

Interrogating theoretical models of neural computation with deep inference

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

¹ 1 Abstract

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

²¹ 2 Introduction

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

27 when configured appropriately, give rise to a measurable signature of a computation. The work of
28 analyzing a model then becomes the inverse problem: given a computation of interest, how can we
29 reason about these suitable parameter configurations – their likely values, their uniquenesses and
30 degeneracies, their attractor states and phase transitions, and more?

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

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

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

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

67 To address these three challenges, we developed an inference methodology – ‘emergent property
68 inference’ – which learns a distribution over parameter configurations in a theoretical model. Crit-
69 ically, this distribution is such that draws from the distribution (parameter configurations) corre-
70 spond to systems of equations that give rise to a specified emergent property. First, we stipulate a
71 bijective deep neural network that induces a flexible family of probability distributions over model
72 parameterizations with a probability density we can calculate [15, ?, ?]. Second, we quantify the
73 notion of emergent properties as a set of moment constraints on datasets generated by the model.
74 Thus, an emergent property is not a single data realization, but a phenomenon or a feature of the
75 model, which is the central object of interest to the theorist (unlike say the statistical neurosci-
76 entist). Conditioning on an emergent property requires a variant of deep probabilistic inference
77 methods, which we have previously introduced [16]. Third, because we cannot assume the theo-
78 retical model has explicit likelihood on data or the emergent property of interest, we use stochas-
79 tic gradient techniques in the spirit of likelihood free variational inference [17]. Taken together,
80 emergent property inference (EPI) provides a methodology for inferring and then reasoning about
81 parameter configurations that give rise to particular emergent phenomena in theoretical models.
82 To clarify the technical details of EPI, we use it to analyze network syncing in a classic model of
83 the stomatogastric ganglion [18].

84 Equipped with this methodology, we then investigated three models of current importance in theo-
85 retical neuroscience. These models were chosen to demonstrate generality through ranges of biolog-
86 ical realism (conductance-based biophysics to recurrent neural networks), neural system function
87 (pattern generation to abstract cognitive function), and network scale (four to infinite neurons).
88 First, we use EPI to produce a set of verifiable hypotheses of input-responsivity in a four neuron-
89 type dynamical model of primary visual cortex; we then validate these hypotheses in the model.
90 Second, we demonstrated how the systematic application of EPI to levels of task performance can

91 generate experimentally testable hypotheses regarding connectivity in superior colliculus. Third,
 92 we use EPI to uncover the sources of bias in a low-rank recurrent neural network executing a toy
 93 mathematical computation. The novel scientific insights offered by EPI contextualize and clarify
 94 the previous studies exploring these models [18, 19, 20, 21] and more generally offer a quantitative
 95 grounding for theoretical models going forward, pointing a way to how rigorous statistical inference
 96 can enhance theoretical neuroscience at large.

97 We note that, during our preparation and early presentation of this work [22, 23], another work
 98 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
 99 circuits [24]. We are excited by this broad problem being recognized by the community, and we
 100 emphasize that these works offer complementary neuroscientific contributions and use different
 101 technical methodologies. Scientifically, our work has focused primarily on systems-level theoretical
 102 models, while their focus has been on lower-level cellular models. Secondly, there are several key
 103 technical differences in the approaches (see Section A.1.4) perhaps most notably is our focus on
 104 the emergent property – the measurable signal of the computation in question, vs their focus
 105 on observed datasets; both certainly are worthy pursuits. The existence of these complementary
 106 methodologies emphasizes the increased importance and timeliness of both works.

107 3 Results

108 3.1 Motivating emergent property inference of theoretical models

109 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-
 110 erty inference. First, the theorist designs or chooses an existing model that, it is hypothesized,
 111 captures the computation of interest. To ground this process in a well-known example, consider
 112 the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple
 113 rhythmic muscle activation patterns for digestion [25]. A model of the STG [18] is shown schemat-
 114 ically in Figure 1A, and note that the behavior of this model will be critically dependent on its
 115 parameterization – the choices of conductance parameters $z = [g_{el}, g_{synA}]$. Specifically, the two
 116 fast neurons (f_1 and f_2) mutually inhibit one another, and oscillate at a faster frequency than
 117 the mutually inhibiting slow neurons (s_1 and s_2), and the hub neuron (hub) couples with the fast
 118 or slow population or both. Second, once the model is selected, the theorist defines the emergent
 119 property, the measurable signal of scientific interest. To continue our running STG example, one
 120 such emergent property is the phenomenon of *network syncing* – in certain parameter regimes, the

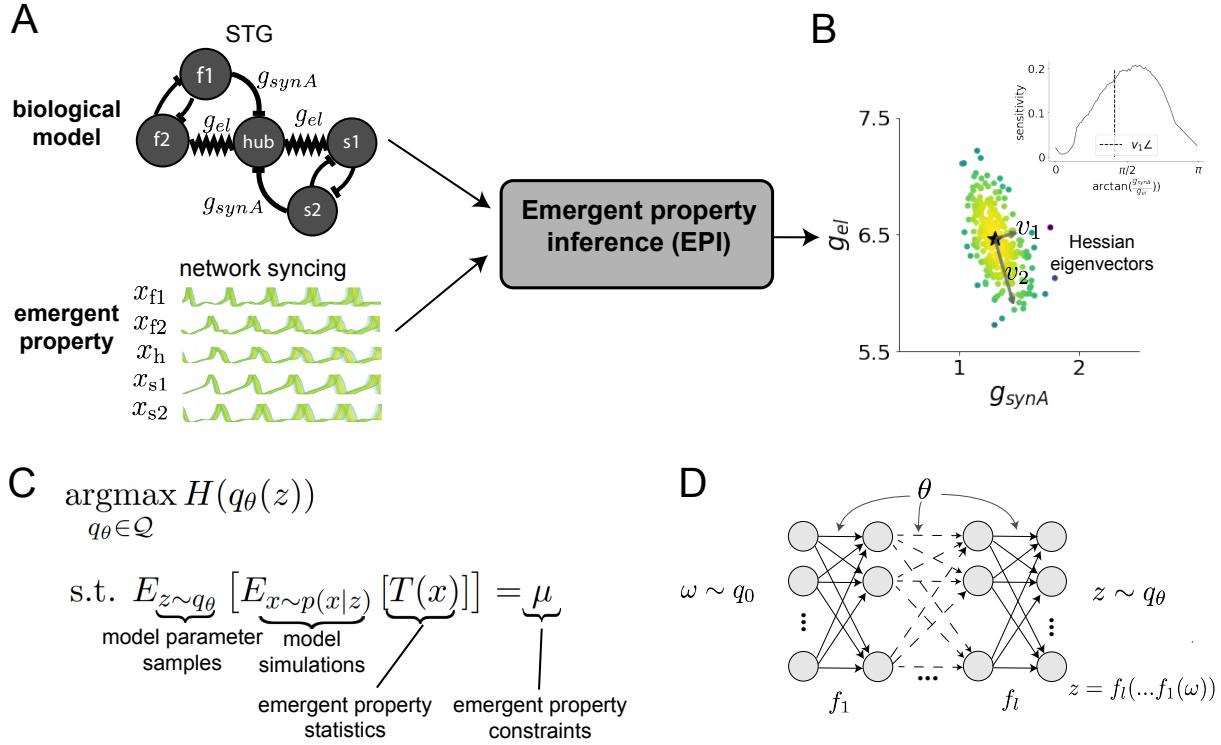


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. A. For a choice of model (STG) and emergent property (network syncing), emergent property inference (EPI) learns a posterior distribution of the model parameters $z = [g_{el}, g_{synA}]^\top$ conditioned on network syncing. B. An EPI distribution of STG model parameters producing network syncing. The eigenvectors of the Hessian at the mode of the inferred distribution are indicated as v_1 and v_2 . (Inset) Sensitivity of the system with respect to network syncing along all dimensions of parameter space away from the mode. (see Section A.2.1). C. EPI learns a distribution $q_\theta(z)$ of model parameters that produce an emergent property: the emergent property statistics $T(x)$ are fixed in expectation over parameter distribution samples $z \sim q_\theta(z)$ to particular values μ . EPI distributions maximize randomness via entropy, although other measures are sensible. D. Deep probability distributions map a latent random variable $\omega \sim q_0$, where q_0 is chosen to be simple distribution such as an isotropic gaussian, through a highly expressive function family $f_\theta(\omega) = f_l(\dots f_1(\omega))$ parameterized by the neural network weights and biases $\theta \in \Theta$. This mapping induces an implicit probability model $q(g_\theta(\omega)) \in \mathcal{Q}$

frequency of the hub neuron matches that of the fast and slow populations at an intermediate frequency. This emergent property is shown in Figure 1A at a frequency of 0.55Hz. Third, qualitative parameter analysis ensues: since precise mathematical analysis is intractable in this model, a brute force sweep of parameters is done. Subsequently, a qualitative description is formulated to describe of the different parameter configurations that lead to the emergent property. In this last step lies the opportunity for a precise quantification of the emergent property as a statistical feature of the model. Equipped with this methodology, we can infer a probability distribution over parameter configurations that produce this quantified emergent property.

