*(z)y = \underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta(z)), \text{ s.t. } E_{z \sim q_\theta} [E_{x \sim p(x|z)} [T(x)]] = \mu \quad (27)$$

648 The lagrangian objective is

$$q_\theta^* = \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \alpha^\top (E_{z \sim q_\theta} [\tilde{T}(z)] - \mu) \quad (28)$$

649 As the lagrangian optimization proceeds,  $\alpha$  should converge to  $\tilde{\eta}(x)$ ) through its adaptations in  
 650 each epoch. More formally,  $\tilde{\eta}(x) \leftrightarrow \tilde{\eta}(\mu)$  is referred to as the backward mapping, and is formally  
 651 hard to identify [50]. Since this backward mapping is deterministic, conceptually, we can replace  
 652  $p(z | x)$  with  $p(z | \mu)$ . More commonly, we write  $p(z | \mathcal{B})$  for clarity where  $\mathcal{B}$  more explicitly  
 653 captures the moment constraints of the sufficient statistics.

654 **A.2 Theoretical models**

655 **Draft in progress:**

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

658 **A.2.1 Stomatogastric ganglion**

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

$$C_m \frac{\partial x_m}{\partial t} = -[h_{leak} + h_{Ca} + h_K + h_{hyp} + h_{elec} + h_{syn}] \quad (29)$$

660 The membrane potential of each neuron is affected by the leak, calcium, potassium, hyperpolar-  
 661 ization, electrical and synaptic currents, respectively. The capacitance of the cell membrane was  
 662 set to  $C_m = 1nF$ . Each current is a function of the the neuron's membrane potential  $x_m$  and  
 663 the parameters of the circuit such as  $g_{el}$  and  $g_{syn}$ , whose affect on the circuit is considered in the  
 664 motivational example of EPI in Fig. 1. Specifically, the currents are the difference in the neuron's  
 665 membrane potential and that current type's reversal potential multiplied by a conductance:

$$h_{leak} = g_{leak}(x_m - V_{leak}) \quad (30)$$

$$h_{elec} = g_{el}(x_m^{post} - x_m^{pre}) \quad (31)$$

$$h_{syn} = g_{syn}S_\infty^{pre}(x_m^{post} - V_{syn}) \quad (32)$$

$$h_{Ca} = g_{Ca}M_\infty(x_m - V_{Ca}) \quad (33)$$

$$h_K = g_KN(x_m - V_K) \quad (34)$$

$$h_{hyp} = g_hH(x_m - V_{hyp}) \quad (35)$$

671 The reversal potentials were set to  $V_{leak} = -40mV$ ,  $V_{Ca} = 100mV$ ,  $V_K = -80mV$ ,  $V_{hyp} = -20mV$ ,  
 672 and  $V_{syn} = -75mV$ . The other conductance parameters were fixed to  $g_{leak} = 1 \times 10^{-4}\mu S$ .  $g_{Ca}$ ,

673  $g_K$ , and  $g_{hyp}$  had different values based on fast, intermediate (hub) or slow neuron. Fast:  $g_{Ca} =$   
 674  $1.9 \times 10^{-2}$ ,  $g_K = 3.9 \times 10^{-2}$ , and  $g_{hyp} = 2.5 \times 10^{-2}$ . Intermediate:  $g_{Ca} = 1.7 \times 10^{-2}$ ,  
 675  $g_K = 1.9 \times 10^{-2}$ , and  $g_{hyp} = 8.0 \times 10^{-3}$ . Intermediate:  $g_{Ca} = 8.5 \times 10^{-3}$ ,  $g_K = 1.5 \times 10^{-2}$ , and  $g_{hyp} = 1.0 \times 10^{-2}$ .

