

# Interrogating theoretical models of neural computation with deep inference

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

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

<sup>2</sup> A cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures  
<sup>3</sup> a hypothesized neural mechanism. Such models are valuable when they give rise to an experimen-  
<sup>4</sup> tally observed phenomenon – whether behavioral or in terms of neural activity – and thus can offer  
<sup>5</sup> insights into neural computation. The operation of these circuits, like all models, critically depends  
<sup>6</sup> on the choices of model parameters. Historically, the gold standard has been to analytically derive  
<sup>7</sup> the relationship between model parameters and computational properties. However, this enterprise  
<sup>8</sup> quickly becomes infeasible as biologically realistic constraints are included into the model increas-  
<sup>9</sup> ing its complexity, often resulting in *ad hoc* approaches to understanding the relationship between  
<sup>10</sup> model and computation. We bring recent machine learning techniques – the use of deep generative  
<sup>11</sup> models for probabilistic inference – to bear on this problem, learning distributions of parameters  
<sup>12</sup> that produce the specified properties of computation. Importantly, the techniques we introduce  
<sup>13</sup> offer a principled means to understand the implications of model parameter choices on compu-  
<sup>14</sup> tational properties of interest. We motivate this methodology with a worked example analyzing  
<sup>15</sup> sensitivity in the stomatogastric ganglion. We then use it to generate insights into neuron-type  
<sup>16</sup> input-responsivity in a model of primary visual cortex, a new understanding of rapid task switch-  
<sup>17</sup> ing in superior colliculus models, and attribution of error in recurrent neural networks solving a  
<sup>18</sup> simple mathematical task. More generally, this work suggests a departure from realism vs tractabil-  
<sup>19</sup> ity considerations, towards the use of modern machine learning for sophisticated interrogation of  
<sup>20</sup> biologically relevant models.

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

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

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

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

44 Of course, this same tradeoff has been confronted in many scientific fields characterized by the  
45 need to do inference in complex models. In response, the machine learning community has made  
46 remarkable progress in recent years, via the use of deep neural networks as a powerful inference  
47 engine: a flexible function family that can map observed phenomena (in this case the measurable  
48 signal of some computation) back to probability distributions quantifying the likely parameter  
49 configurations. One celebrated example of this approach from machine learning, of which we  
50 draw key inspiration for this work, is the variational autoencoder [6, 7], which uses a deep neural  
51 network to induce an (approximate) posterior distribution on hidden variables in a latent variable  
52 model, given data. Indeed, these tools have been used to great success in neuroscience as well,  
53 in particular for interrogating parameters (sometimes treated as hidden states) in models of both  
54 cortical population activity [8, 9, 10, 11] and animal behavior [12, 13, 14]. These works have used  
55 deep neural networks to expand the expressivity and accuracy of statistical models of neural data  
56 [15].

57 However, these inference tools have not significantly influenced the study of theoretical neuroscience  
58 models, for at least three reasons. First, at a practical level, the nonlinearities and dynamics of

many theoretical models are such that conventional inference tools typically produce a narrow set of insights into these models. Indeed, only in the last few years has deep learning research advanced to a point of relevance to this class of problem. Second, the object of interest from a theoretical model is not typically data itself, but rather a qualitative phenomenon – inspection of model behavior, or better, a measurable signature of some computation – an *emergent property* of the model. Third, because theoreticians work carefully to construct a model that has biological relevance, such a model as a result often does not fit cleanly into the framing of a statistical model. Technically, because many such models stipulate a noisy system of differential equations that can only be sampled or realized through forward simulation, they lack the explicit likelihood and priors central to the probabilistic modeling toolkit.

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

Equipped with this methodology, we then investigated three models of current importance in theoretical neuroscience. These models were chosen to demonstrate generality through ranges of biological realism (from conductance-based biophysics to recurrent neural networks), neural sys-

tem function (from pattern generation to abstract cognitive function), and network scale (from four to infinite neurons). First, we use EPI to produce a set of verifiable hypotheses of input-responsivity in a four neuron-type dynamical model of primary visual cortex; we then validate these hypotheses in the model. Second, we demonstrated how the systematic application of EPI to levels of task performance can generate experimentally testable hypotheses regarding connectivity in superior colliculus. Third, we use EPI to uncover the sources of error in a low-rank recurrent neural network executing a simple mathematical task. The novel scientific insights offered by EPI contextualize and clarify the previous studies exploring these models [23, 24, 25, 26] and more generally, suggests a departure from realism vs tractability considerations towards the use of modern machine learning for sophisticated interrogation of biologically relevant models.

We note that, during our preparation and early presentation of this work [27, 28], another work has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural circuits [29]. We are encouraged by this general problem being recognized by others in the community, and we emphasize that these works offer complementary neuroscientific contributions (different theoretical models of focus) and use different technical methodologies (ours is built on our prior work [20], theirs similarly [30]). These distinct methodologies and scientific investigations emphasize the increased importance and timeliness of both works.

## 3 Results

### 3.1 Motivating emergent property inference of theoretical models

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



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

121 with the fast or slow population or both.

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

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

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

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

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

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

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

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

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

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

161 which completes the quantification of the emergent property.

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

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

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

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

EPI optimizes the weights and biases  $\theta$  of the deep neural network (which induces the probability distribution) by iteratively solving Equation 3. The optimization is complete when the sampled models with parameters  $z \sim q_\theta$  produce activity consistent with the specified emergent property. Such convergence is evaluated with a hypothesis test that the mean of each emergent property statistic is not different than its emergent property value (see Section A.1.2). In relation to broader methodology, inspection of the EPI objective reveals a natural relationship to posterior inference. Specifically, EPI executes variational inference in an exponential family model, the sufficient statistics and mean parameter of which are defined by  $T(x)$  and  $\mu$ , respectively (see Section A.1.4). Equipped with this method, we now prove out the value of EPI by using it to investigate and produce novel insights about three prominent models in neuroscience.

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

Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output function have succeeded in explaining a host of experimentally documented phenomena. In a regime characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to paradoxical responses [4], selective amplification [36], surround suppression [37] and normalization [38]. Despite their strong predictive power, E-I circuit models rely on the assumption that inhibition is composed of distinct elements (parvalbumin (P), somatostatin(S), VIP (V)) composing 80% of GABAergic interneurons in V1 [39, 40, 41] and that these inhibitory cell types follow specific connectivity patterns (Fig. 2A) [42]. Recent theoretical advances [24, 43, 44], have only started to address the consequences of this multiplicity in the dynamics of V1, strongly relying on linear theoretical tools. Here, we use EPI to go beyond linear theory by systematically generating and evaluating hypotheses of circuit model function using distributions of parameters producing various neuron-type population responses.

Specifically, we consider a four-dimensional circuit model with dynamical state given by the firing

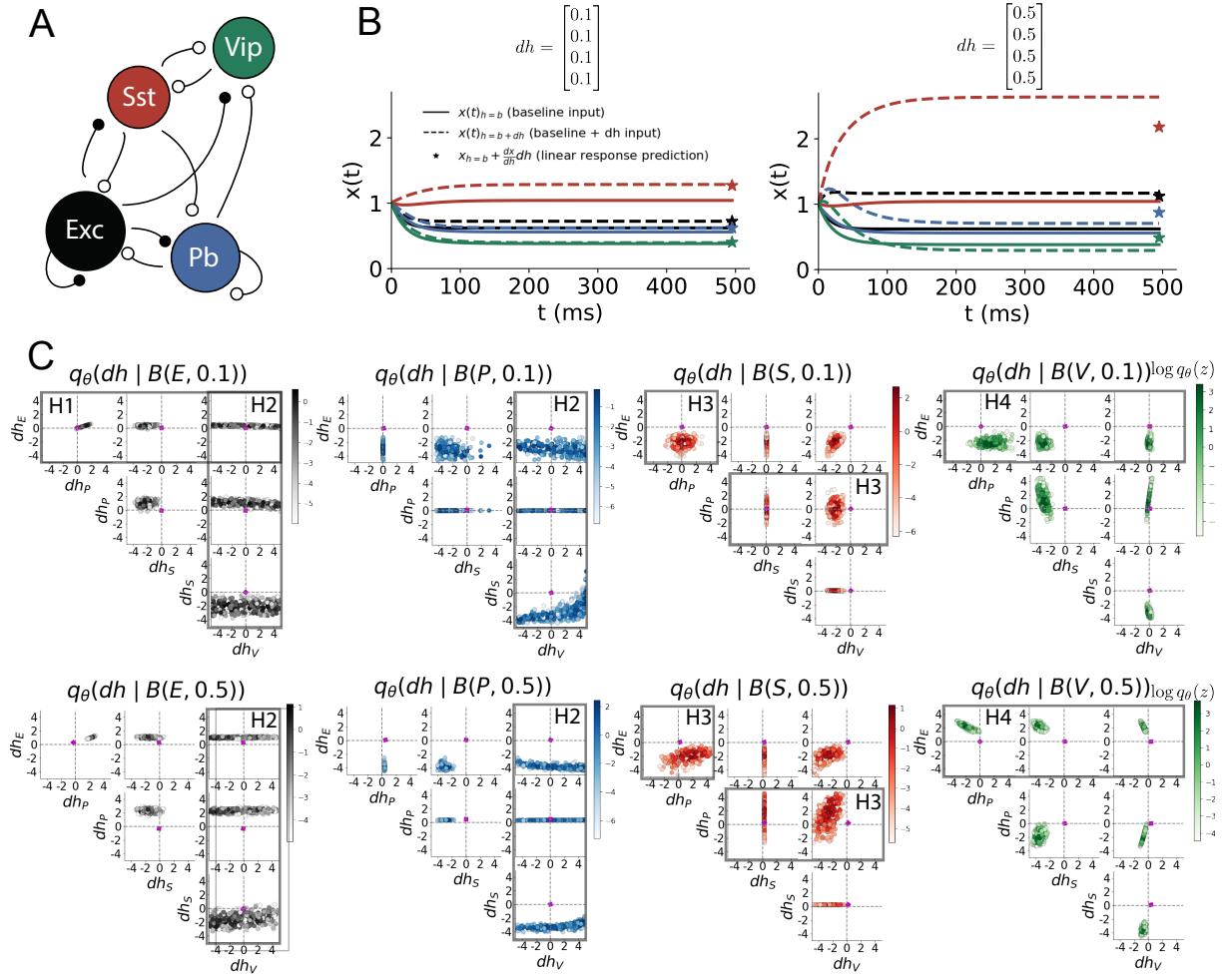


Figure 2: Hypothesis generation through EPI in a V1 model. A. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons. Some neuron-types largely do not form synaptic projections to others (excitatory and inhibitory projections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

rate  $x$  of each neuron-type population  $x = [x_E, x_P, x_S, x_V]^\top$ . Given a time constant of  $\tau = 20$  ms and a power  $n = 2$ , the dynamics are driven by the rectified ( $\|\cdot\|_+$ ) and exponentiated sum of recurrent ( $Wx$ ) and external  $h$  inputs:

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

The effective connectivity weights  $W$  were obtained from experimental recordings of publicly available datasets of mouse V1 [45, 46] (see Section A.2.2). The input  $h = b + dh$  is comprised of a baseline input  $b = [b_E, b_P, b_S, b_V]^\top$  and a differential input  $dh = [dh_E, dh_P, dh_S, dh_V]^\top$  to each neuron-type population. Throughout subsequent analyses, the baseline input is  $b = [1, 1, 1, 1]^\top$ .