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

3.2 A deep generative modeling approach to emergent property inference

Emergent property inference (EPI) systematizes the three-step procedure of the previous section. First, we consider the model as a coupled set of differential (and potentially stochastic) equations [18]. In the running STG example, the dynamical state $x = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$ is the membrane potential for each neuron, which evolves according to the biophysical conductance-based equation:

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

150 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
 151 and $z = [g_{\text{el}}, g_{\text{synA}}]$ (see Section A.2.1).

153 Second, we define the emergent property, which as above is network syncing: the phase locking
 154 of the population and its oscillation at an intermediate frequency of our choosing (Figure 1A
 155 bottom). Quantifying this phenomenon is straightforward: we define network syncing to be that
 156 the spiking frequency of each neuron is close to an intermediate frequency of 0.55Hz. Thus, our
 157 measurable signature of computation – the firing frequencies of each neuron $\omega_{f1}(x)$, $\omega_{f2}(x)$, etc.– are
 158 statistics of the membrane potential activity x which we insist be near a particular value 0.55Hz.
 159 Mathematically, we achieve this via constraints on the mean and variance of $\omega_i(x)$ for each neuron
 160 $i \in \{f1, f2, \text{hub}, s1, s2\}$, and thus:

$$E[T(x)] \triangleq E \begin{bmatrix} \omega_{f1}(x) \\ \vdots \\ (\omega_{f1}(x) - 0.55)^2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0.55 \\ \vdots \\ 0.025^2 \\ \vdots \end{bmatrix} \triangleq \mu, \quad (2)$$

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 model of
 166 the normalizing flow class [15, ?, ?] – neural networks which transform a simple distribution into a
 167 suitably complicated distribution (as is needed here). This deep model is represented in Figure 1E
 168 (and see Methods for more detail). Then, mathematically, we must solve the following optimization
 169 program:

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

170 where $T(x), \mu$ are defined as in Equation 3. The purpose of each element in this program is detailed
 171 in Figure 1D. Finally, we recognize that many distributions in \mathcal{Q} will respect the emergent property
 172 constraints, so we require a normative principle to select amongst them. This principle is captured
 173 in Equation 3 by the primal objective H . Here we chose Shannon entropy to model parameter
 174 distributions with minimal assumptions beyond some chosen structure [26, 27, 16, 28], but we

emphasize that the EPI method is unaffected by this choice (the results of course will depend on this choice). Stating such a problem is easy enough; finding a tractable and suitably flexible family of probability distributions (\mathcal{Q}) is hard.

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

3.3 Comprehensive input-responsivity in a nonlinear sensory system

In studies of primary visual cortex (V1), theoretical models with two populations (excitatory (E) and inhibitory (I) neurons) have reproduced a host of experimentally documented phenomena. In particular regimes of excitation and inhibition, these E/I models exhibit the paradoxical effect [4], selective amplification [29], surround suppression [30], and sensory integrative properties [31]. Extending this using experimental evidence of three genetically-defined classes of inhibitory neurons [32, 33], recent work [19] has investigated a four-population model – excitatory (E), parvalbumin (P), somatostatin (S), and vasointestinal peptide (V) neurons – as shown in Fig. 2A. The dynamical state of this model is the firing rate of each neuron-type population $x = [x_E, x_P, x_S, x_V]^\top$, which evolves according to rectified and exponentiated dynamics:

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

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

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

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

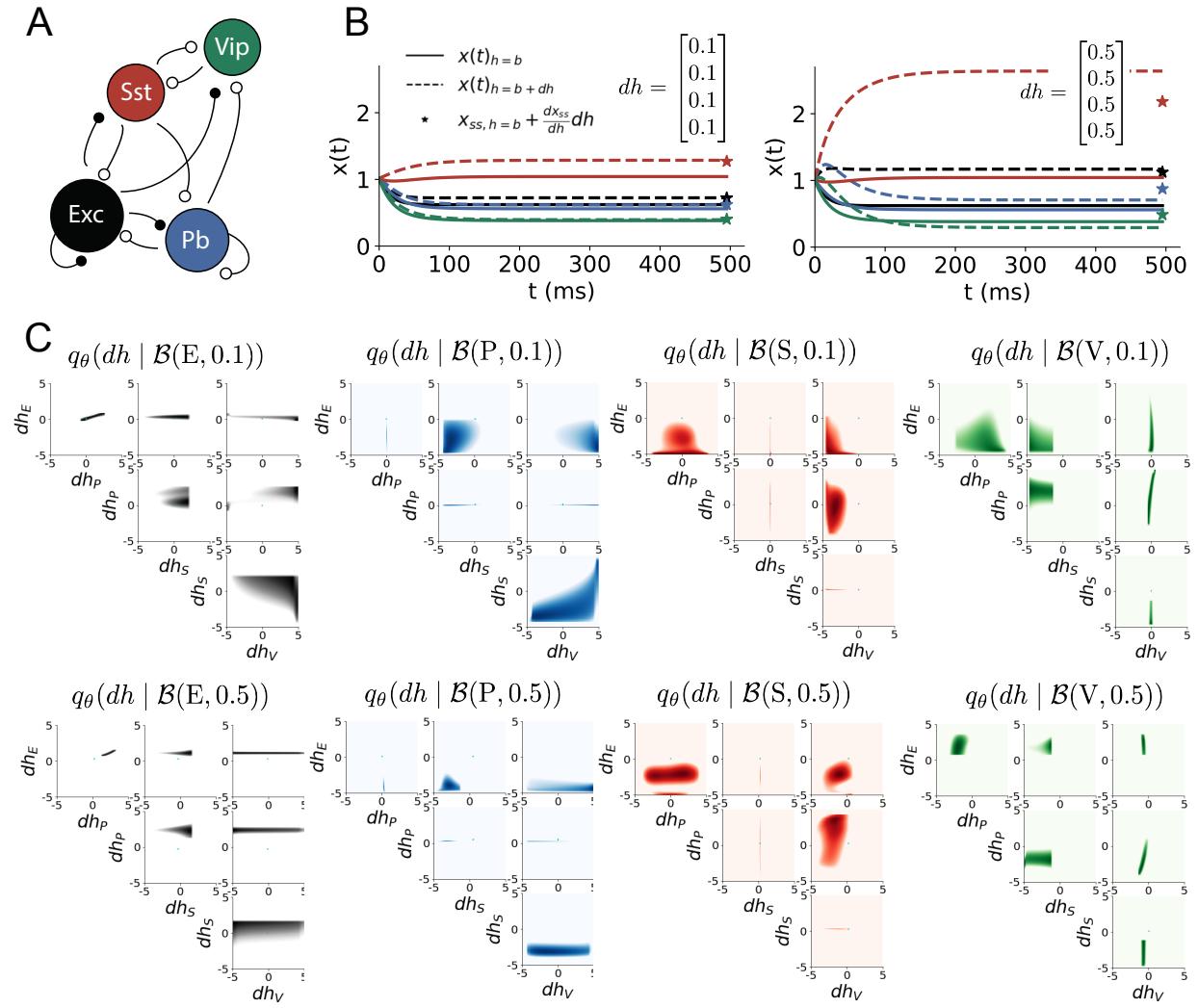


Figure 2: Exploring neuron-type responsivity in V1. A. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and vip (green) neurons. Some neuron-types largely do not form synaptic projections to others (excitatory and inhibitory projections filled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

Having established our model, we now turn to defining the emergent property. We begin with the linearized response of the system $\frac{dx_{ss}}{dh}$ at fixed points x_{ss} . While this linearization accurately predicts differential responses $dx_{ss} = [dx_{E,ss} \ dx_{P,ss} \ dx_{S,ss} \ dx_{V,ss}]$ for small differential inputs to each population $dh = [0.1 \ 0.1 \ 0.1 \ 0.1]$ (Fig. 2B, left), it can be misleading in such a nonlinear model for a greater differential input strength $dh = [0.1 \ 0.1 \ 0.1 \ 0.1]$ (Fig. 3B, right). In fact, the linearly predicted response for the V-population to $dh = [0.5 \ 0.5 \ 0.5 \ 0.5]$ was actually in the opposite direction of the true response (Fig. 2B, right, green). This shows that currently available approaches to deriving the steady state response of this system are limited.