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

$$M_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_m - v_1}{v_2} \right) \right) \quad (36)$$

$$\frac{\partial N}{\partial t} = \lambda_N (N_\infty - N) \quad (37)$$

$$N_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_m - v_3}{v_4} \right) \right) \quad (38)$$

$$\lambda_N = \phi_N \cosh \left( \frac{x_m - v_3}{2v_4} \right) \quad (39)$$

$$\frac{\partial H}{\partial t} = \frac{(H_\infty - H)}{\tau_h} \quad (40)$$

$$H_\infty = \frac{1}{1 + \exp \left( \frac{x_m + v_5}{v_6} \right)} \quad (41)$$

$$\tau_h = 272 - \left( \frac{-1499}{1 + \exp \left( \frac{-x_m + v_7}{v_8} \right)} \right) \quad (42)$$

684 where we set  $v_1 = 0mV$ ,  $v_2 = 20mV$ ,  $v_3 = 0mV$ ,  $v_4 = 15mV$ ,  $v_5 = 78.3mV$ ,  $v_6 = 10.5mV$ ,  
 685  $v_7 = -42.2mV$ ,  $v_8 = 87.3mV$ ,  $v_9 = 5mV$ , and  $v_{th} = -25mV$ . These are the same parameter  
 686 values used in [19].

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

$$S_\infty = \frac{1}{1 + \exp \left( \frac{v_{th} - x_m}{v_9} \right)} \quad (43)$$

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

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

694 Our original approach to measuring frequency was to take the max of the fast Fourier transform  
 695 (FFT) of the simulated time series. There are a few key considerations here. One is resolution

in frequency space. Each FFT entry will correspond to a signal frequency of  $\frac{F_s k}{N}$ , where  $N$  is the number of samples used for the FFT,  $F_s = \frac{1}{dt}$ , and  $k \in [0, 1, \dots, N - 1]$ . Our resolution is improved by increasing  $N$  and decreasing  $dt$ . Increasing  $N = T - b$ , where  $b$  is some fixed number of buffer burn-in initialization samples, necessitates an increase in simulation time steps  $T$ , which directly increases computational cost. Increasing  $F_s$  (decreasing  $dt$ ) increases system approximation accuracy, but requires more time steps before a full cycle is observed. At the level of  $dt = 0.025$ , thousands of temporal samples were required for resolution of .01Hz. These challenges in frequency resolution with the discrete Fourier transform motivated the use of an alternative basis of complex exponentials. Instead, we used a basis of complex exponentials with frequencies from 0.0-1.0 Hz at 0.01Hz resolution,  $\Phi = [0.0, 0.01, \dots, 1.0]^\top$

Another consideration was that the frequency spectra of the hub neuron has several peaks. This 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  $x_i \in \mathcal{R}^N$ , where  $i \in \{\text{f1, f2, hub, s1, s2}\}$ . The “frequency identification” vector is

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

. The frequency is then calculated as  $\Omega_i = u_i^\top \Phi$  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 0.55Hz, while the second moments are set to

719 0.025Hz<sup>2</sup>.

$$E \begin{bmatrix} \Omega_{f1} \\ \Omega_{f2} \\ \Omega_{hub} \\ \Omega_{s1} \\ \Omega_{s2} \\ (\Omega_{f1} - 0.55)^2 \\ (\Omega_{f2} - 0.55)^2 \\ (\Omega_{hub} - 0.55)^2 \\ (\Omega_{s1} - 0.55)^2 \\ (\Omega_{s2} - 0.55)^2 \end{bmatrix} = \begin{bmatrix} 0.55 \\ 0.55 \\ 0.55 \\ 0.55 \\ 0.55 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \end{bmatrix} \quad (44)$$

720 For EPI in Fig 2C, we used a real NVP architecture with two coupling layers. Each coupling layer  
 721 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  
 722 have shown the EPI optimization that converged with maximum entropy across 2 random seeds  
 723 and augmented Lagrangian coefficient initializations of  $c_0=0, 2$ , and 5.

724 **A.2.2 Primary visual cortex**

725 **Draft in progress:**

726 The dynamics of each neural populations average rate  $x = \begin{bmatrix} x_E \\ x_P \\ x_S \\ x_V \end{bmatrix}$  are given by:

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n \quad (45)$$

727 Some neuron-types largely lack synaptic projections to other neuron-types [52], and it is popular  
 728 to only consider a subset of the effective connectivities [20].