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

To get a more comprehensive picture of the input-responsivity of each neuron-type beyond linear theory, we used EPI to learn a distribution of the differential inputs to each population  $dh$  that produce an increase of  $y \in \{0.1, 0.5\}$  in the rate of each neuron-type population  $\alpha \in \{E, P, S, V\}$ . We want to know the differential inputs  $dh$  that result in a differential steady state  $dx_\alpha$  (the change in  $x_\alpha$  when receiving input  $h = b + dh$  with respect to the baseline  $h = b$ ) of value  $y$  with some small, arbitrarily chosen amount of variance 0.01<sup>2</sup>. These statements amount to the emergent property

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

We maintain the notation  $\mathcal{B}(\cdot)$  throughout the rest of the study as short hand for emergent property, which represents a different signature of computation in each application. In each column of Figure 2C visualizes the inferred distribution, available through EPI, of  $dh$  corresponding to an excitatory (red), parvalbumin (blue), somatostatin (red) and VIP (green) neuron-type increase, while each row corresponds to amounts of increase 0.1 and 0.5. For each pair of parameters, we show the two-dimensional marginal distribution of samples colored by  $\log q_\theta(dh \mid \mathcal{B}(\alpha, y))$ . The

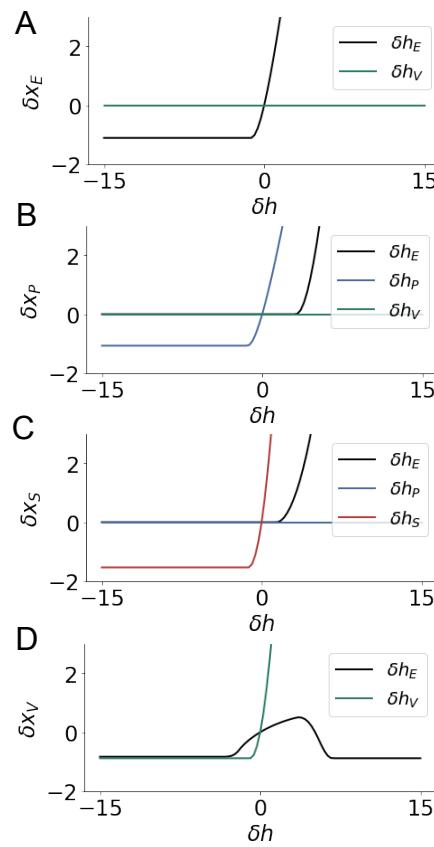


Figure 3: Confirming EPI generated hypotheses in V1. A. Differential responses by the E-population to changes in individual input  $\delta h_\alpha \hat{u}_\alpha$  away from the mode of the EPI distribution  $dh^*$ . B-D Same plots for the P-, S-, and V-populations. Labels H1, H2, H3, and H4 indicate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

231 inferred distributions immediately suggest four hypotheses:

232

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

236 H2: the E- and P-populations should be largely unaffected by input to the V-population (Fig.  
237 2C H2 gray boxes indicate high variance in  $dh_V$  when  $\alpha \in \{E, P\}$ );

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

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

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

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

<sup>245</sup>  $\delta x_\alpha$  is the change in steady state response to the system with input  $h = b + dh^* + \delta h_\alpha \hat{u}_\alpha$  compared  
<sup>246</sup> to  $h = b + dh^*$ , where  $\hat{u}_\alpha$  is a unit vector in the dimension of  $\alpha$ . The EPI-generated hypotheses are  
<sup>247</sup> confirmed.

<sup>248</sup> H1: the neuron-type responses are sensitive to their direct inputs (Fig. 3A black, 3B blue,  
<sup>249</sup> 3C red, 3D green);

<sup>250</sup> H2: the E- and P-populations are not affected by  $\delta h_V$  (Fig. 3A green, 3B green);

<sup>251</sup> H3: the S-population is not affected by  $\delta h_P$  (Fig. 3C blue);

<sup>252</sup> H4: the V-population exhibits a nonmonotonic response to  $\delta h_E$  (Fig. 3D black), and is in  
<sup>253</sup> fact the on population to do so (Fig. 3A-C black).

<sup>254</sup> These hypotheses were in stark contrast to what was available to us via traditional analytical linear  
<sup>255</sup> prediction (Fig. 2C, magenta). To this point, we have shown the utility of EPI on relatively low-  
<sup>256</sup> level emergent properties like network syncing and differential neuron-type population responses.  
<sup>257</sup> In the remainder of the study, we focus on using EPI to understand models of more abstract  
<sup>258</sup> cognitive function.

### <sup>259</sup> 3.4 Identifying neural mechanisms of behavioral learning

<sup>260</sup> In a rapid task switching experiment [47], rats were explicitly cued on each trial to either orient  
<sup>261</sup> towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A)  
<sup>262</sup> task (Fig. 4a). Neural recordings in the midbrain superior colliculus (SC) exhibited two populations  
<sup>263</sup> of neurons that simultaneously represented both task context (Pro or Anti) and motor response  
<sup>264</sup> (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons [25]. Duan  
<sup>265</sup> et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is a four-  
<sup>266</sup> population dynamical system. Here, the neuron-type populations are functionally-defined as the  
<sup>267</sup> Pro- and Anti-populations in each hemisphere (left (L) and right (R)). The Pro- or Anti-populations  
<sup>268</sup> receive an input determined by the cue, and then the left and right populations receive an input  
<sup>269</sup> based on the side of the light stimulus. Activities were bounded between 0 and 1, so that a high  
<sup>270</sup> output of the Pro population in a given hemisphere corresponds to the contralateral response. An  
<sup>271</sup> additional stipulation is that when one Pro population responds with a high-output, the opposite  
<sup>272</sup> Pro population must respond with a low output. Finally, this circuit operates in the presence of  
<sup>273</sup> Gaussian noise resulting in trial-to-trial variability (see Section A.2.3). The connectivity matrix is  
<sup>274</sup> parameterized by the geometry of the population arrangement (Fig. 4B).

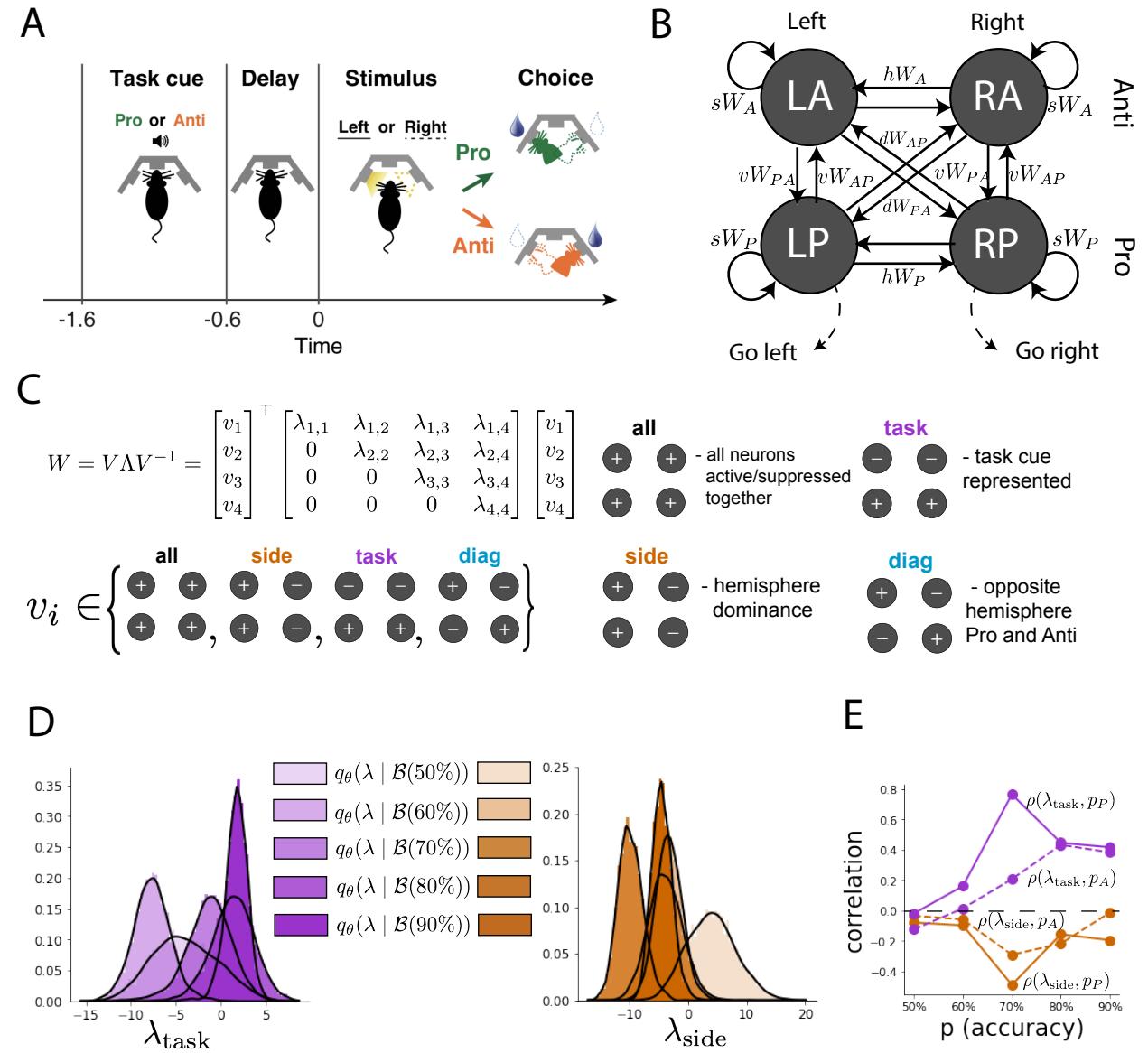


Figure 4: EPI reveals changes in SC [25] connectivity that control task accuracy. A. Rapid task switching behavioral paradigm (see text). B. Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters:  $sW$  - self,  $hW$  - horizontal,  $vW$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