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

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

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

We evaluate these hypotheses by taking steps in individual neuron-type input Δh_α away from the modes of the inferred distributions

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

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

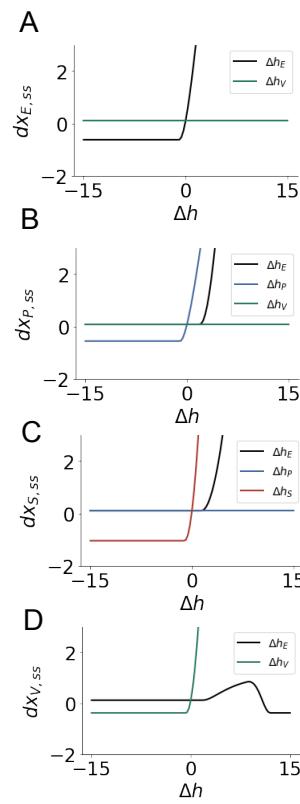


Figure 3: Confirming EPI generated hypotheses in V1. A. Differential responses by the E-population to changes in individual input $\Delta h_\alpha u_\alpha$ away from the mode of the EPI distribution dh^* . B-D Same plots for the P-, S-, and V-populations for the inputs for which hypotheses were formulated.

232 To this point, we have shown the utility of EPI on relatively low-level emergent properties like
 233 network syncing and differential neuron-type population responses. In the remainder of the study,
 234 we focus on using EPI to understand models of more abstract cognitive function.

235 3.4 Identifying neural mechanisms of behavioral learning.

236 Identifying measurable biological changes that result in improved behavior is important for neuro-
 237 science, since they may indicate how the learning brain adapts. In a rapid task switching exper-
 238 iment [?], where rats were to respond right (R) or left (L) to the side of a light stimulus in the
 239 pro (P) task, and oppositely in the anti (A) task predicated by an auditory cue (Fig. 3A), neural
 240 recordings exhibited two population of neurons in each hemisphere of superior colliculus (SC) that
 241 simultaneously represented both task condition and motor response: the Pro/contralateral and
 242 Anti/ipsilateral neurons [20]. Duan et al. proposed a model of SC that, like the V1 model analyzed
 243 in the previous section, is a four-population dynamical system. Here, the neuron-type populations
 244 are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right
 245 (R)). The Pro- or Anti-populations receive an input determined by the cue, and then the left and
 246 right populations receive an input based on the side of the light stimulus. Activities were bounded

from 0-1, so that a high output (1) of the Pro population in a given hemisphere corresponds to the contralateral response. An additional stipulation is that when one Pro population responds with a high-output, the opposite Pro population must respond with a low output (0). Finally, this circuit operates in the presence of gaussian noise resulting in trial-to-trial variability (see Section A.2.3). The connectivity matrix is parameterized by the geometry of the population arrangement (Fig. 3B).

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

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

3.5 Characterizing the sources of bias in RNN computation

So far, each biologically realistic model we have studied was designed from fundamental biophysical principles, genetically- or functionally-defined neuron types. At a more abstract level of modeling,

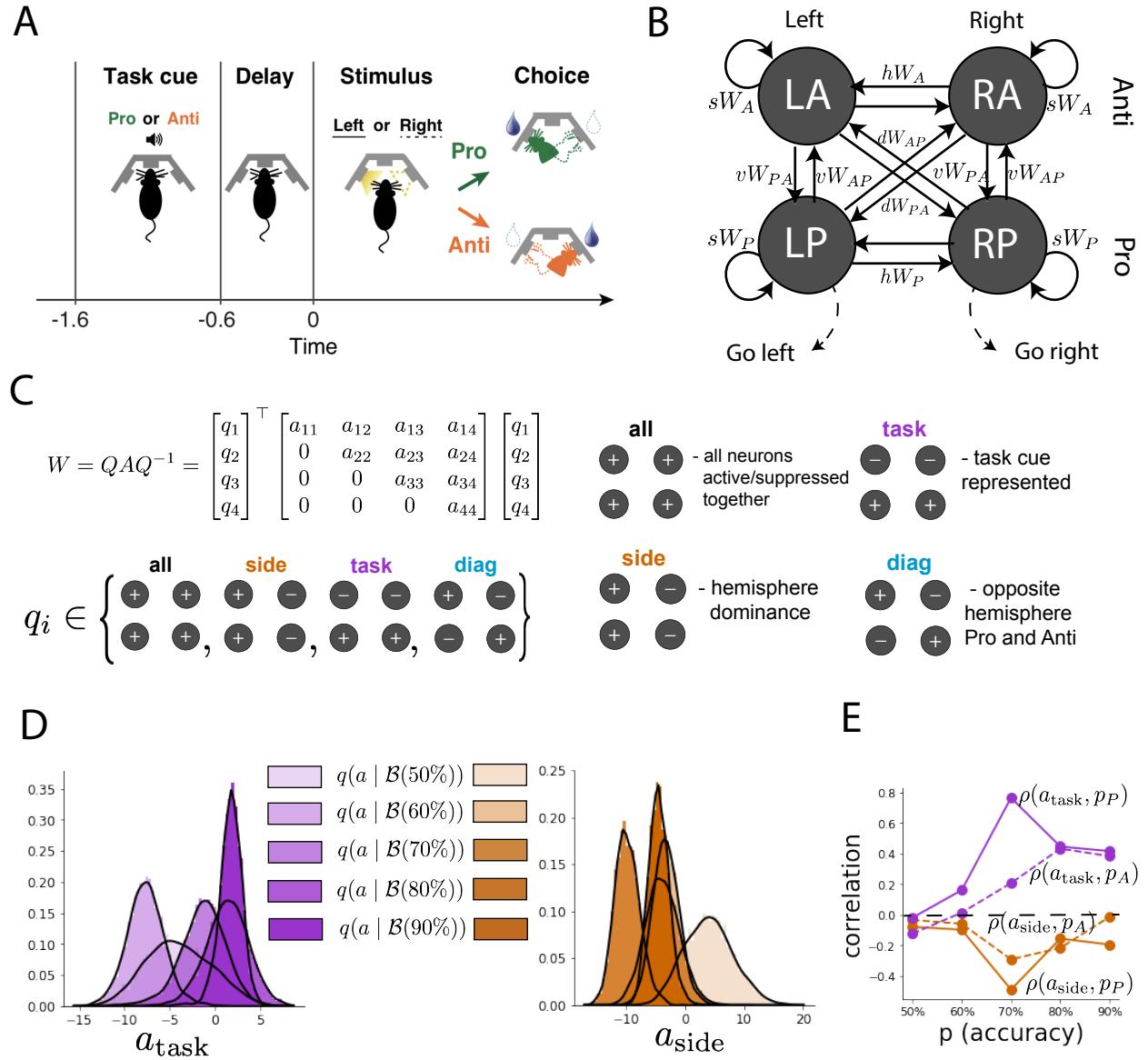


Figure 4: EPI reveals changes in SC [20] connectivity that result in greater task accuracy. A. Rapid task switching behavioral paradigm. In the Pro (Anti) condition indicated by an auditory cue, rats respond by poking into a side port to the same (opposite) side as the light stimulus that is provided after a delay to receive a reward. B. Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. C. The Schur decomposition of the weight matrix $W = QAQ^{-1}$ is a unique decomposition with orthogonal Q and upper triangular A . The invariant Schur eigenmodes (symmetry of W) are labeled by their hypothesized role in computation: q_{all} , q_{task} , q_{side} , and q_{diag} . The values of A are what change for different realizations of W . D. The marginal EPI distributions of the Schur eigenvalues at each level of task accuracy. E. The correlation of Schur eigenvalue with task performance in each learned EPI distribution.

278 recurrent neural networks (RNNs) are high-dimensional models of computation, which have become
 279 increasingly popular in neuroscience research [34]. Typically, RNNs are trained to do a task from a
 280 systems neuroscience experiment, and then the unit activations of the trained RNN are compared
 281 to recorded neural activity. A monumental challenge for this line of work is to link findings at
 282 this level of abstraction with interpretable biophysical mechanisms in the brain. Here we leverage
 283 recent theoretical work to run EPI on interpretable parameterizations of RNN connectivity solving
 284 a toy problem.