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & 0 \\ W_{PE} & W_{PP} & W_{PS} & 0 \\ W_{SE} & 0 & 0 & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & 0 \end{bmatrix} \quad (46)$$

729 (TODO: ask Ken about how to introduce these values and what to ref). Estimates of the of the  
 730 probability of connection and strength of connection from the Allen institute result in an estimate

<sup>731</sup> of the effective connectivity:

$$W = \begin{bmatrix} 0.0576 & 0.19728 & 0.13144 & 0 \\ 0.58855 & 0.30668 & 0.4285 & 0 \\ 0.15652 & 0 & 0 & 0.2 \\ 0.13755 & 0.0902 & 0.4004 & 0 \end{bmatrix} \quad (47)$$

<sup>732</sup> We look at how this four-dimensional nonlinear dynamical model of V1 responds to different inputs,  
<sup>733</sup> and compare the predictions of the linear response to the approximate posteriors obtained through  
<sup>734</sup> EPI. The input to the system is the sum of a baseline input  $b = [1 \ 1 \ 1 \ 1]^\top$  and a differential  
<sup>735</sup> input  $dh$ :

$$h = b + dh \quad (48)$$

<sup>736</sup> All simulations of this system had  $T = 100$  time points, a time step  $dt = 5\text{ms}$ , and time constant  
<sup>737</sup>  $\tau = 20\text{ms}$ . And the system was initialized to a random draw  $x(0)_i \sim \mathcal{N}(1, 0.01)$ .

<sup>738</sup> We can describe the dynamics of this system more generally by

$$\dot{x}_i = -x_i + f(u_i) \quad (49)$$

<sup>739</sup> where the input to each neuron is

$$u_i = \sum_j W_{ij}x_j + h_i \quad (50)$$

<sup>740</sup> Let  $F_{ij} = \gamma_i \delta(i, j)$ , where  $\gamma_i = f'(u_i)$ . Then, the linear response is

$$\frac{\partial x_{ss}}{\partial h} = F(W \frac{\partial x_{ss}}{\partial h} + I) \quad (51)$$

<sup>741</sup> which is calculable by

$$\frac{\partial x_{ss}}{\partial h} = (F^{-1} - W)^{-1} \quad (52)$$

<sup>742</sup> The emergent property we considered was the first and second moments of the change in rate  $dr$   
<sup>743</sup> between the baseline input  $h = b$  and  $h = b + dh$ . We use the following notation to indicate that  
<sup>744</sup> the emergent property statistics were set to the following values:

$$\mathcal{B}(\alpha, y) \leftrightarrow E \begin{bmatrix} dx_{\alpha,ss} \\ (dx_{\alpha,ss} - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix} \quad (53)$$

<sup>745</sup> 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  
<sup>746</sup> initial seeds using an architecture of four coupling layers, each with two hidden layers of 10 units.

<sup>747</sup> We set  $c_0 = 10^5$ .

748 **A.2.3 Superior colliculus**

749 **Draft in progress:**

750 There are four total units: two in each hemisphere corresponding to the PRO/CONTRA and  
 751 ANTI/IPSI populations. Each unit has an activity ( $x_i$ ) and internal variable ( $u_i$ ) related by

$$x_i(t) = \left( \frac{1}{2} \tanh \left( \frac{u_i(t) - \epsilon}{\zeta} \right) + \frac{1}{2} \right) \quad (54)$$

752  $\epsilon = 0.05$  and  $\zeta = 0.5$  control the position and shape of the nonlinearity, respectively.