275 Here, we used EPI to learn distributions of the SC weight matrix parameters  $z = W$  conditioned  
 276 on of various levels of rapid task switching accuracy  $\mathcal{B}(p)$  for  $p \in \{50\%, 60\%, 70\%, 80\%, 90\%\}$  (see  
 277 Section A.2.3). Following the approach in Duan et al., we decomposed the connectivity matrix  
 278  $W = V\Lambda V^{-1}$  in such a way (the Schur decomposition) that the basis vectors  $v_i$  are the same for all  
 279  $W$  (Fig. 4C). These basis vectors have intuitive roles in processing for this task, and are accordingly  
 280 named the *all* mode - all neurons co-fluctuate, *side* mode - one side dominates the other, *task* mode  
 281 - the Pro or Anti populations dominate the other, and *diag* mode - Pro- and Anti-populations of  
 282 opposite hemispheres dominate the opposite pair. The corresponding eigenvalues (e.g.  $\lambda_{\text{task}}$ , which  
 283 change according to  $W$ ) indicate the degree to which activity along that mode is increased or  
 284 decreased by  $W$ .

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

### 296 3.5 Linking RNN connectivity to error

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

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

302 where  $x(t)$  is the network activity,  $W$  is the network connectivity,  $\phi(\cdot) = \tanh(\cdot)$ , and  $h(t)$  is the  
 303 input to the system. Such RNNs are trained to do a task from a systems neuroscience experiment,  
 304 and then the unit activations of the trained RNN are compared to recorded neural activity. Fully-

305 connected RNNs with tens of thousands of parameters are challenging to characterize [49], especially  
 306 making statistical inferences about their parameterization. Alternatively, we consider a rank-1,  $N$ -  
 307 neuron RNN with connectivity

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

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

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

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

316 The RNN is trained to solve this task by producing readout activity that is on average the posterior  
 317 mean  $\mu_{post}$ , and activity whose variability is the posterior variance  $\sigma_{post}^2$  (Fig. 5A, a setup inspired  
 318 by [50]). To solve this Gaussian posterior conditioning task, the RNN response to a constant input  
 319  $h(t) = yw + (n - M_n)$  must equal the posterior mean along readout vector  $r$ , where

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

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

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

326 We specify a substantial amount of variance in these emergent property statistics, so that the  
 327 inferred distribution results in RNNs with a variety errors in their solutions to the gaussian posterior  
 328 conditioning problem.

329 We used EPI to learn distributions of RNN connectivity properties  $z = [g \ M_m \ M_n]$  executing  
 330 Gaussian posterior conditioning given an input of  $y = 2$  (see Section A.2.4) (Fig. 5B). The true  
 331 Gaussian conditioning posterior for an input of  $y = 2$  is  $\mu_{\text{post}} = 3$  and  $\sigma_{\text{post}} = 0.5$ . We examined the  
 332 nature of the over- and under-estimation of the posterior means (Fig. 5B, left) and variances (Fig.  
 333 5B, right) in the inferred distributions. There is rough symmetry in the  $M_m$ - $M_n$  plane, suggesting  
 334 a degeneracy in the product of  $M_m$  and  $M_n$  (Fig. 5B). The product of  $M_m$  and  $M_n$  strongly  
 335 determines the posterior mean (Fig. 5B, left), and the random strength  $g$  is the most influential  
 336 variable on the chaotic variance (Fig. 5B, right). Neither of these observations were obvious from  
 337 what mathematical analysis is available in networks of this type (see Section A.2.4). While the  
 338 relationship of the random strength to chaotic variance (and resultingly posterior variance) is well-  
 339 known [3], the distribution admits a novel hypothesis: the estimation of the posterior mean by the  
 340 RNN increases with the product of  $M_m$  and  $M_n$ .

341 Testing these now in finite-size networks. Will almost certainly confirm the hypothesis about  
 342 product of  $M_m$  and  $M_n$ . This novel procedure of doing inference in interpretable parameterizations  
 343 of RNNs conditioned on the emergent property of task execution is straightforwardly generalizable  
 344 to other tasks like noisy integration and context-dependent decision making (Fig. S4).

## 345 4 Discussion

### 346 4.1 EPI is a general tool for theoretical neuroscience

347 Biologically realistic models of neural circuits are comprised of complex nonlinear differential equa-  
 348 tions, making traditional theoretical analysis and statistical inference intractable. In contrast, EPI  
 349 is capable of learning distributions of parameters in such models producing measurable signatures  
 350 of computation. We have demonstrated its utility on biological models (STG), intermediate-level  
 351 models of interacting genetically- and functionally-defined neuron-types (V1, SC), and the most  
 352 abstract of models (RNNs). We are able to condition both deterministic and stochastic models on  
 353 low-level emergent properties like spiking frequency of membrane potentials, as well as high-level  
 354 cognitive function like posterior conditioning. Technically, EPI is tractable when the emergent  
 355 property statistics are continuously differentiable with respect to the model parameters, which is  
 356 very often the case; this emphasizes the general applicability of EPI.

357 In this study, we have focused on applying EPI to low dimensional parameter spaces of models

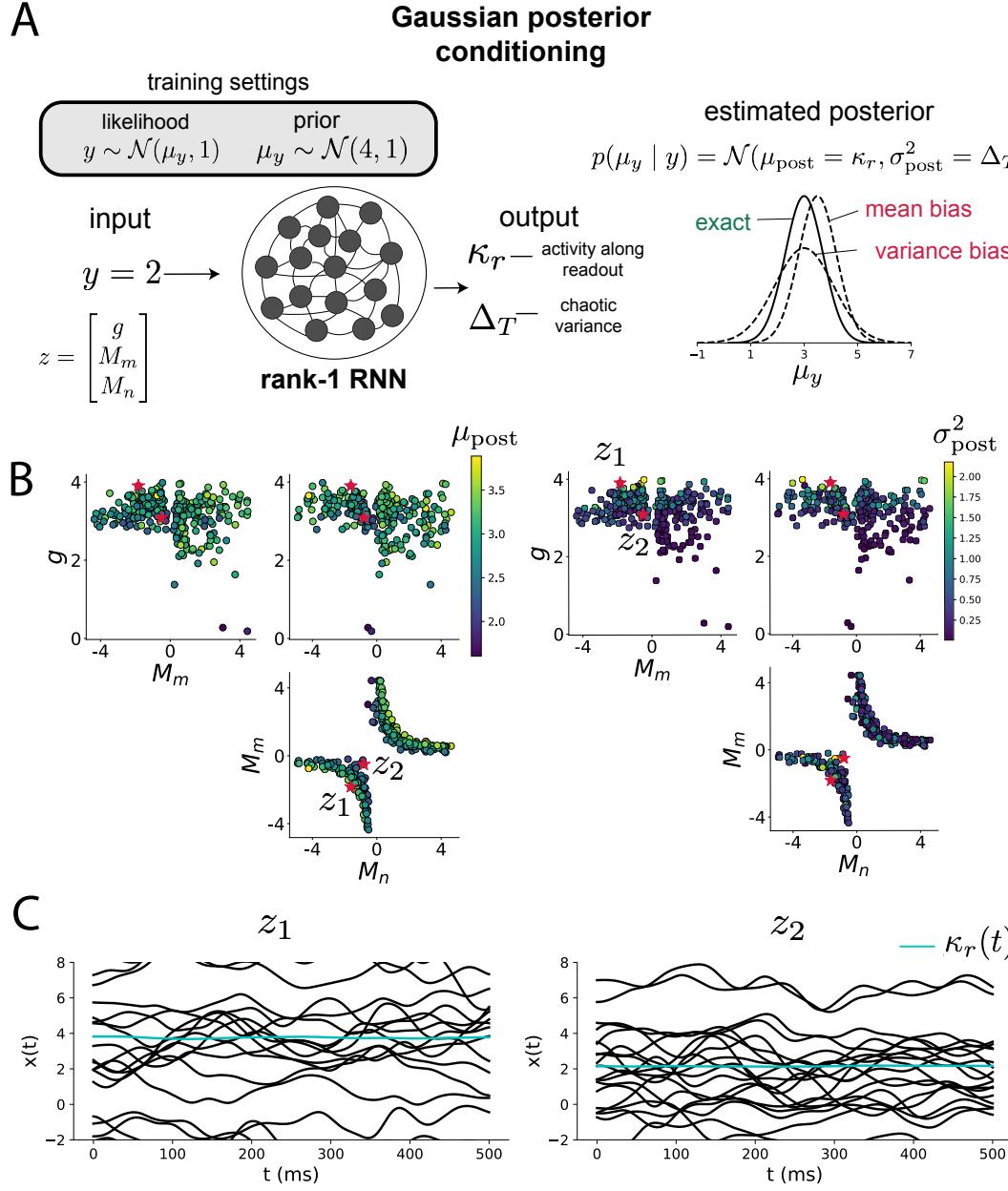


Figure 5: Sources of error in an RNN solving a simple task. A. (left) A rank-1 RNN executing a Gaussian posterior conditioning computation on  $\mu_y$ . (right) Error in this computation can come from over- or under-estimating the posterior mean or variance. B. EPI distribution of rank-1 RNNs executing Gaussian posterior conditioning. Samples are colored by (left) posterior mean  $\mu_{\text{post}} = \kappa_r$  and (right) posterior variance  $\sigma_{\text{post}}^2 = \Delta T$ . C. Finite-size networks with parameters  $z_1$  and  $z_2$  sampled from the distribution attempt the computation and have the errors expected from their parameter values. Activity along readout  $\kappa_r$  (cyan).