285 Importantly, recent work establishes such a link between RNN connectivity weights and the re-
 286 sulting dynamical responses of the network using dynamic mean field theory (DMFT) for neural
 287 networks [3]. Specifically, DMFT describes the properties of activity in infinite-size neural networks
 288 given a distribution on the connectivity weights. This theory has been extended from random neural
 289 networks to low rank RNNs, which have low-dimensional parameterizations of RNN connectivity
 290 via the pairwise correlations of the low-rank vectors (i.e. the low-rank “geometry”) [21]. For
 291 example, the connectivity of a rank-1 RNN J is the sum of a random component with strength
 292 determined by g and a structured component determined by the outer product of vectors m and n :

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

293 where the activity x evolves as

$$\frac{\partial x}{\partial t} = -x(t) + J\phi(x(t)) + I(t) \quad (9)$$

294 $I(t)$ is some input, ϕ is the tanh nonlinearity, and $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$. The entries of m and n are
 295 drawn from gaussian distributions $m_i \sim \mathcal{N}(M_m, 1)$ and $n_i \sim \mathcal{N}(M_n, 1)$, whose parameters M_m and
 296 M_n determine their degree of correlation.

297 Mastrogiuseppe et al. are able to design low-rank connectivities via the pairwise correlations of such
 298 low-rank vectors that solve tasks from behavioral neuroscience. An important detail is that a non-
 299 linear system of equations solver must be used to obtain the task-relevant variables of interest from
 300 the derived consistency equations (see Section A.2.4). However, we can consider the DMFT equa-
 301 tion solver as a black box that takes in a low-rank parameterization z (e.g. $z = [g \ M_m \ M_n]$) and
 302 outputs task-relevant response variables (e.g. average network activity μ , the temporal variability
 303 in the network Δ_T , or network activity along a given dimension κ). Furthermore, we recognize that
 304 the solution produced by the solver is differentiable with respect to the input parameters. Thus, we
 305 are able to combine this DMFT with EPI to learn distributions on such connectivity parameters

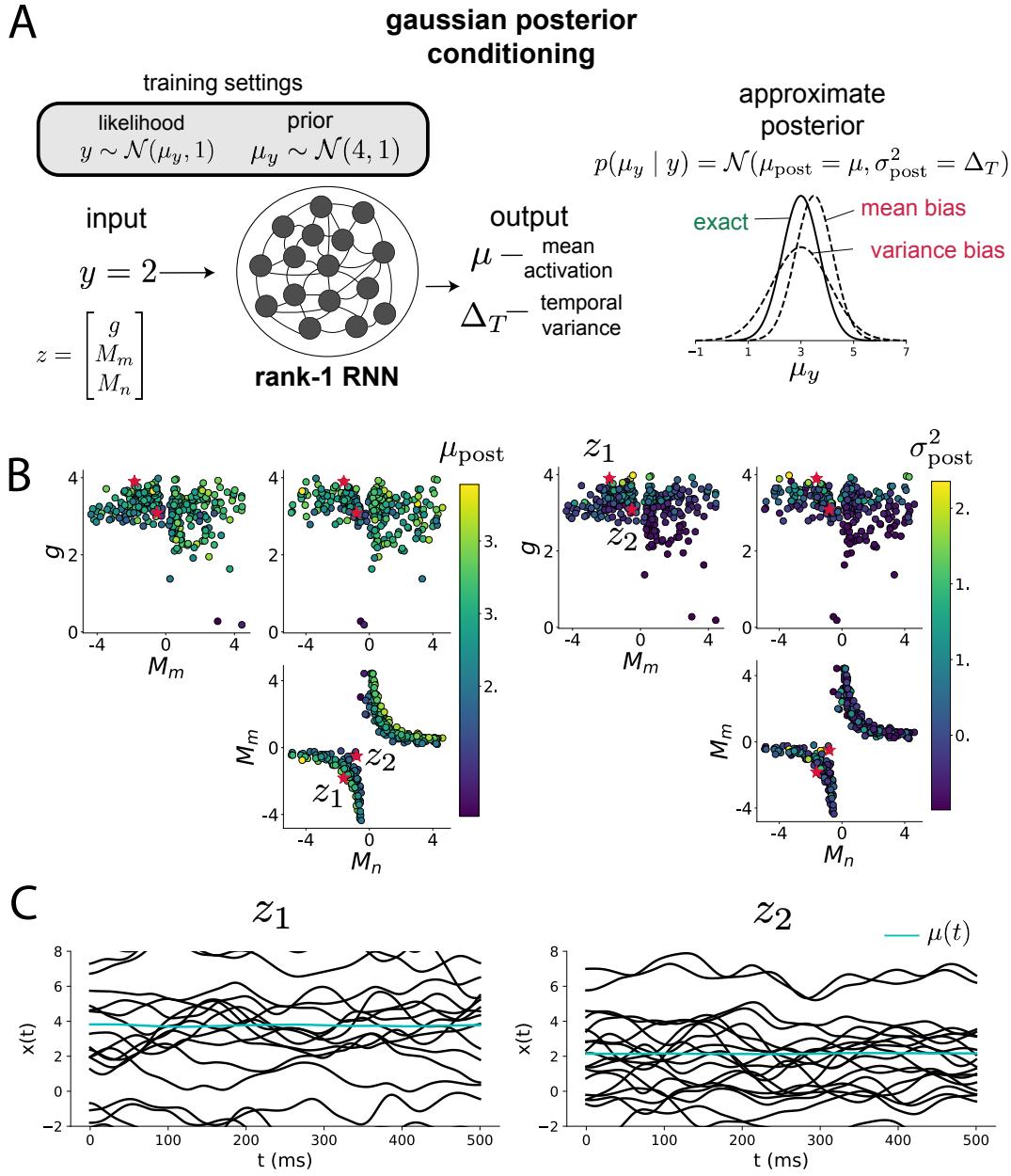


Figure 5: Sources of bias in RNN computation. A. (left) A rank-1 RNN running approximate Bayesian inference on μ_y assuming a gaussian likelihood variance of 1 and a prior of $\mathcal{N}(4, 1)$. (center) The rank-1 RNN represents the computed gaussian posterior mean μ_{post} and variance σ_{post}^2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

306 of RNNs that execute neuroscientific tasks via an emergent property defined on the task-relevant
 307 responses produced by DMFT.

308 For our toy problem, we consider the emergent property of gaussian posterior conditioning. Specif-
 309 ically, we ask the RNN to calculate the parameters of a gaussian posterior distribution on the mean
 310 of a gaussian likelihood μ_y given a single observation of y and a gaussian prior $p(\mu_y) = \mathcal{N}(4, 1)$
 311 (Fig. 5A). Assuming the variance of the gaussian likelihood is 1, the true posterior for an input of
 312 $y = 2$ is $p(\mu_y | y) = \mathcal{N}(3, 0.5)$. We used EPI to learn distributions of RNNs producing the correct
 313 posterior mean and variance in their mean activity $\mu = \mu_{\text{post}}$ and temporal variance $\Delta_T = \sigma_{\text{post}}^2$
 314 given an input of $y = 2$. (see Section A.2.4) (Fig. 5B).

315 When specifying the emergent property of gaussian posterior conditioning, we allowed a substan-
 316 tial amount of variability in the second moment constraints of the network mean μ and temporal
 317 variance Δ_T . This resulted in a distribution of rank-1 RNN parameterizations having a wide vari-
 318 ety biases in the resulting μ_{post} and σ_{post}^2 (under- or over-estimations of the posterior means and
 319 variances). We can examine the nature of the biases in this toy computation by visualizing the
 320 produced posterior means (Fig. 5B, left) and variances (Fig. 5B, right) in the inferred distribution.
 321 The inferred distribution has roughly symmetric in the M_m - M_n plane, suggesting there is a degen-
 322 eracy in the product of M_m and M_n (Fig. 5B). The product of M_m and M_n almost completely
 323 determines the posterior mean (Fig. 5B, left), and the random strength g is the most influential
 324 variable on the temporal variance (Fig. 5B, right). Neither of these observations were obvious from
 325 the consistency equations afforded by DMFT (see Section A.2.4).