753 We can order the elements of  $x_i$  and  $u_i$  into vectors  $x$  and  $u$  with elements

$$x = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad u = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix} \quad (55)$$

754 The internal variables follow dynamics:

$$\tau \frac{\partial u}{\partial t} = -u + Wx + h + \sigma \partial B \quad (56)$$

755 with time constant  $\tau = 0.09s$  and gaussian noise  $\sigma \partial B$  controlled by the magnitude of  $\sigma = 1.0$ . The  
 756 weight matrix has 8 parameters  $sW_P$ ,  $sW_A$ ,  $vW_{PA}$ ,  $vW_{AP}$ ,  $hW_P$ ,  $hW_A$ ,  $dW_{PA}$ , and  $dW_{AP}$  (Fig.  
 757 4B).

$$W = \begin{bmatrix} sW_P & vW_{PA} & hW_P & dW_{PA} \\ vW_{AP} & sW_A & dW_{AP} & hW_A \\ hW_P & dW_{PA} & sW_P & vW_{PA} \\ dW_{AP} & hW_A & vW_{AP} & sW_A \end{bmatrix} \quad (57)$$

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

$$h = h_{\text{rule}} + h_{\text{choice-period}} + h_{\text{light}} \quad (58)$$

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

$$h_{P,\text{rule}}(t) = \begin{cases} I_{P,\text{rule}} \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (59)$$

760

$$h_{A,\text{rule}}(t) = \begin{cases} I_{A,\text{rule}} \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (60)$$

761 a choice-period input,

$$h_{\text{choice}}(t) = \begin{cases} I_{\text{choice}} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (61)$$

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

$$h_{\text{light}}(t) = \begin{cases} I_{\text{light}} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Left} \\ I_{\text{light}} \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Right} \\ 0, & t \leq 1.2s \end{cases} \quad (62)$$

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

764 TODO: this is probably a good place to explain the intuition behind the naming of the Schur  
765 eigenmodes.

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

$$\hat{p}_i = E_{\sigma \partial B} [x_{LP,ss} | s = L, c = P, z_i] = \frac{1}{M} \sum_{j=1}^M x_{LP,ss}(s = L, c = P, z_i, \sigma \partial B_j) \quad (63)$$

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

$$E_{z_i \sim q_\phi} [E_{\sigma \partial B} [x_{LP,ss} | s = L, c = P, z_i]] = E_{z_i \sim q_\phi} [\hat{p}_i] = p_{LP} \quad (64)$$

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

$$E_{z \sim q_\phi} [\sigma_{err}^2] = 0 \quad (65)$$

772

$$\sigma_{err}^2 = Var_{\sigma \partial B} [x_{LP,ss} | s = L, c = P, z_i] - \hat{p}_i(1 - \hat{p}_i) \quad (66)$$

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

$$E_{z \sim q_\phi} [d_P] = 1 \quad (67)$$

775

$$E_{\sigma \partial W} [(x_{LP,ss} - x_{RP,ss})^2 | s = L, c = P, z_i] \quad (68)$$

776 We refer to networks obeying these constraints as Bernoulli, winner-take-all networks. Since the  
 777 maximum variance of a random variable bounded from 0 to 1 is the Bernoulli variance ( $\hat{p}(1 - \hat{p})$ ),  
 778 and the maximum squared difference between two variables bounded from 0 to 1 is 1, we do not  
 779 need to control the second moment of these test statistics. In reality, these variables are dynamical  
 780 system states and can only exponentially decay (or saturate) to 0 (or 1), so the Bernoulli variance  
 781 error and squared difference constraints can only be undershot. This is important to be mindful  
 782 of when evaluating the convergence criteria. Instead of using our usual hypothesis testing criteria  
 783 for convergence to the emergent property, we set a slack variable threshold for these technically  
 784 infeasible constraints to 0.05.

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

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

$$\mathcal{B}(p) \leftrightarrow \begin{bmatrix} \hat{p}_P \\ \hat{p}_A \\ (\hat{p}_P - p)^2 \\ (\hat{p}_A - p)^2 \\ \sigma_{P,err}^2 \\ \sigma_{A,err}^2 \\ d_P \\ d_A \end{bmatrix} = \begin{bmatrix} p \\ p \\ 0.15^2 \\ 0.15^2 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad (69)$$

793 For each accuracy level  $p$ , we ran EPI for 10 different random seeds and selected the maximum  
 794 entropy solution using an architecture of 10 planar flows with  $c_0 = 2$ .