358 with low dimensional dynamical states. These choices were made to present the reader with a  
359 series of interpretable conclusions, which is more challenging in high dimensional spaces. In fact,  
360 EPI should scale reasonably to high dimensional parameter spaces, as the underlying technology has  
361 produced state-of-the-art performance on high-dimensional tasks such as texture generation [20]. Of  
362 course, increasing the dimensionality of the dynamical state of the model makes optimization more  
363 expensive, and there is a practical limit there as with any machine learning approach. Although,  
364 theoretical approaches (e.g. [26]) can be used to reason about the wholistic activity of such high  
365 dimensional systems by introducing some degree of additional structure into the model.

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

## 377 4.2 Novel hypotheses from EPI

378 In neuroscience, machine learning has primarily been used to revealed structure in large-scale neural  
379 datasets [51, 52, 53, 54, 55, 56] (see review, [15]). Such careful inference procedures are developed  
380 for these statistical models allowing precise, quantitative reasoning, which clarifies the way data  
381 informs knowledge of the model parameters. However, these inferable statistical models lack re-  
382 semblance to the underlying biology, making it unclear how to go from the structure revealed by  
383 these methods, to the neural mechanisms giving rise to it. In contrast, theoretical neuroscience has  
384 focused on careful mechanistic modeling and the production of emergent properties of computation.  
385 The careful steps of 1.) model design and 2.) emergent property definition, are followed by 3.)  
386 practical inference methods resulting in an opaque characterization of the way model parameters  
387 govern computation. In this work, we replaced this opaque procedure of parameter identification  
388 in theoretical neuroscience with emergent property inference, opening the door to careful inference

389 in careful models of neural computation.

390 Biologically realistic models of neural circuits often prove formidable to analyze. For example,  
391 consider the fact that we do not fully understand the (only) four-dimensional models of V1 [24]  
392 and SC [25]. Because analytical approaches to studying nonlinear dynamical systems become  
393 increasingly complicated when stepping from two-dimensional to three- or four-dimensional systems  
394 in the absence of restrictive simplifying assumptions [57], it is unsurprising that these models pose  
395 a challenge. In Section 3.3, we showed that EPI was far more informative about neuron-type  
396 input-responsivity than the predictions afforded through the available linear analytic methods. By  
397 flexibly conditioning this V1 model on different emergent properties, we performed an exploratory  
398 analysis of a *model* rather than a dataset, which generated a set of testable hypotheses, which  
399 were proved out. Of course, exploratory analyses can be directed towards formulating hypotheses  
400 of a specific form. For example, when interested in model parameter changes with behavioral  
401 performance, one can use EPI to condition on various levels of task accuracy as we did in Section  
402 3.4. This analysis identified experimentally testable predictions (proved out *in-silico*) of patterns  
403 of effective connectivity in SC that should be correlated with increased performance.

404 In our final analysis, we presented a novel procedure for doing statistical inference on interpretable  
405 parameterizations of RNNs executing simple tasks. Specifically, we analyzed RNNs solving a pos-  
406 terior conditioning problem in the spirit of [50]. This methodology relies on recently extended  
407 theory of responses in random neural networks with minimal structure [26]. While we focused on  
408 rank-1 RNNs, which were sufficient for solving this task, we can more generally use this approach  
409 to analyze rank-2 and greater RNNs. The ability to apply the probabilistic model selection toolkit  
410 to such black box models should prove invaluable as their use in neuroscience increases.

## 411 References

- 412 [1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 413 [2] John J Hopfield. Neural networks and physical systems with emergent collective computational  
414 abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 415 [3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural  
416 networks. *Physical review letters*, 61(3):259, 1988.

- [4] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17(11):4382–4388, 1997.
- [5] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- [6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.
- [7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. *International Conference on Machine Learning*, 2014.
- [8] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical neural population models through nonlinear embeddings. In *Advances in neural information processing systems*, pages 163–171, 2016.
- [9] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- [10] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders. *bioRxiv*, page 399246, 2018.
- [11] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, page 1, 2018.
- [12] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- [13] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- [14] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall, Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:

- 446 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural*  
447 *Information Processing Systems*, 2019.
- 448 [15] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-  
449 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,  
450 2018.
- 451 [16] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R  
452 Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology,  
453 and beyond. *The Journal of chemical physics*, 143(1):07B201\_1, 2015.
- 454 [17] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.  
455 *International Conference on Machine Learning*, 2015.
- 456 [18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
457 *arXiv preprint arXiv:1605.08803*, 2016.
- 458 [19] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density  
459 estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.
- 460 [20] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow  
461 networks. *International Conference on Learning Representations*, 2017.
- 462 [21] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-  
463 free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–  
464 5533, 2017.
- 465 [22] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness,  
466 and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.
- 467 [23] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an  
468 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.  
469 *Neuron*, 77(5):845–858, 2013.
- 470 [24] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and vi-  
471 sual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,  
472 115(3):1399–1409, 2016.

- 473 [25] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J  
474 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor  
475 routing. *bioRxiv*, page 245613, 2018.
- 476 [26] Francesca Mastrogiovanni and Srdjan Ostoic. Linking connectivity, dynamics, and computa-  
477 tions in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- 478 [27] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-  
479 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience  
480 Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 481 [28] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,  
482 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with  
483 degenerate solution networks. *Bernstein Conference*, 2019.
- 484 [29] Jan-Matthis Lueckmann, Pedro Goncalves, Chaitanya Chintaluri, William F Podlaski, Gia-  
485 como Bassetto, Tim P Vogels, and Jakob H Macke. Amortised inference for mechanistic models  
486 of neural dynamics. In *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon,  
487 Portugal*, 2019.
- 488 [30] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-  
489 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural  
490 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 491 [31] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-  
492 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 493 [32] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit  
494 parameters. *Nature neuroscience*, 7(12):1345, 2004.
- 495 [33] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,  
496 1957.
- 497 [34] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:  
498 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 499 [35] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise  
500 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.

- 501 [36] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of  
502 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 503 [37] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory  
504 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–  
505 592, 2009.
- 506 [38] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear  
507 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,  
508 85(2):402–417, 2015.
- 509 [39] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and  
510 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,  
511 5(10):793, 2004.
- 512 [40] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of  
513 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-  
514 biology*, 71(1):45–61, 2011.
- 515 [41] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex:  
516 From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 517 [42] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-  
518 bition of inhibition in visual cortex: the logic of connections between molecularly distinct  
519 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 520 [43] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing Wang.  
521 Paradoxical response reversal of top- down modulation in cortical circuits with three interneu-  
522 ron types. *Elife*, 6:1–15, 2017.
- 523 [44] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying  
524 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 525 [45] (2018) Allen Institute for Brain Science. Layer 4 model of v1. available from:  
526 <https://portal.brain-map.org/explore/models/l4-mv1>.
- 527 [46] Yazan N Billeh, Binghuang Cai, Sergey L Gratiy, Kael Dai, Ramakrishnan Iyer, Nathan W  
528 Gouwens, Reza Abbasi-Asl, Xiaoxuan Jia, Joshua H Siegle, Shawn R Olsen, et al. Systematic

- 529 integration of structural and functional data into multi-scale models of mouse primary visual  
530 cortex. *bioRxiv*, page 662189, 2019.
- 531 [47] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain  
532 regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.
- 533 [48] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current  
534 opinion in neurobiology*, 46:1–6, 2017.
- 535 [49] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-  
536 dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.
- 537 [50] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like  
538 dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*,  
539 page 696088, 2019.
- 540 [51] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,  
541 13(8):1713–1720, 2001.
- 542 [52] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.  
543 A statistical paradigm for neural spike train decoding applied to position prediction from  
544 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–  
545 7425, 1998.
- 546 [53] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding  
547 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 548 [54] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and  
549 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis  
550 of neural population activity. In *Advances in neural information processing systems*, pages  
551 1881–1888, 2009.
- 552 [55] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W  
553 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.  
554 *Science*, 349(6244):184–187, 2015.
- 555 [56] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable  
556 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th Inter-  
557 national Conference on Machine Learning*, 2019.

- 558 [57] Steven H Strogatz. Nonlinear dynamics and chaos: with applications to physics. *Biology,*  
 559 *Chemistry, and Engineering (Studies in Nonlinearity)*, Perseus, Cambridge, UK, 1994.
- 560 [58] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial*  
 561 *Intelligence and Statistics*, pages 814–822, 2014.
- 562 [59] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and  
 563 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 564 [60] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.  
 565 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 566 [61] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for  
 567 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

568 **A Methods**

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

570 Emergent property inference (EPI) learns distributions of theoretical model parameters that pro-  
 571 duce emergent properties of interest by combining ideas from maximum entropy flow networks  
 572 (MEFNs) [20] and likelihood-free variational inference (LFVI) [21]. Consider model parameteri-  
 573 zation  $z$  and data  $x$  which has an intractable likelihood  $p(x | z)$  defined by a model simulator of  
 574 which samples are available  $x \sim p(x | z)$ . EPI optimizes a distribution  $q_\theta(z)$  (itself parameterized  
 575 by  $\theta$ ) of model parameters  $z$  to produce an emergent property of interest  $\mathcal{B}$ ,

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

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

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

$$z = f_\theta(\omega) = f_l(\dots, f_1(w)) \quad (13)$$

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

$$\begin{aligned} q_\theta^*(z) &= \operatorname{argmax}_{q_\theta \in \mathcal{Q}} H(q_\theta(z)) \\ \text{s.t. } \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] &= \mu \end{aligned} \quad (14)$$

587 Since we are optimizing parameters  $\theta$  of our deep probability distribution with respect to the entropy  
 588  $H(q_\theta(z))$ , we will need to take gradients with respect to the log probability density of samples from  
 589 the deep probability distribution.

$$H(q_\theta(z)) = \int -q_\theta(z) \log(q_\theta(z)) dz = \mathbb{E}_{z \sim q_\theta} [-\log(q_\theta(z))] = \mathbb{E}_{w \sim q_0} [-\log(q_\theta(f_\theta(w)))] \quad (15)$$

590

$$\nabla_\theta H(q_\theta(z)) = \mathbb{E}_{w \sim q_0} [-\nabla_\theta \log(q_\theta(f_\theta(w)))] \quad (16)$$

591 This optimization is done using the approach of MEFN [20], using architectures for deep proba-  
 592 bility distributions, called normalizing flows (see Section A.1.3), conferring a tractable calculation  
 593 of sample probability. In EPI, this methodology for learning maximum entropy distributions is  
 594 repurposed toward variational learning of model parameter distributions. Similar to LFVI [21], we  
 595 are motivated to do variational learning in models with intractable likelihood functions, in which  
 596 standard methods like stochastic gradient variational Bayes [6] or black box variational inference[58]  
 597 are not tractable. Furthermore, EPI focuses on setting mathematically defined emergent property  
 598 statistics to emergent property values of interest, whereas LFVI is focused on learning directly from  
 599 datasets. Optimizing this objective is a technological challenge, the details of which we elaborate  
 600 in Section A.1.2. Before going through those details, we ground this optimization in a toy example.