326 When working with DMFT, it's important to check that finite-size realizations of these infinite-
 327 size networks match the theoretical predictions. We check 2,000-neuron realizations of drawn
 328 parameters z_1 and z_2 from the inferred distribution. z_1 has relatively high g and high $M_m M_n$,
 329 whereas z_2 has relatively low g and low $M_m M_n$. Confirming our intuition, z_1 overestimates the
 330 posterior mean, since mean activity $\mu(t) > 3$ (Fig. 5C, left cyan). In turn, z_2 underestimates the
 331 posterior mean, since $\mu(t) < 3$ (Fig. 5C, right cyan). Finally, z_1 results in evidently greater temporal
 332 variance than z_2 . This novel procedure of doing inference in interpretable parameterizations of
 333 RNNs conditioned on task execution is straightforwardly generalizable to other tasks like noisy
 334 integration and context-dependent decision making (Fig. S1).

335 4 Discussion**336 4.1 EPI is a general tool for theoretical neuroscience.**

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

347 In this study, we have focused on applying EPI to low dimensional parameter spaces of models
348 with low dimensional dynamical state. These choices were made to present the reader with a series
349 of interpretable conclusions, which is more challenging in high dimensional spaces. In fact, EPI
350 should scale reasonably to high dimensional parameter spaces, as the underlying technology has
351 produced state-of-the-art performance on high-dimensional tasks such as texture generation [16].
352 Of course, increasing the dimensionality of the dynamical state of the model makes optimization
353 more expensive, and there is a practical limit there as with any machine learning approach. For
354 systems with high dimensional state, we recommend using theoretical approaches (e.g. [21]) to
355 reason about reduced parameterizations of such high-dimensional systems.

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

366 (e.g. Section A.2.1). In such cases, optimization will be expensive.

367 **4.2 Novel hypotheses from EPI**

368 Machine learning has played an effective, multifaceted role in neuroscientific progress. Primarily,
369 it has revealed structure in large-scale neural datasets [35, 36, 37, 38, 39, 40] (see review, [14]).
370 Secondarily, trained algorithms of varying degrees of biological relevance are beginning to be viewed
371 as fully-observable computational systems comparable to the brain [41, 42]. Theorists can use deep
372 learning for probabilistic inference to understand their models and their behavior.

373 For example, consider the fact that we do not yet understand just a four-dimensional, deterministic
374 model of V1 [19]. This should not be surprising, since analytic approaches to studying nonlinear
375 dynamical systems become increasingly complicated when stepping from two-dimensional to three-
376 or four-dimensional systems in the absence of restrictive simplifying assumptions [43]. We promote
377 the recognition of analytic difficulty, and alternatively the use of EPI to gain the desired model
378 insights. In Section 3.3, we showed that EPI was far more informative about neuron-type input
379 responsibility than the predictions afforded through analysis. By flexibly conditioning this V1 model
380 on different emergent properties, we performed an exploratory analysis of a *model* rather than a
381 dataset, which generated and proved out a set of testable predictions.

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

389 In our final analysis, we present a novel procedure for doing statistical inference on interpretable
390 parameterizations of RNNs executing tasks from behavioral neuroscience. This methodology relies
391 on recently extended theory of responses in random neural networks with minimal structure [21].
392 With this methodology, we can finally open the probabilistic model selection toolkit reasoning
393 about the connectivity of RNNs solving tasks.

394 References

- 395 [1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 396 [2] John J Hopfield. Neurons with graded response have collective computational properties like
397 those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092,
398 1984.
- 399 [3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
400 networks. *Physical review letters*, 61(3):259, 1988.
- 401 [4] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
402 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
403 17(11):4382–4388, 1997.
- 404 [5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Confer-
405 ence on Learning Representations*, 2014.
- 406 [6] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation
407 and variational inference in deep latent gaussian models. *International Conference on Machine
408 Learning*, 2014.
- 409 [7] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
410 neural population models through nonlinear embeddings. In *Advances in neural information
411 processing systems*, pages 163–171, 2016.
- 412 [8] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear
413 dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 414 [9] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders.
415 *bioRxiv*, page 399246, 2018.
- 416 [10] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
417 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg,
418 et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature
419 methods*, page 1, 2018.
- 420 [11] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
421 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
422 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

- [12] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- [13] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall, Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet: nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural Information Processing Systems*, 2019.
- [14] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241, 2018.
- [15] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 2015.
- [16] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow networks. *International Conference on Learning Representations*, 2017.
- [17] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.
- [18] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. *Neuron*, 77(5):845–858, 2013.
- [19] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*, 115(3):1399–1409, 2016.
- [20] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor routing. *bioRxiv*, page 245613, 2018.
- [21] Francesca Mastrogiovanni and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.

- 451 [22] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
452 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience*
453 *Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 454 [23] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
455 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with
456 degenerate solution networks. *Bernstein Conference*, 2019.
- 457 [24] Jan-Matthis Lueckmann, Pedro Goncalves, Chaitanya Chintaluri, William F Podlaski, Gia-
458 como Bassetto, Tim P Vogels, and Jakob H Macke. Amortised inference for mechanistic models
459 of neural dynamics. In *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon,*
460 *Portugal*, 2019.
- 461 [25] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
462 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 463 [26] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
464 1957.
- 465 [27] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
466 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 467 [28] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
468 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 469 [29] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
470 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 471 [30] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
472 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
473 592, 2009.
- 474 [31] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
475 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,
476 85(2):402–417, 2015.
- 477 [32] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
478 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
479 5(10):793, 2004.

- [33] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neurobiology*, 71(1):45–61, 2011.
- [34] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, 46:1–6, 2017.
- [35] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*, 13(8):1713–1720, 2001.
- [36] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–7425, 1998.
- [37] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- [38] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.
- [39] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.
- [40] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- [41] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.
- [42] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 2019.
- [43] Steven H Strogatz. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (Studies in Nonlinearity), Perseus, Cambridge, UK, 1994.

- 509 [44] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial
510 Intelligence and Statistics*, pages 814–822, 2014.
- 511 [45] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
512 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 513 [46] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
514 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 515 [47] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
516 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
517 interneurons. *Nature neuroscience*, 16(8):1068, 2013.

518 **A Methods**

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

520 Emergent property inference (EPI) learns distributions of theoretical model parameters that pro-
521 duce emergent properties of interest. EPI combines ideas from likelihood-free variational inference
522 [17] and maximum entropy flow networks [16]. A maximum entropy flow network is used as a deep
523 probability distribution for the parameters, while these samples often parameterize a differentiable
524 model simulator, which may lack a tractable likelihood function.

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

$$\mathcal{B} \leftrightarrow E_{z \sim q_\theta} [E_{x \sim p(x|z)} [T(x)]] = \mu \quad (10)$$

530 by making an approximation $q_\theta(z)$ to $p(z | \mathcal{B})$ (see Section A.1.5). So, over the DSN distribution
531 $q_\theta(z)$ of model $p(x | z)$ for behavior \mathcal{B} , the emergent properties $T(x)$ are constrained in expectation
532 to μ .

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

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

538 In most settings (especially those relevant to theoretical neuroscience) the likelihood of the behavior
 539 with respect to the model parameters $p(T(x) | z)$ is unknown or intractable, requiring an alternative
 540 to stochastic gradient variational Bayes [5] or black box variational inference[44]. These types
 541 of methods called likelihood-free variational inference (LFVI, [17]) skate around the intractable
 542 likelihood function in situations where there is a differentiable simulator. Akin to LFVI, DSNs are
 543 optimized with the following objective for a given theoretical model, emergent property statistics
 544 $T(x)$, and emergent property constraints μ :

$$\begin{aligned} q_\theta^*(z) &= \underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta(z)) \\ \text{s.t. } E_{z \sim q_\theta} [E_{x \sim p(x|z)} [T(x)]] &= \mu \end{aligned} \tag{11}$$

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

548 A.1.1 Example: 2D LDS

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

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

550 that produce a band of oscillations. To do EPI with the dynamics matrix elements as the free
 551 parameters $z = [a_1, a_2, a_3, a_4]$, and fixing $\tau = 1$, such that the posterior yields a band of oscillations,
 552 the emergent property statistics $T(x)$ are chosen to contain the first- and second-moments of the
 553 oscillatory frequency Ω and the growth/decay factor d of the oscillating system. To learn the
 554 distribution of real entries of A that yield a distribution of d with mean zero with variance 0.25^2 ,
 555 and oscillation frequency Ω with mean 1 Hz with variance $(0.1\text{Hz})^2$, then we would select the real
 556 part of the complex conjugate eigenvalues $\text{real}(\lambda_1) = d$ (via an arbitrary choice of eigenvalue of the
 557 dynamics matrix λ_1) and the positive imaginary component of one of the eigenvalues $\text{imag}(\lambda_1) =$
 558 $2\pi\Omega$ as the emergent property statistics. Those emergent property statistics are then constrained

559 to

$$\mu = E \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\Omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\Omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \quad (12)$$

560 where $\Omega = 1\text{Hz}$. Unlike the models we study in the paper which calculate $E_{x \sim p(x|z)} [T(x)]$ via
 561 forward simulation, we have a closed form for the eigenvalues of the dynamics matrix. λ can be
 562 calculated using the quadratic formula:

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

563 where λ_1 is the eigenvalue of $\frac{1}{\tau}A$ with greatest real part. Even though $E_{x \sim p(x|z)} [T(x)]$ is calculable
 564 directly via a closed form function and does not require simulation, we cannot derive the distribution
 565 q_θ^* directly. This is due to the formally hard problem of the backward mapping: finding the natural
 566 parameters η from the mean parameters μ of an exponential family distribution [45]. Instead, we
 567 can use EPI to learn the linear system parameters producing such a band of oscillations (Fig. S2B).