795 **A.2.4 Rank-1 RNN**796 **Draft in progress:**797 The network dynamics of neuron  $i$ 's rate  $x$  evolve according to:

$$\dot{x}_i(t) = -x_i(t) + \sum_{j=1}^N J_{ij}\phi(x_j(t)) + I_i \quad (70)$$

798 where the connectivity is comprised of a random and structured component:

$$J_{ij} = g\chi_{ij} + P_{ij} \quad (71)$$

799 The random all-to-all component has elements drawn from  $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ , and the structured  
800 component is a sum of  $r$  unit rank terms:

$$P_{ij} = \sum_{k=1}^r \frac{m_i^{(k)} n_j^{(k)}}{N} \quad (72)$$

801 We use this theory to compute  $T(x)$  when running EPI.802 Rank-1 vectors  $m$  and  $n$  have elements drawn

$$m_i \sim \mathcal{N}(M_m, \Sigma_m)$$

803

$$n_i \sim \mathcal{N}(M_n, \Sigma_n)$$

804 The current has the following statistics:

$$I = M_I + \frac{\Sigma_{mI}}{\Sigma_m} x_1 + \frac{\Sigma_{nI}}{\Sigma_n} x_2 + \Sigma_\perp h$$

805 where  $x_1$ ,  $x_2$ , and  $h$  are standard normal random variables.806 The  $\ddot{\Delta}$  equation is broken into the equation for  $\Delta_0$  and  $\Delta_\infty$  by the autocorrelation dynamics  
807 assertions.

$$\ddot{\Delta}(\tau) = -\frac{\partial V}{\partial \Delta}$$

808

$$\ddot{\Delta} = \Delta - \{g^2 \langle [\phi_i(t)\phi_i(t+\tau)] \rangle + \Sigma_m^2 \kappa^2 + 2\Sigma_{mI} \kappa + \Sigma_I^2\}$$

809 We can write out the potential function by integrating the negated RHS.

$$V(\Delta, \Delta_0) = \int \mathcal{D}\Delta \frac{\partial V(\Delta, \Delta_0)}{\partial \Delta}$$

810

$$V(\Delta, \Delta_0) = -\frac{\Delta^2}{2} + g^2 \langle [\Phi_i(t)\Phi_i(t+\tau)] \rangle + (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI} \kappa + \Sigma_I^2) \Delta + C$$

811 We assume that as time goes to infinity, the potential relaxes to a steady state.

$$\frac{\partial V(\Delta_\infty, \Delta_0)}{\partial \Delta} = 0$$

$$812 \quad \frac{\partial V(\Delta_\infty, \Delta_0)}{\partial \Delta} = -\Delta + \{g^2 \langle [\phi_i(t)\phi_i(t+\infty)] \rangle + \Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2\} = 0$$

$$813 \quad \Delta_\infty = g^2 \langle [\phi_i(t)\phi_i(t+\infty)] \rangle + \Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2$$

$$814 \quad \Delta_\infty = g^2 \int \mathcal{D}z \left[ \int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty}x + \sqrt{\Delta_\infty}z) \right]^2 + \Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2$$

815 Also, we assume that the energy of the system is preserved throughout the entirety of its evolution.

$$816 \quad V(\Delta_0, \Delta_0) = V(\Delta_\infty, \Delta_0)$$

$$-\frac{\Delta_0^2}{2} + g^2 \langle [\Phi_i(t)\Phi_i(t)] \rangle + (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2) \Delta_0 + C = -\frac{\Delta_\infty^2}{2} + g^2 \langle [\Phi_i(t)\Phi_i(t)] \rangle + (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2) \Delta_\infty + C$$

$$817 \quad \frac{\Delta_0^2 - \Delta_\infty^2}{2} = g^2 (\langle [\Phi_i(t)\Phi_i(t)] \rangle - \langle [\Phi_i(t)\Phi_i(t)] \rangle) + (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)(\Delta_0 - \Delta_\infty)$$

$$818 \quad \frac{\Delta_0^2 - \Delta_\infty^2}{2} = g^2 \left( \int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0}z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty}x + \sqrt{\Delta_\infty}z) \right)$$

$$+ (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)(\Delta_0 - \Delta_\infty)$$