### 601 A.1.1 Example: 2D LDS

602 To gain intuition for EPI, consider a two-dimensional linear dynamical system model

$$\tau \frac{dx}{dt} = Ax \quad (17)$$

603 with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \quad (18)$$

604 To do EPI with the dynamics matrix elements as the free parameters  $z = [a_1 \ a_2 \ a_3 \ a_4]$  (fixing  
 605  $\tau = 1$ ), the emergent property statistics  $T(x)$  were chosen to contain the first- and second-moments

of the oscillatory frequency  $\omega$  and the growth/decay factor  $d$  of the oscillating system. To learn the distribution of real entries of  $A$  that yield a distribution of  $d$  with mean zero with variance  $0.25^2$ , and oscillation frequency  $\omega$  with mean 1 Hz with variance  $(0.1\text{Hz})^2$ , we selected the real part of the eigenvalue  $\text{real}(\lambda_1) = d$  and imaginary component of  $\text{imag}(\lambda_1) = 2\pi\omega$  as the emergent property statistics.  $\lambda_1$  is the eigenvalue of greatest real part when there is zero imaginary component, and alternatively of positive imaginary component, when the eigenvalues are complex conjugate pairs.

Those emergent property statistics were then constrained to

$$\mu = \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \quad (19)$$

where  $\omega = 1\text{Hz}$ . Unlike the models we presented in the main text, which calculate  $\mathbb{E}_{x \sim p(x|z)} [T(x)]$  via forward simulation, we have a closed form for  $\lambda_1$  of the dynamics matrix. The eigenvalues can be calculated using the quadratic formula:

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2} \quad (20)$$

where  $\lambda_1$  is the eigenvalue of  $\frac{1}{\tau}A$  with greatest real part.

Importantly, even though  $\mathbb{E}_{x \sim p(x|z)} [T(x)]$  is calculable directly via a closed form function and does not require simulation, we cannot derive the distribution  $q_\theta^*$  directly. This is due to the formally hard problem of the backward mapping: finding the natural parameters  $\eta$  from the mean parameters  $\mu$  of an exponential family distribution [59]. Instead, we can use EPI to learn the linear system parameters producing such a band of oscillations (Fig. S1B).

Even this relatively simple system has nontrivial (though intuitively sensible) structure in the parameter distribution. To validate our method (further than that of the underlying technology on a ground truth solution [20]) we analytically derived the contours of the probability density from the emergent property statistics and values (Fig. S2). In the  $a_1 - a_4$  plane, the black line at  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$ , and the dotted black line at the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$ , and the grey line at twice the standard deviation  $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.5$  follow the contour of probability density of the samples. (Fig. 2A). The distribution precisely reflects the desired statistical constraints and model degeneracy in the sum of  $a_1$  and  $a_4$ . Intuitively, the parameters equivalent with respect to emergent property statistic  $\text{real}(\lambda_1)$  have similar log densities.

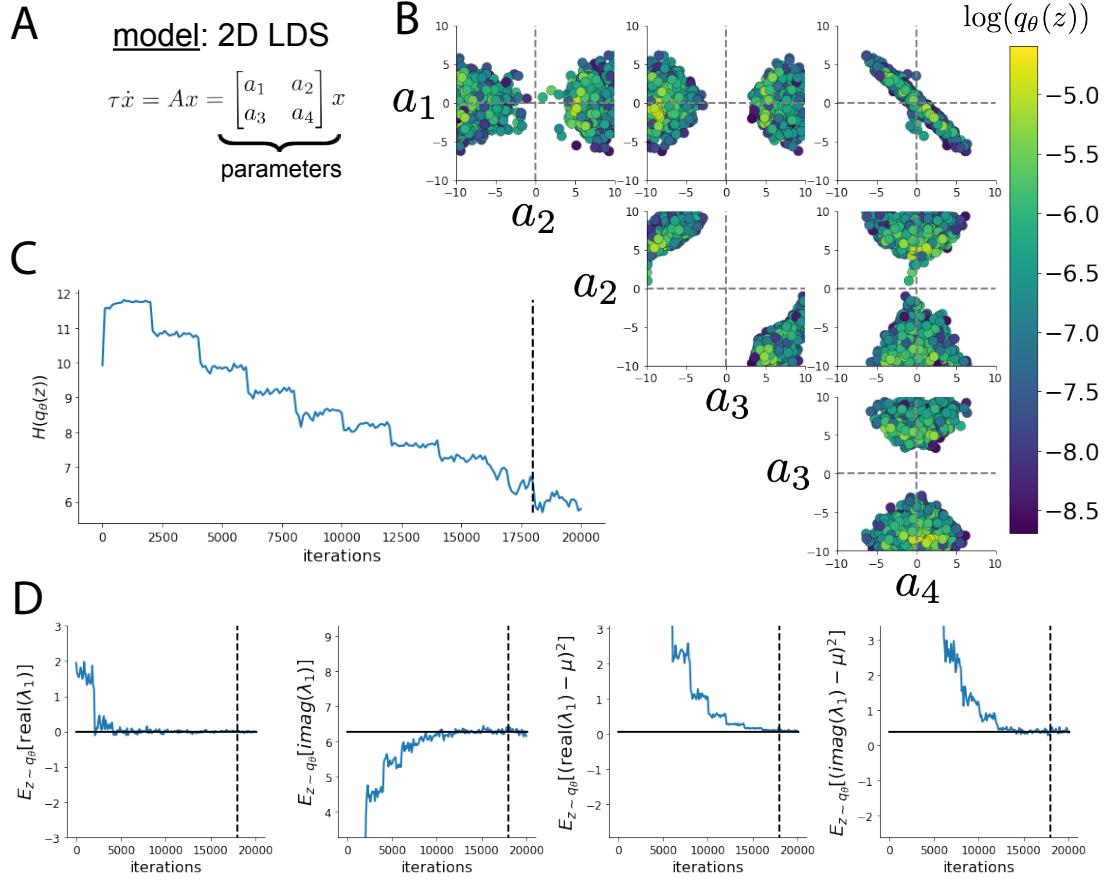


Fig. S1: A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix  $A$  are the parameters. B. The DSN distribution for a two-dimensional linear dynamical system with  $\tau = 1$  that produces an average of 1Hz oscillations with some small amount of variance. C. Entropy throughout the optimization. At the beginning of each augmented Lagrangian epoch (5,000 iterations), the entropy dipped due to the shifted optimization manifold where emergent property constraint satisfaction is increasingly weighted. D. Emergent property moments throughout optimization. At the beginning of each augmented Lagrangian epoch, the emergent property moments move closer to their constraints.

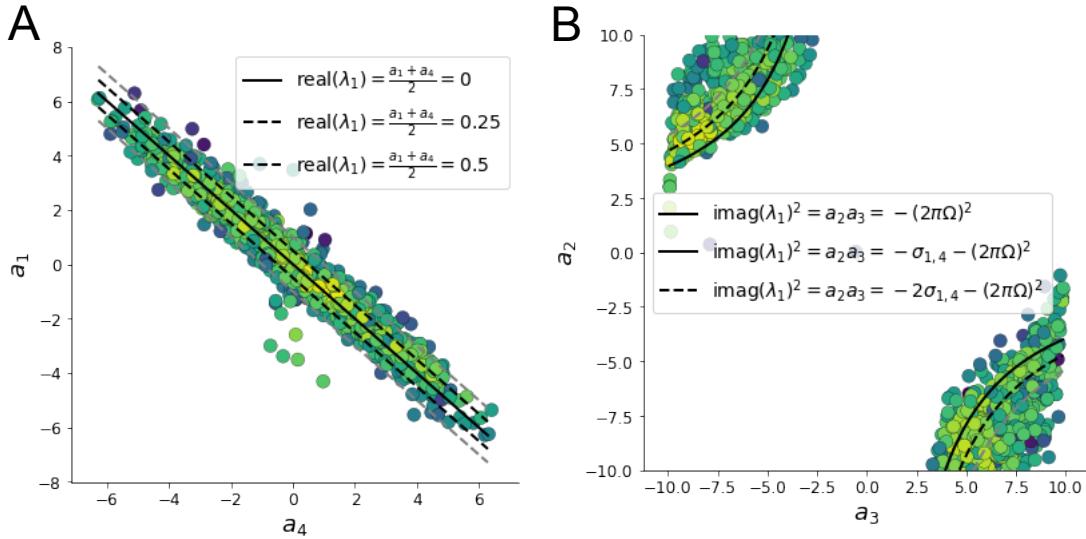


Fig. S2: A. Probability contours in the  $a_1 - a_4$  plane can be derived from the relationship to emergent property statistic of growth/decay factor. B. Probability contours in the  $a_2 - a_3$  plane can be derived from relationship to the emergent property statistic of oscillation frequency.