568 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the
 569 parameter distribution. To validate our method (further than that of the underlying technology
 570 on a ground truth solution [16]) we can analytically derive the contours of the probability density
 571 from the emergent property statistics and values (Fig. S3). In the $a_1 - a_4$ plane, is a black line
 572 at $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, a dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 1$, and a
 573 grey line at twice the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 2$ (Fig. S3A). Here the lines denote the
 574 set of solutions at fixed behaviors, which overlay the posterior obtained through EPI. The learned
 575 DSN distribution precisely reflects the desired statistical constraints and model degeneracy in the
 576 sum of a_1 and a_4 . Intuitively, the parameters equivalent with respect to emergent property statistic
 577 $\text{real}(\lambda_1)$ have similar log densities.

578 To explain the structure in the bimodality of the DSN posterior, we can look at the imaginary
 579 component of λ_1 . When $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, we have

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

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

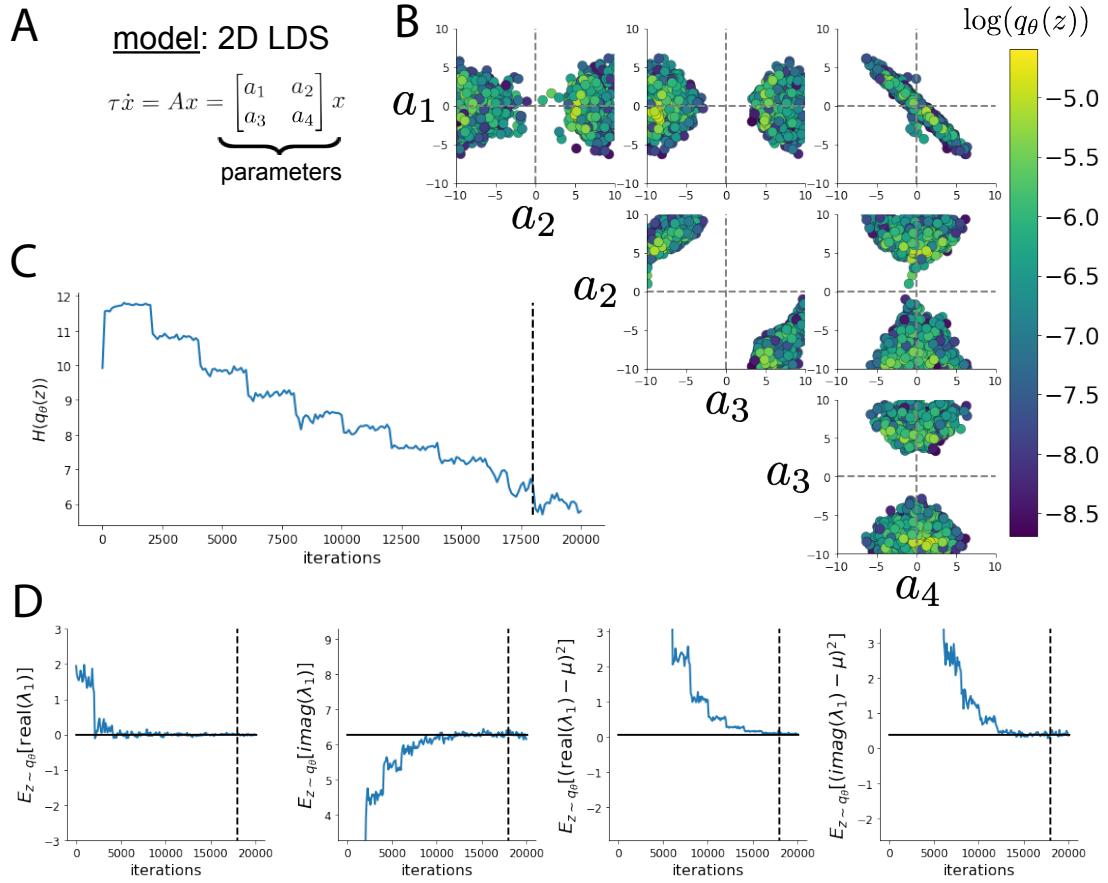


Fig. S2: A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The DSN distribution for a 2D LDS with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. C. Entropy throughout the optimization. At the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

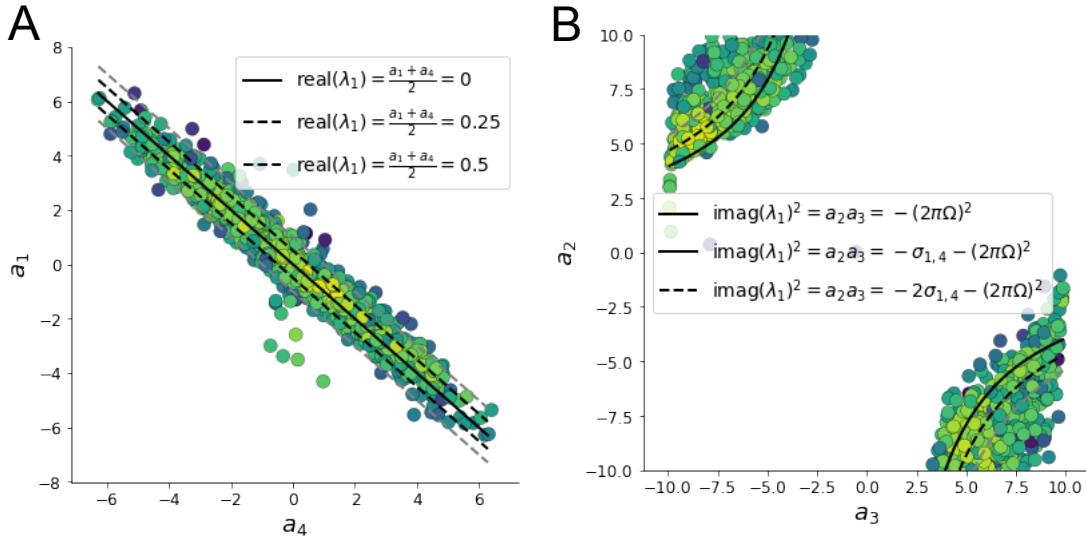


Fig. S3: A. Probability contours in the $a_1 - a_4$ plane can be derived from the relationship to emergent property statistic of growth/decay factor. B. Probability contours in the $a_2 - a_3$ plane can be derived from relationship to the emergent property statistic of oscillation frequency.

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

582 Since we constrained $E_{q_\theta} [\text{imag}(\lambda)] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation
 583 $\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 = (2\pi)^2$ for various $a_1 a_4$ (Fig. S3A). If $\sigma_{1,4} = E_{q_\theta} (|a_1 a_4 - E_{q_\theta}[a_1 a_4]|)$,
 584 then we plot the contours as $a_1 a_4 = 0$ (black), $a_1 a_4 = -\sigma_{1,4}$ (black dotted), and $a_1 a_4 = -2\sigma_{1,4}$
 585 (grey dotted) (Fig. S3B). This validates the curved structure of the inferred distribution learned
 586 through EPI. We take steps in negative standard deviation of $a_1 a_4$ (dotted and gray lines), since
 587 there are few positive values $a_1 a_4$ in the posterior. Subtler model-behavior combinations will have
 588 even more complexity, further motivating the use of EPI for understanding these systems. Indeed,
 589 we sample a distribution of systems oscillating near 1Hz (Fig. S4).