### 819 Consistency equations:

820

$$\mu = F(\mu, \kappa, \Delta_0, \Delta_\infty) = M_m \kappa + M_I$$

$$\kappa = G(\mu, \kappa, \Delta_0, \Delta_\infty) = M_n \langle [\phi_i] \rangle + \Sigma_{nI} \langle [\phi'_i] \rangle$$

$$\frac{\Delta_0^2 - \Delta_\infty^2}{2} = H(\mu, \kappa, \Delta_0, \Delta_\infty) = g^2 \left( \int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0}z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty}x + \sqrt{\Delta_\infty}z) \right)$$

$$+ (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)(\Delta_0 - \Delta_\infty)$$

$$\Delta_\infty = L(\mu, \kappa, \Delta_0, \Delta_\infty) = g^2 \int \mathcal{D}z \left[ \int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty}x + \sqrt{\Delta_\infty}z) \right]^2 + \Sigma_m^2 \kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2$$

$$(73)$$

821 We can solve these equations by simulating the following Langevin dynamical system.

$$x(t) = \frac{\Delta_0(t)^2 - \Delta_\infty(t)^2}{2}$$

$$\Delta_0(t) = \sqrt{2x(t) + \Delta_\infty(t)^2}$$

$$\dot{\mu}(t) = -\mu(t) + F(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t))$$

$$\dot{\kappa}(t) = -\kappa + G(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t))$$

$$\dot{x}(t) = -x(t) + H(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t))$$

$$\dot{\Delta}_\infty(t) = -\Delta_\infty(t) + L(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t))$$

$$(74)$$

<sup>822</sup> Then, the temporal variance is simply calculated via

$$\Delta_T = \Delta_0 - \Delta_\infty \quad (75)$$

<sup>823</sup> TODO Need to explain the warm starting for the aficionados.

<sup>824</sup> TODO explain the density network architectures used.