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1 a_4 - a_2 a_3}{\tau}}, & \text{if } a_1 a_4 < a_2 a_3 \\ 0 & \text{otherwise} \end{cases} \quad (21)$$

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

$$\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 \quad (22)$$

635 Since we constrained  $\mathbb{E}_{z \sim q_\theta} [\text{imag}(\lambda)] = 2\pi$  (with  $\omega = 1$ ), we can plot contours of the equation  
 636  $\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 = (2\pi)^2$  for various  $a_1 a_4$  (Fig. S2A). If  $\sigma_{1,4} = \mathbb{E}_{z \sim q_\theta} [|a_1 a_4 - E_{q_\theta}[a_1 a_4]|]$ ,  
 637 then we plot the contours as  $a_1 a_4 = 0$  (black),  $a_1 a_4 = -\sigma_{1,4}$  (black dotted), and  $a_1 a_4 = -2\sigma_{1,4}$   
 638 (grey dotted) (Fig. S2B). This validates the curved structure of the inferred distribution learned  
 639 through EPI. We take steps in negative standard deviation of  $a_1 a_4$  (dotted and gray lines), since  
 640 there are few positive values  $a_1 a_4$  in the learned distribution. Subtler model-emergent property  
 641 combinations will have even more complexity, further motivating the use of EPI for understanding  
 642 these systems. As we expect, the distribution results in samples of two-dimensional linear systems  
 643 oscillating near 1Hz (Fig. S3).

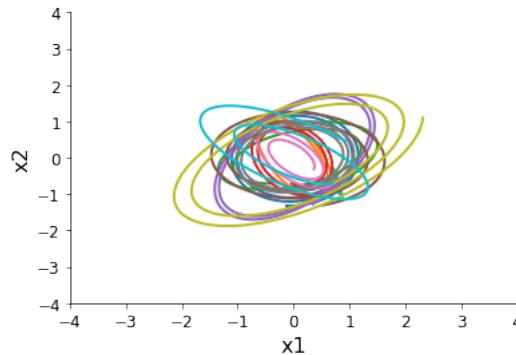


Fig. S3: Sampled dynamical system trajectories from the EPI distribution. Each trajectory is initialized at  $x(0) = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$ .

<sup>644</sup> **A.1.2 Augmented Lagrangian optimization**

<sup>645</sup> To optimize  $q_\theta(z)$  in Equation 14, the constrained optimization is performed using the augmented  
<sup>646</sup> Lagrangian method. The following objective is minimized:

$$L(\theta; \eta, c) = -H(q_\theta) + \eta^\top R(\theta) + \frac{c}{2} \|R(\theta)\|^2 \quad (23)$$

<sup>647</sup> where  $R(\theta) = \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x) - \mu]]$ ,  $\eta \in \mathbb{R}^m$  are the Lagrange multipliers (which are closely  
<sup>648</sup> related to the natural parameters of exponential families (see Section A.1.4)) and  $c$  is the penalty  
<sup>649</sup> coefficient. For a fixed  $(\eta, c)$ ,  $\theta$  is optimized with stochastic gradient descent. A low value of  
<sup>650</sup>  $c$  is used initially, and increased during each augmented Lagrangian epoch, which is a period of  
<sup>651</sup> optimization with fixed  $\eta$  and  $c$  for a given number of stochastic optimization iterations. Similarly,  
<sup>652</sup>  $\eta$  is tuned each epoch based on the constraint violations. For the linear two-dimensional system  
<sup>653</sup> (Fig. S1C), optimization hyperparameters are initialized to  $c_1 = 10^{-4}$  and  $\eta_1 = \mathbf{0}$ . The penalty  
<sup>654</sup> coefficient is updated based on the result of a hypothesis test regarding the reduction in constraint  
<sup>655</sup> violation. The p-value of  $E[\|R(\theta_{k+1})\|] > \gamma \mathbb{E}[\|R(\theta_k)\|]$  is computed, and  $c_{k+1}$  is updated to  $\beta c_k$   
<sup>656</sup> with probability  $1 - p$ . Throughout the study,  $\beta = 4.0$  and  $\gamma = 0.25$  were used. The other update  
<sup>657</sup> rule is  $\eta_{k+1} = \eta_k + c_k \frac{1}{n} \sum_{i=1}^n (T(x^{(i)}) - \mu)$ . In this example, each augmented Lagrangian epoch ran  
<sup>658</sup> for 2,000 iterations. We consider the optimization to have converged when a null hypothesis test of  
<sup>659</sup> constraint violations being zero is accepted for all constraints at a significance threshold 0.05. This  
<sup>660</sup> is the dotted line on the plots below depicting the optimization cutoff of EPI for the 2-dimensional  
<sup>661</sup> linear system.

<sup>662</sup> The intention is that  $c$  and  $\eta$  start at values encouraging entropic growth early in optimization.

663 Then, as they increase in magnitude with each training epoch, the constraint satisfaction terms  
 664 are increasingly weighted, resulting in a decrease in entropy. If the optimization is left to continue  
 665 running, and structural pathologies in the distribution may be introduced.

666 **A.1.3 Normalizing flows**

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

$$q(z') = q(f^{-1}(z')) \left| \det \frac{\partial f^{-1}(z')}{\partial z'} \right| = q(z) \left| \det \frac{\partial f(z)}{\partial z} \right|^{-1} \quad (24)$$

670 However, this computation has cubic complexity in dimensionality for fully connected layers. By  
 671 restricting our layers to normalizing flows [17] – bijective functions with fast log determinant Ja-  
 672 cobian computations, we can tractably optimize deep generative models with objectives that are a  
 673 function of sample density, like entropy. Most of our analyses use real NVP [60], which have proven  
 674 effective in our architecture searches, and have the advantageous features of fast sampling and fast  
 675 probability density evaluation.

676 **A.1.4 Emergent property inference as variational inference in an exponential family**

677 Consider the goal of doing variational inference with an exponential family posterior distribution  
 678  $p(z | x)$ . We use the following abbreviated notation to collect the base measure  $b(z)$  and sufficient  
 679 statistics  $T(z)$  into  $\tilde{T}(z)$  and likewise concatenate a 1 onto the end of the natural parameter  $\tilde{\eta}(x)$ .  
 680 The log normalizing constant  $A(\eta(x))$  remains unchanged.

$$\begin{aligned} p(z | x) &= b(z) \exp \left( \eta(x)^\top T(z) - A(\eta(x)) \right) = \exp \left( \begin{bmatrix} \eta(x) \\ 1 \end{bmatrix}^\top \begin{bmatrix} T(z) \\ b(z) \end{bmatrix} - A(\eta(x)) \right) \\ &= \exp \left( \tilde{\eta}(x)^\top \tilde{T}(z) - A(\eta(x)) \right) \end{aligned} \quad (25)$$

681 Variational inference with an exponential family posterior distribution uses optimization to mini-  
 682 mize the following divergence [61]:

$$q_\theta^* = \operatorname{argmin}_{q_\theta \in Q} KL(q_\theta || p(z | x)) \quad (26)$$

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

$$KL(q_\theta \parallel p(z \mid x)) = \mathbb{E}_{z \sim q_\theta} [\log(q_\theta(z))] - \mathbb{E}_{z \sim q_\theta} [\log(p(z \mid x))] \quad (27)$$

685

$$= -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} [\tilde{\eta}(x)^\top \tilde{T}(z) - A(\eta(x))] \quad (28)$$

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

$$\operatorname{argmin}_{q_\theta \in Q} KL(q_\theta \parallel p(z \mid x)) = \operatorname{argmin}_{q_\theta \in Q} -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} [\tilde{\eta}(x)^\top \tilde{T}(z)] \quad (29)$$

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

$$= \operatorname{argmin}_{q_\theta \in Q} -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} [\tilde{\eta}(x)^\top (\tilde{T}(z) - \mu)] + \tilde{\eta}(x)^\top \mu \quad (30)$$

690

$$= \operatorname{argmin}_{q_\theta \in Q} -H(q_\theta) - \mathbb{E}_{z \sim q_\theta} [\tilde{\eta}(x)^\top (\tilde{T}(z) - \mu)] \quad (31)$$

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

$$q_\theta^*(z) = \operatorname{argmax}_{q_\theta \in Q} H(q_\theta(z)), \text{ s.t. } \mathbb{E}_{z \sim q_\theta} [\mathbb{E}_{x \sim p(x|z)} [T(x)]] = \mu \quad (32)$$

692 The Lagrangian objective (without the augmentation) is

$$q_\theta^* = \operatorname{argmin}_{q_\theta \in Q} -H(q_\theta) + \eta_{\text{opt}}^\top (\mathbb{E}_{z \sim q_\theta} [\tilde{T}(z)] - \mu) \quad (33)$$

693 As the optimization proceeds,  $\eta_{\text{opt}}^\top$  should converge to the natural parameter  $\tilde{\eta}(x)$  through its  
 694 adaptations in each epoch (see Section A.1.2).

695 The derivation of the natural parameter  $\tilde{\eta}(x)$  of an exponential family distribution from its mean  
 696 parameter  $\mu$  is referred to as the backward mapping and is formally hard to identify [59]. Since  
 697 this backward mapping is deterministic, we can replace the notation of  $p(z \mid x)$  with  $p(z \mid \mathcal{B})$   
 698 conceptualizing an inferred distribution that obeys emergent property  $\mathcal{B}$  (see Section A.1).