590 A.1.2 Augmented Lagrangian optimization

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

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

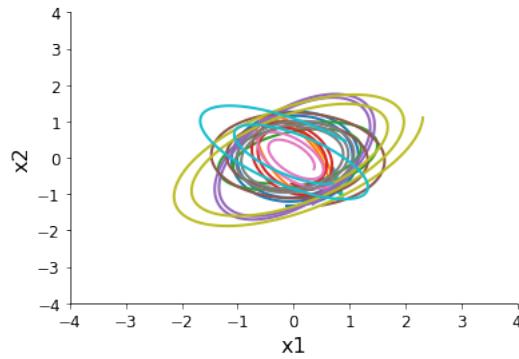


Fig. S4: Sampled dynamical system trajectories from the EPI distribution. Each trajectory is initialized at $x(0) = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$.

593 where $\delta(\theta) = E_{z \sim q_\theta} [E_{x \sim p(x|z)} [T(x) - \mu]]$, $\alpha \in \mathcal{R}^m$ are the Lagrange multipliers and c is the penalty
 594 coefficient. For a fixed (α, c) , θ is optimized with stochastic gradient descent. A low value of c is
 595 used initially, and increased during each augmented Lagrangian epoch – a period of optimization
 596 with fixed α and c for a given number of stochastic optimization iterations. Similarly, α is tuned
 597 each epoch based on the constraint violations. For the linear 2-dimensional system (Fig. S2C)
 598 optimization hyperparameters are initialized to $c_1 = 10^{-4}$ and $\alpha_1 = 0$. The penalty coefficient
 599 is updated based on a hypothesis test regarding the reduction in constraint violation. The p-
 600 value of $E[|\delta(\theta_{k+1})|] > \gamma E[|\delta(\theta_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability
 601 $1 - p$. Throughout the project, $\beta = 4.0$ and $\gamma = 0.25$ is used. The other update rule is $\alpha_{k+1} =$
 602 $\alpha_k + c_k \frac{1}{n} \sum_{i=1}^n (T(x^{(i)}) - \mu)$. In this example, each augmented Lagrangian epoch ran for 2,000
 603 iterations. We consider the optimization to have converged when a null hypothesis test of constraint
 604 violations being zero is accepted for all constraints at a significance threshold 0.05. This is the dotted
 605 line on the plots below depicting the optimization cutoff of EPI optimization for the 2-dimensional
 606 linear system. If the optimization is left to continue running, entropy usually decreases, and
 607 structural pathologies in the distribution may be introduced.

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

614 **A.1.3 Normalizing flows**

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

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

$$618 \quad \nabla_\theta H(q_\theta(z)) = E_{\omega \sim q_0} [-\nabla_\theta \log(q_\theta(f_\theta(\omega)))] \quad (18)$$

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

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

622 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 623 restricting our layers to normalizing flows [15] – bijective functions with fast log determinant ja-
 624 cobian computations, we can tractably optimize deep generative models with objectives that are a
 625 function of sample density, like entropy. Most of our analyses use real NVP [46], which have proven
 626 effective in our architecture searches, and have the advantageous features of fast sampling and fast
 627 density evaluation.

628 **A.1.4 Related work**

629 (To come)

630

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

632 (To come)

633

634 **A.2 Theoretical models**

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

637 **A.2.1 Stomatogastric ganglion**

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

$$C_m \frac{\partial x_m}{\partial t} = -[h_{leak}(x; z) + h_{Ca}(x; z) + h_K(x; z) + h_{hyp}(x; z) + h_{elec}(x; z) + h_{syn}(x; z)] \quad (20)$$

639 The membrane potential of each neuron is affected by the leak, calcium, potassium, hyperpolariza-
 640 tion, electrical and synaptic currents, respectively. The capacitance of the cell membrane was set to
 641 $C_m = 1nF$. Each current is a function of the neuron's membrane potential x_m and the parameters
 642 of the circuit such as g_{el} and g_{syn} , whose effect on the circuit is considered in the motivational
 643 example of EPI in Fig. 1. Specifically, the currents are the difference in the neuron's membrane
 644 potential and that current type's reversal potential multiplied by a conductance:

$$h_{leak}(x; z) = g_{leak}(x_m - V_{leak}) \quad (21)$$

645

$$h_{elec}(x; z) = g_{el}(x_m^{post} - x_m^{pre}) \quad (22)$$

646

$$h_{syn}(x; z) = g_{syn}S_\infty^{pre}(x_m^{post} - V_{syn}) \quad (23)$$

647

$$h_{Ca}(x; z) = g_{Ca}M_\infty(x_m - V_{Ca}) \quad (24)$$

648

$$h_K(x; z) = g_KN(x_m - V_K) \quad (25)$$

649

$$h_{hyp}(x; z) = g_hH(x_m - V_{hyp}) \quad (26)$$

650 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 651 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 652 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. Fast: $g_{Ca} =$
 653 1.9×10^{-2} , $g_K = 3.9 \times 10^{-2}$, and $g_{hyp} = 2.5 \times 10^{-2}$. Intermediate: $g_{Ca} = 1.7 \times 10^{-2}$, $g_K = 1.9 \times 10^{-2}$,
 654 and $g_{hyp} = 8.0 \times 10^{-3}$. Intermediate: $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

655 Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating
 656 dynamics dependent on steady-state gating variables M_∞ , N_∞ and H_∞ , respectively.

$$M_\infty = 0.5 \left(1 + \tanh \left(\frac{x_m - v_1}{v_2} \right) \right) \quad (27)$$

657
$$\frac{\partial N}{\partial t} = \lambda_N(N_\infty - N) \quad (28)$$

658
$$N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_m - v_3}{v_4} \right) \right) \quad (29)$$

659
$$\lambda_N = \phi_N \cosh \left(\frac{x_m - v_3}{2v_4} \right) \quad (30)$$

660
$$\frac{\partial H}{\partial t} = \frac{(H_\infty - H)}{\tau_h} \quad (31)$$

661
$$H_\infty = \frac{1}{1 + \exp \left(\frac{x_m + v_5}{v_6} \right)} \quad (32)$$

662
$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp \left(\frac{-x_m + v_7}{v_8} \right)} \right) \quad (33)$$

663 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 664 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$. These are the same parameter
 665 values used in [18].

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

$$S_\infty = \frac{1}{1 + \exp \left(\frac{v_{th} - x_m}{v_9} \right)} \quad (34)$$

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

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

673 Our original approach to measuring frequency was to take the max of the fast Fourier transform
 674 (FFT) of the simulated time series. There are a few key considerations here. One is resolution
 675 in frequency space. Each FFT entry will correspond to a signal frequency of $\frac{F_s k}{N}$, where N is
 676 the number of samples used for the FFT, $F_s = \frac{1}{dt}$, and $k \in [0, 1, \dots, N - 1]$. Our resolution is
 677 improved by increasing N and decreasing dt . Increasing $N = T - b$, where b is some fixed number
 678 of buffer burn-in initialization samples, necessitates an increase in simulation time steps T , which
 679 directly increases computational cost. Increasing F_s (decreasing dt) increases system approximation
 680 accuracy, but requires more time steps before a full cycle is observed. At the level of $dt = 0.025$,
 681 thousands of temporal samples were required for resolution of .01Hz. These challenges in frequency

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

Another consideration was that the frequency spectra of the hub neuron has several peaks. This was due to high-frequency sub-threshold activity. The maximum frequency was often not the firing frequency. Accordingly, subthreshold activity was set to zero, and the whole signal was low-pass filtered with a moving average window of length 20. The signal was subsequently mean centered. After this pre-processing, the maximum frequency in the filter bank accurately reflected the firing frequency.

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

$$u_i = \frac{|\mathcal{X}_i|^\alpha}{\sum_{k=1}^N |\mathcal{X}_i(k)|^\alpha} \quad (35)$$

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

Network syncing, like all other emergent properties in this work, are defined by the emergent property statistics and values. The emergent property statistics are the first- and second-moments of the firing frequencies. The first moments are set to 0.55Hz, while the second moments are set to 0.025Hz².

$$E \begin{bmatrix} \Omega_{\text{f1}} \\ \Omega_{\text{f2}} \\ \Omega_{\text{hub}} \\ \Omega_{\text{s1}} \\ \Omega_{\text{s2}} \\ (\Omega_{\text{f1}} - 0.55)^2 \\ (\Omega_{\text{f2}} - 0.55)^2 \\ (\Omega_{\text{hub}} - 0.55)^2 \\ (\Omega_{\text{s1}} - 0.55)^2 \\ (\Omega_{\text{s2}} - 0.55)^2 \end{bmatrix} = \begin{bmatrix} 0.55 \\ 0.55 \\ 0.55 \\ 0.55 \\ 0.55 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \end{bmatrix} \quad (36)$$

For EPI in Fig 2C, we used a real NVP architecture with two coupling layers. Each coupling layer had two hidden layers of 10 units each, and we mapped onto a support of $z \in \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 \\ 8 \end{bmatrix} \right]$. We

701 have shown the EPI optimization that converged with maximum entropy across 2 random seeds
 702 and augmented Lagrangian coefficient initializations of $c_0=0$, 2, and 5.