<sup>825</sup> **A.3 Supplementary Figures**

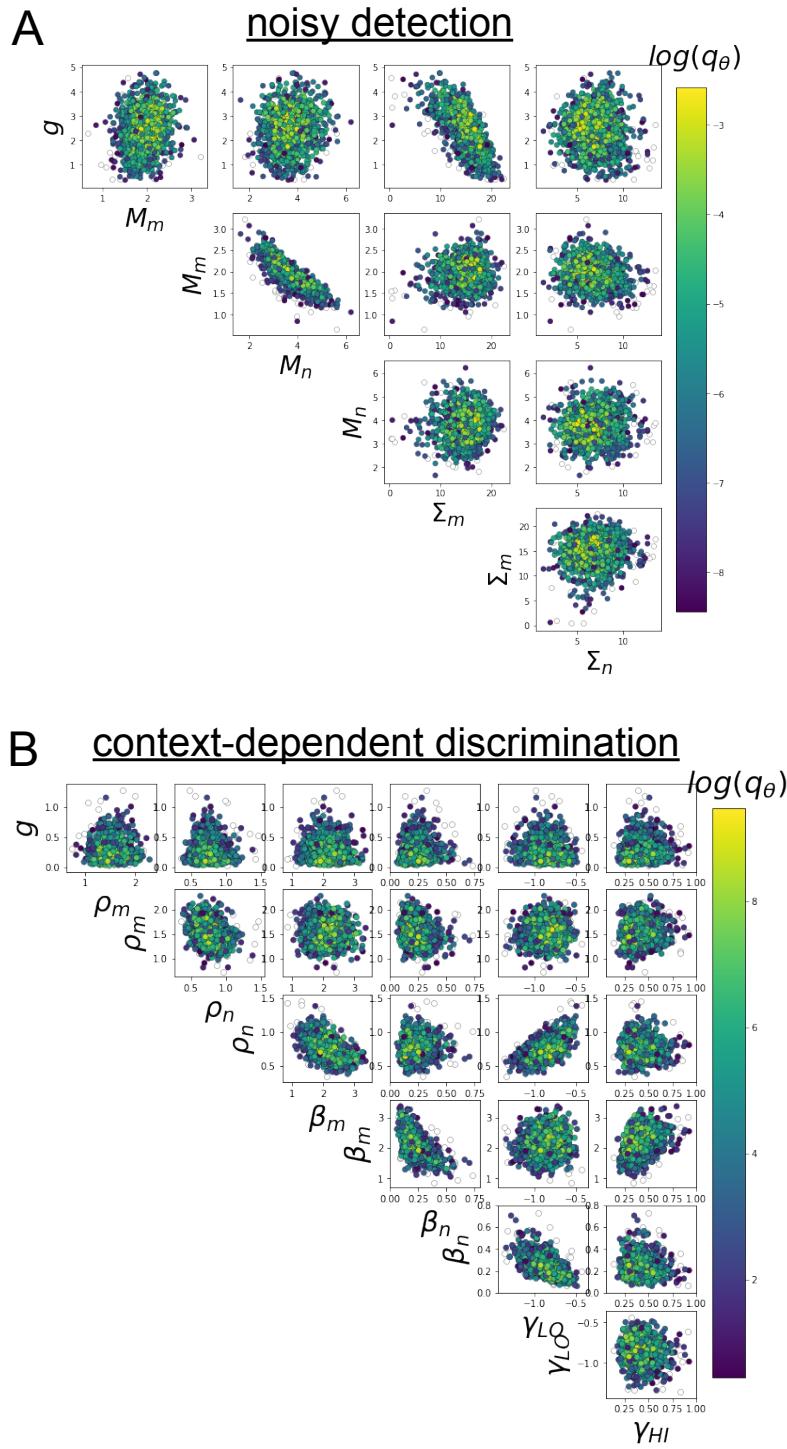


Fig. S1: A. EPI for rank-1 networks doing discrimination. B. EPI for rank-2 networks doing context-dependent discrimination.