## 699 A.2 Theoretical models

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

702 **A.2.1 Stomatogastric ganglion**

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

$$C_m \frac{dx_m}{dt} = -[h_{leak}(x; z) + h_{Ca}(x; z) + h_K(x; z) + h_{hyp}(x; z) + h_{elec}(x; z) + h_{syn}(x; z)] \quad (34)$$

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

$$h_{leak}(x; z) = g_{leak}(x_m - V_{leak}) \quad (35)$$

$$h_{elec}(x; z) = g_{el}(x_m^{post} - x_m^{pre}) \quad (36)$$

$$h_{syn}(x; z) = g_{syn}S_\infty^{pre}(x_m^{post} - V_{syn}) \quad (37)$$

$$h_{Ca}(x; z) = g_{Ca}M_\infty(x_m - V_{Ca}) \quad (38)$$

$$h_K(x; z) = g_KN(x_m - V_K) \quad (39)$$

$$h_{hyp}(x; z) = g_hH(x_m - V_{hyp}) \quad (40)$$

710 The reversal potentials were set to  $V_{leak} = -40mV$ ,  $V_{Ca} = 100mV$ ,  $V_K = -80mV$ ,  $V_{hyp} = -20mV$ ,  
 711 and  $V_{syn} = -75mV$ . The other conductance parameters were fixed to  $g_{leak} = 1 \times 10^{-4}\mu S$ .  $g_{Ca}$ ,  
 712  $g_K$ , and  $g_{hyp}$  had different values based on fast, intermediate (hub) or slow neuron. Fast:  $g_{Ca} =$   
 713  $1.9 \times 10^{-2}$ ,  $g_K = 3.9 \times 10^{-2}$ , and  $g_{hyp} = 2.5 \times 10^{-2}$ . Intermediate:  $g_{Ca} = 1.7 \times 10^{-2}$ ,  $g_K = 1.9 \times 10^{-2}$ ,  
 714 and  $g_{hyp} = 8.0 \times 10^{-3}$ . Intermediate:  $g_{Ca} = 8.5 \times 10^{-3}$ ,  $g_K = 1.5 \times 10^{-2}$ , and  $g_{hyp} = 1.0 \times 10^{-2}$ .

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

$$M_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_m - v_1}{v_2} \right) \right) \quad (41)$$

$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (42)$$

$$N_\infty = 0.5 \left( 1 + \tanh \left( \frac{x_m - v_3}{v_4} \right) \right) \quad (43)$$

$$\lambda_N = \phi_N \cosh \left( \frac{x_m - v_3}{2v_4} \right) \quad (44)$$

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (45)$$

$$H_\infty = \frac{1}{1 + \exp\left(\frac{x_m + v_5}{v_6}\right)} \quad (46)$$

$$\tau_h = 272 - \left( \frac{-1499}{1 + \exp\left(\frac{-x_m + v_7}{v_8}\right)} \right) \quad (47)$$

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

Finally, there is a synaptic gating variable as well:

$$S_\infty = \frac{1}{1 + \exp\left(\frac{v_{th} - x_m}{v_9}\right)} \quad (48)$$

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

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

Poor resolution afforded by the discrete Fourier transform motivated the use of an alternative basis of complex exponentials to measure spiking frequency. Instead, we used a basis of complex exponentials with frequencies from 0.0-1.0 Hz at 0.01Hz resolution,  $\Phi = [0.0, 0.01, \dots, 1.0]^\top$

Another consideration was that the frequency spectra of the neurons had 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, we used a sum-of-powers normalization. Let  $\mathcal{X}_i \in \mathcal{C}^{|\Phi|}$  be the complex exponential filter bank dot products with the signal  $x_i \in \mathbb{R}^N$ , where  $i \in \{f1, f2, \text{hub}, s1, s2\}$ . The “frequency identification” vector is

$$v_i = \frac{|\mathcal{X}_i|^\beta}{\sum_{k=1}^N |\mathcal{X}_i(k)|^\beta} \quad (49)$$

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

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

$$E \begin{bmatrix} \omega_{f1} \\ \omega_{f2} \\ \omega_{hub} \\ \omega_{s1} \\ \omega_{s2} \\ (\omega_{f1} - 0.542)^2 \\ (\omega_{f2} - 0.542)^2 \\ (\omega_{hub} - 0.542)^2 \\ (\omega_{s1} - 0.542)^2 \\ (\omega_{s2} - 0.542)^2 \end{bmatrix} = \begin{bmatrix} 0.542 \\ 0.542 \\ 0.542 \\ 0.542 \\ 0.542 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \end{bmatrix} \quad (50)$$

755 For EPI in Fig 2C, we used a real NVP architecture with two coupling layers. Each coupling layer  
 756 had two hidden layers of 10 units each, and we mapped onto a support of  $z \in \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 \\ 8 \end{bmatrix} \right]$  (the  
 757 same considered in [23]). We have shown the EPI optimization that converged with maximum  
 758 entropy across 5 random seeds and augmented Lagrangian coefficient initializations of  $c_0 \in \{5, 10\}$ .

759 **A.2.2 Primary visual cortex**

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

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n \quad (51)$$

761 Some neuron-types largely lack synaptic projections to other neuron-types [42], and it is popular  
 762 to only consider a subset of the effective connectivities [24, 43, 44].

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & 0 \\ W_{PE} & W_{PP} & W_{PS} & 0 \\ W_{SE} & 0 & 0 & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & 0 \end{bmatrix} \quad (52)$$

<sup>763</sup> By consolidating information from many experimental datasets, Billeh et al. [46] produce estimates  
<sup>764</sup> of the synaptic strength (in mV)

$$M = \begin{bmatrix} 0.36 & 0.48 & 0.31 & 0.28 \\ 1.49 & 0.68 & 0.50 & 0.18 \\ 0.86 & 0.42 & 0.15 & 0.32 \\ 1.31 & 0.41 & 0.52 & 0.37 \end{bmatrix} \quad (53)$$

<sup>765</sup> and connection probability

$$C = \begin{bmatrix} 0.16 & 0.411 & 0.424 & 0.087 \\ 0.395 & .451 & 0.857 & 0.02 \\ 0.182 & 0.03 & 0.082 & 0.625 \\ 0.105 & 0.22 & 0.77 & 0.028 \end{bmatrix} \quad (54)$$

<sup>766</sup> Multiplying these connection probabilities and synaptic efficacies gives us an effective connectivity  
<sup>767</sup> matrix:

$$W_{\text{full}} = C \odot M = \begin{bmatrix} 0.16 & 0.411 & 0.424 & 0.087 \\ 0.395 & .451 & 0.857 & 0.02 \\ 0.182 & 0.03 & 0.082 & 0.625 \\ 0.105 & 0.22 & 0.77 & 0.028 \end{bmatrix} \quad (55)$$

<sup>768</sup> We used the entries of this full effective connectivity matrix that are not considered to be ineffectual  
<sup>769</sup> (Equation 52).

<sup>770</sup> We look at how this four-dimensional nonlinear dynamical model of V1 responds to different inputs,  
<sup>771</sup> and compare the predictions of the linear response to the approximate posteriors obtained through  
<sup>772</sup> EPI. The input to the system is the sum of a baseline input  $b = [1 \ 1 \ 1 \ 1]^T$  and a differential  
<sup>773</sup> input  $dh$ :

$$h = b + dh \quad (56)$$

<sup>774</sup> All simulations of this system had  $T = 100$  time points, a time step  $dt = 5\text{ms}$ , and time constant  
<sup>775</sup>  $\tau = 20\text{ms}$ . And the system was initialized to a random draw  $x(0)_i \sim \mathcal{N}(1, 0.01)$ .

<sup>776</sup> We can describe the dynamics of this system more generally by

$$\dot{x}_i = -x_i + f(u_i) \quad (57)$$

<sup>777</sup> where the input to each neuron is

$$u_i = \sum_j W_{ij}x_j + h_i \quad (58)$$

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

$$\frac{dx_{ss}}{dh} = F(W \frac{dx_{ss}}{dh} + I) \quad (59)$$

779 which is calculable by

$$\frac{dx_{ss}}{dh} = (F^{-1} - W)^{-1} \quad (60)$$

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

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

$$\mathcal{B}(\alpha, y) \triangleq \mathbb{E} \begin{bmatrix} dx_{\alpha,ss} \\ (dx_{\alpha,ss} - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix} \quad (61)$$

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

789 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  
790 initial seeds using an architecture of four coupling layers, each with two hidden layers of 10 units.  
791 We set  $c_0 = 10^5$ . The support of the learned distribution was restricted to  $z_i \in [-5, 5]$ .

### 792 A.2.3 Superior colliculus

793 In the model of Duan et al [25], there are four total units: two in each hemisphere corresponding to  
794 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right  
795 Pro (RP) and right Anti (RA). Each unit has an activity ( $x_\alpha$ ) and internal variable ( $u_\alpha$ ) related  
796 by

$$x_\alpha(t) = \left( \frac{1}{2} \tanh \left( \frac{u_\alpha(t) - \epsilon}{\zeta} \right) + \frac{1}{2} \right) \quad (62)$$

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

<sup>799</sup> We order the elements of  $x$  and  $u$  in the following manner

$$x = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad u = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix} \quad (63)$$

<sup>800</sup> The internal variables follow dynamics:

$$\tau \frac{dv}{dt} = -u + Wx + h + \sigma dB \quad (64)$$

<sup>801</sup> with time constant  $\tau = 0.09s$  and Gaussian noise  $\sigma dB$  controlled by the magnitude of  $\sigma = 1.0$ . The  
<sup>802</sup> weight matrix has 8 parameters  $sW_P$ ,  $sW_A$ ,  $vW_{PA}$ ,  $vW_{AP}$ ,  $hW_P$ ,  $hW_A$ ,  $dW_{PA}$ , and  $dW_{AP}$  (Fig.  
<sup>803</sup> 4B).

$$W = \begin{bmatrix} sW_P & vW_{PA} & hW_P & dW_{PA} \\ vW_{AP} & sW_A & dW_{AP} & hW_A \\ hW_P & dW_{PA} & sW_P & vW_{PA} \\ dW_{AP} & hW_A & vW_{AP} & sW_A \end{bmatrix} \quad (65)$$

<sup>804</sup> The system receives five inputs throughout each trial, which has a total length of 1.8s.

$$h = h_{\text{rule}} + h_{\text{choice-period}} + h_{\text{light}} \quad (66)$$

<sup>805</sup> There are rule-based inputs depending on the condition,

$$h_{P,\text{rule}}(t) = \begin{cases} I_{P,\text{rule}} \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (67)$$

$$h_{A,\text{rule}}(t) = \begin{cases} I_{A,\text{rule}} \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (68)$$

<sup>806</sup> a choice-period input,

$$h_{\text{choice}}(t) = \begin{cases} I_{\text{choice}} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (69)$$

<sup>807</sup> and an input to the right or left-side depending on where the light stimulus is delivered.

$$h_{\text{light}}(t) = \begin{cases} I_{\text{light}} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Left} \\ I_{\text{light}} \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Right} \\ 0, & t \leq 1.2s \end{cases} \quad (70)$$

809 The input parameterization was fixed to  $I_{P,\text{rule}} = 10$ ,  $I_{A,\text{rule}} = 10$ ,  $I_{\text{choice}} = 2$ , and  $I_{\text{light}} = 1$   
 810 To produce a Bernoulli rate of  $p_{LP}$  in the Left, Pro condition, let  $\hat{p}_i$  be the empirical average steady  
 811 state (ss) response (final  $x_{LP}$  at end of task) over  $M=500$  Gaussian noise draws for a given SC  
 812 model parameterization  $z_i$ :

$$\hat{p}_i = \mathbb{E}_{\sigma dB} [x_{LP} | s = L, c = P, z_i] = \frac{1}{M} \sum_{j=1}^M x_{LP}(s = L, c = P, z_i, \sigma dB_j) \quad (71)$$

813 where here  $x_\alpha$  denotes the steady state activity at the end of the trial. For the first constraint, the  
 814 average over posterior samples (from  $q_\theta(z)$ ) to be  $p_{LP}$ :

$$\mathbb{E}_{z_i \sim q_\phi} [\mathbb{E}_{\sigma dB} [x_{LP,ss} | s = L, c = P, z_i]] = \mathbb{E}_{z_i \sim q_\phi} [\hat{p}_i] = p_{LP} \quad (72)$$

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

$$\mathbb{E}_{z \sim q_\phi} [\sigma_{err}^2] = 0 \quad (73)$$

817

$$\sigma_{err}^2 = Var_{\sigma dB} [x_{LP} | s = L, c = P, z_i] - \hat{p}_i(1 - \hat{p}_i) \quad (74)$$

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

$$\mathbb{E}_{z \sim q_\phi} [d_P] = 1 \quad (75)$$

820

$$\mathbb{E}_{\sigma dB} [(x_{LP} - x_{RP})^2 | s = L, c = P, z_i] \quad (76)$$

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

830 Training DSNs to learn distributions of dynamical system parameterizations that produce Bernoulli  
 831 responses at a given rate (with small variance around that rate) was harder to do than expected.

832 There is a pathology in this optimization setup, where the learned distribution of weights is bimodal  
 833 attributing a fraction  $p$  of the samples to an expansive mode (which always sends  $x_{LP}$  to 1), and a  
 834 fraction  $1 - p$  to a decaying mode (which always sends  $x_{LP}$  to 0). This pathology was avoided using  
 835 an inequality constraint prohibiting parameter samples that resulted in low variance of responses  
 836 across noise.

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

$$\mathcal{B}(p) \triangleq \begin{bmatrix} \hat{p}_P \\ \hat{p}_A \\ (\hat{p}_P - p)^2 \\ (\hat{p}_A - p)^2 \\ \sigma_{P,err}^2 \\ \sigma_{A,err}^2 \\ d_P \\ d_A \end{bmatrix} = \begin{bmatrix} p \\ p \\ 0.15^2 \\ 0.15^2 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad (77)$$

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

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

841 Recent work establishes a link between RNN connectivity weights and the resulting dynamical  
 842 responses of the network, using dynamic mean field theory (DMFT) [26]. Specifically, DMFT  
 843 describes the properties of activity in infinite-size neural networks given a distribution on the  
 844 connectivity weights. In such a model, the connectivity of a rank-1 RNN (which was sufficient for  
 845 our task), has weight matrix  $W$ , whis is the sum of a random component with strength determined  
 846 by  $g$  and a structured component determined by the outer product of vectors  $m$  and  $n$ :

$$W = g\chi + \frac{1}{N}mn^\top, \quad (78)$$

847 where the activity  $x$  evolves as and  $I(t)$  is some input,  $\phi$  is the tanh nonlinearity, and  $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ .  
 848 The entries of  $m$  and  $n$  are drawn from Gaussian distributions  $m_i \sim \mathcal{N}(M_m, 1)$  and  $n_i \sim \mathcal{N}(M_n, 1)$ .  
 849 From such a parameterization, this theory produces consistency equations for the dynamic mean  
 850 field variables in terms of parameters like  $g$ ,  $M_m$ , and  $M_n$ , which we study in Section 3.5. That  
 851 is the dynamic mean field variables (e.g. the activity along along a vector  $\kappa_v$ , the total variance

852  $\Delta_0$ , structured variance  $\Delta_\infty$ , and the chaotic variance  $\Delta_T$ ) are written as functions of one another  
853 in terms of connectivity parameters. The values of these variables can be used obtained using a  
854 nonlinear system of equations solver. These dynamic mean field variables are then cast as task-  
855 relevant variables with respect to the context of the provided inputs. Mastrogiuseppe et al. designed  
856 low-rank RNN connectivities via minimalist connectivity parameters to solve canonical tasks from  
857 behavioral neuroscience.

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

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

$$\begin{aligned} \kappa_r &= G_1(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = M_m \kappa_n + y \\ \kappa_n &= G_2(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = M_n \langle [\phi_i] \rangle + \langle [\phi'_i] \rangle \\ \frac{\Delta_0^2 - \Delta_\infty^2}{2} &= G_3(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = g^2 \left( \int \mathcal{D}z \Phi^2(\kappa_r + \sqrt{\Delta_0} z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\kappa_r + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right) \\ &\quad + (\kappa_n^2 + 1)(\Delta_0 - \Delta_\infty) \\ \Delta_\infty &= G_4(\kappa_r, \kappa_n, \Delta_0, \Delta_\infty) = g^2 \int \mathcal{D}z \left[ \int \mathcal{D}x \phi(\kappa_r + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 + \kappa_n^2 + 1 \end{aligned} \tag{79}$$

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

$$\begin{aligned} l(t) &= \frac{\Delta_0(t)^2 - \Delta_\infty(t)^2}{2} \\ \Delta_0(t) &= \sqrt{2x(t) + \Delta_\infty(t)^2} \\ \frac{d\kappa_r(t)}{dt} &= -\kappa_r(t) + F(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\ \frac{d\kappa_n(t)}{dt} &= -\kappa_n + G(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\ \frac{dI(t)}{dt} &= -l(t) + H(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \\ \frac{d\Delta_\infty(t)}{dt} &= -\Delta_\infty(t) + L(\kappa_r(t), \kappa_n(t), \Delta_0(t), \Delta_\infty(t)) \end{aligned} \tag{80}$$

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

$$\Delta_T = \Delta_0 - \Delta_\infty \quad (81)$$

871 **A.3 Supplementary Figures**

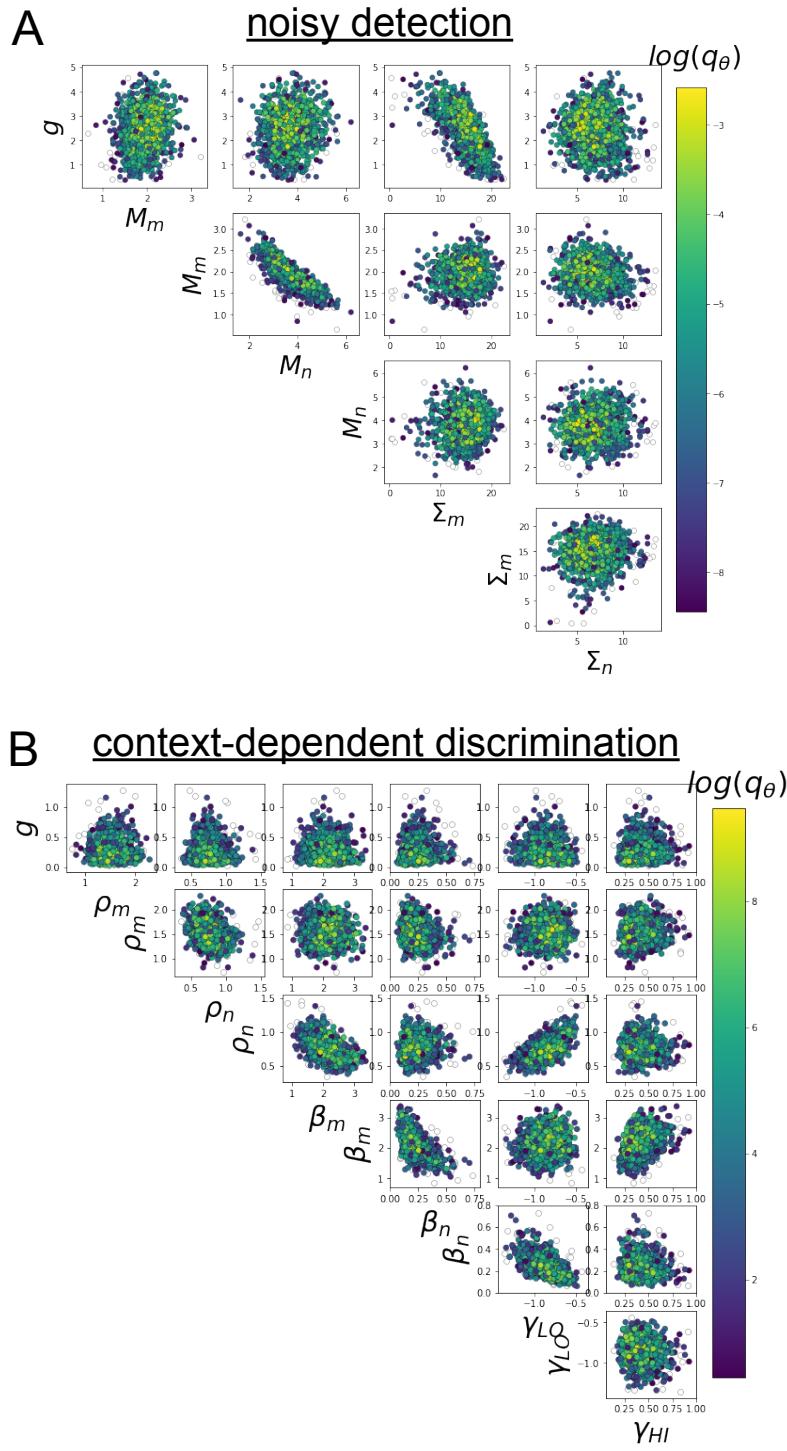


Fig. S4: A. EPI for rank-1 networks doing discrimination. B. EPI for rank-2 networks doing context-dependent discrimination.