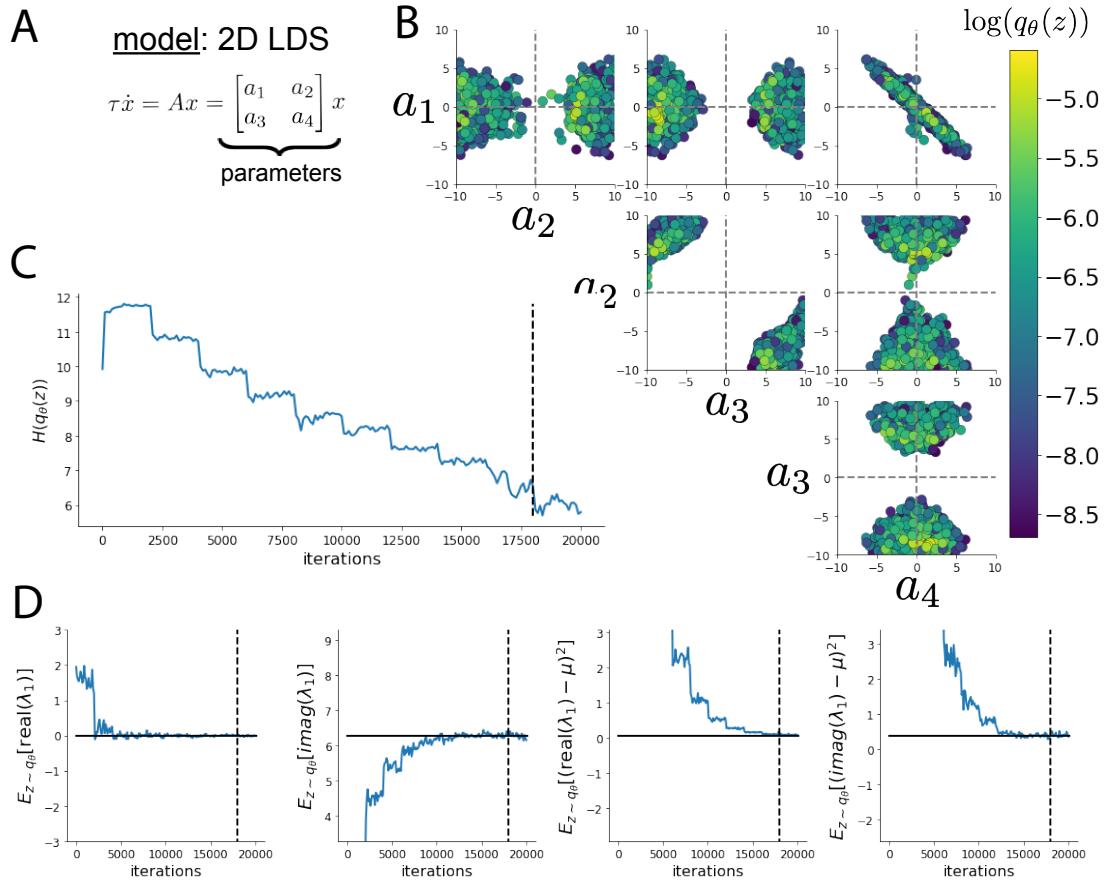


Fig. S2: A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix  $A$  are the parameters. B. The DSN distribution for a 2D LDS with  $\tau = 1$  that produces an average of 1Hz oscillations with some small amount of variance. C. Entropy throughout the optimization. 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.

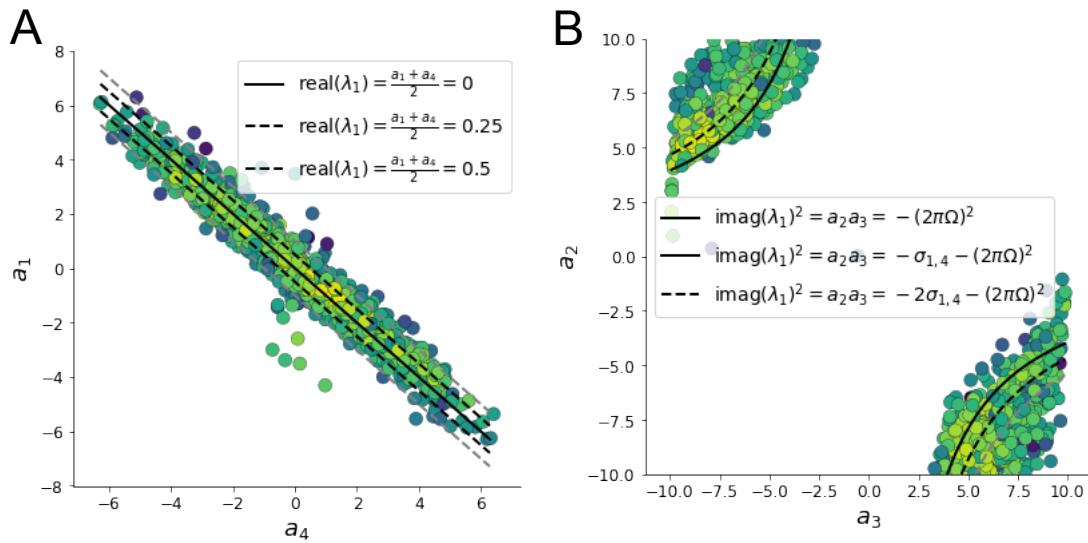


Fig. S3: A. Probability contours in the  $a_1 - a_4$  plane can be derived from the relationship to emergent property statistic of growth/decay factor. B. Probability contours in the  $a_2 - a_3$  plane can be derived from relationship to the emergent property statistic of oscillation frequency.