703 **A.2.2 Primary visual cortex**

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

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n \quad (37)$$

705 Some neuron-types largely lack synaptic projections to other neuron-types [47], and it is popular
 706 to only consider a subset of the effective connectivities [19].

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & 0 \\ W_{PE} & W_{PP} & W_{PS} & 0 \\ W_{SE} & 0 & 0 & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & 0 \end{bmatrix} \quad (38)$$

707 Estimates of the probability of connection and strength of connection from the Allen institute
 708 result in an estimate of the effective connectivity [?]:

$$W = \begin{bmatrix} 0.0576 & 0.19728 & 0.13144 & 0 \\ 0.58855 & 0.30668 & 0.4285 & 0 \\ 0.15652 & 0 & 0 & 0.2 \\ 0.13755 & 0.0902 & 0.4004 & 0 \end{bmatrix} \quad (39)$$

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

$$h = b + dh \quad (40)$$

713 All simulations of this system had $T = 100$ time points, a time step $dt = 5\text{ms}$, and time constant
 714 $\tau = 20\text{ms}$. And the system was initialized to a random draw $x(0)_i \sim \mathcal{N}(1, 0.01)$.

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

$$\dot{x}_i = -x_i + f(u_i) \quad (41)$$

716 where the input to each neuron is

$$u_i = \sum_j W_{ij}x_j + h_i \quad (42)$$

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

$$\frac{\partial x_{ss}}{\partial h} = F(W \frac{\partial x_{ss}}{\partial h} + I) \quad (43)$$

718 which is calculable by

$$\frac{\partial x_{ss}}{\partial h} = (F^{-1} - W)^{-1} \quad (44)$$

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

$$\mathcal{B}(\alpha, y) \leftrightarrow E \begin{bmatrix} dx_{\alpha,ss} \\ (dx_{\alpha,ss} - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix} \quad (45)$$

722 In the final analysis for this model, we sweep the input one neuron at a time away from the mode
 723 of each inferred distributions $dh^* = z^* = \text{argmax}_z \log q_\theta(z | \mathcal{B}(\alpha, 0.1))$. The differential responses
 724 $dx_{\alpha,ss}$ are examined at perturbed inputs $h = b + dh^* + \Delta h_\alpha u_\alpha$ where u_α is a unit vector in the
 725 dimension of α and $\Delta h_\alpha \in [-15, 15]$.

726 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
 727 initial seeds using an architecture of four coupling layers, each with two hidden layers of 10 units.
 728 We set $c_0 = 10^5$. The support of the learned distribution was restricted to $z_i \in [-5, 5]$.

729 A.2.3 Superior colliculus

730 There are four total units: two in each hemisphere corresponding to the Pro/contralateral and
 731 Anti/ipsilateral populations. Each unit has an activity (x_i) and internal variable (u_i) related by

$$x_i(t) = \left(\frac{1}{2} \tanh \left(\frac{v_i(t) - \epsilon}{\zeta} \right) + \frac{1}{2} \right) \quad (46)$$

732 $\epsilon = 0.05$ and $\zeta = 0.5$ control the position and shape of the nonlinearity, respectively.

733 We can order the elements of x_i and v_i into vectors x and v with elements

$$x = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad v = \begin{bmatrix} v_{LP} \\ v_{LA} \\ v_{RP} \\ v_{RA} \end{bmatrix} \quad (47)$$

⁷³⁴ The internal variables follow dynamics:

$$\tau \frac{\partial v}{\partial t} = -v + Wx + h + \sigma \partial B \quad (48)$$

⁷³⁵ with time constant $\tau = 0.09s$ and gaussian noise $\sigma \partial B$ controlled by the magnitude of $\sigma = 1.0$. The
⁷³⁶ weight matrix has 8 parameters sW_P , sW_A , vW_{PA} , vW_{AP} , hW_P , hW_A , dW_{PA} , and dW_{AP} (Fig.
⁷³⁷ 4B).

$$W = \begin{bmatrix} sW_P & vW_{PA} & hW_P & dW_{PA} \\ vW_{AP} & sW_A & dW_{AP} & hW_A \\ hW_P & dW_{PA} & sW_P & vW_{PA} \\ dW_{AP} & hW_A & vW_{AP} & sW_A \end{bmatrix} \quad (49)$$

⁷³⁸ The system receives five inputs throughout each trial, which has a total length of 1.8s.

$$h = h_{\text{rule}} + h_{\text{choice-period}} + h_{\text{light}} \quad (50)$$

⁷³⁹ There are rule-based inputs depending on the condition,

$$h_{P,\text{rule}}(t) = \begin{cases} I_{P,\text{rule}} \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (51)$$

$$h_{A,\text{rule}}(t) = \begin{cases} I_{A,\text{rule}} \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (52)$$

⁷⁴¹ a choice-period input,

$$h_{\text{choice}}(t) = \begin{cases} I_{\text{choice}} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (53)$$

⁷⁴² and an input to the right or left-side depending on where the light stimulus is delivered.

$$h_{\text{light}}(t) = \begin{cases} I_{\text{light}} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Left} \\ I_{\text{light}} \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Right} \\ 0, & t \leq 1.2s \end{cases} \quad (54)$$

⁷⁴³ The input parameterization was fixed to $I_{P,\text{rule}} = 10$, $I_{A,\text{rule}} = 10$, $I_{\text{choice}} = 2$, and $I_{\text{light}} = 1$

⁷⁴⁴ To produce a Bernoulli rate of p_{LP} in the Left, Pro condition (we can generalize this to either cue,
⁷⁴⁵ or stimulus condition), let \hat{p}_i be the empirical average steady state (ss) response (final x_{LP} at end
⁷⁴⁶ of task) over M=500 gaussian noise draws for a given SC model parameterization z_i :

$$\hat{p}_i = E_{\sigma \partial B} [x_{LP,ss} | s = L, c = P, z_i] = \frac{1}{M} \sum_{j=1}^M x_{LP,ss}(s = L, c = P, z_i, \sigma \partial B_j) \quad (55)$$

⁷⁴⁷ For the first constraint, the average over posterior samples (from $q_\theta(z)$) to be p_{LP} :

$$E_{z_i \sim q_\phi} [E_{\sigma \partial B} [x_{LP,ss} | s = L, c = P, z_i]] = E_{z_i \sim q_\phi} [\hat{p}_i] = p_{LP} \quad (56)$$

⁷⁴⁸ We can then ask that the variance of the steady state responses across gaussian draws, is the
⁷⁴⁹ Bernoulli variance for the empirical rate \hat{p}_i .

$$E_{z \sim q_\phi} [\sigma_{err}^2] = 0 \quad (57)$$

⁷⁵⁰

$$\sigma_{err}^2 = Var_{\sigma \partial B} [x_{LP,ss} | s = L, c = P, z_i] - \hat{p}_i(1 - \hat{p}_i) \quad (58)$$

⁷⁵¹ We have an additional constraint that the Pro neuron on the opposite hemisphere should have the
⁷⁵² opposite value. We can enforce this with a final constraint:

$$E_{z \sim q_\phi} [d_P] = 1 \quad (59)$$

⁷⁵³

$$E_{\sigma \partial W} [(x_{LP,ss} - x_{RP,ss})^2 | s = L, c = P, z_i] \quad (60)$$

⁷⁵⁴ We refer to networks obeying these constraints as Bernoulli, winner-take-all networks. Since the
⁷⁵⁵ maximum variance of a random variable bounded from 0 to 1 is the Bernoulli variance ($\hat{p}(1 - \hat{p})$),
⁷⁵⁶ and the maximum squared difference between two variables bounded from 0 to 1 is 1, we do not
⁷⁵⁷ need to control the second moment of these test statistics. In reality, these variables are dynamical
⁷⁵⁸ system states and can only exponentially decay (or saturate) to 0 (or 1), so the Bernoulli variance
⁷⁵⁹ error and squared difference constraints can only be undershot. This is important to be mindful
⁷⁶⁰ of when evaluating the convergence criteria. Instead of using our usual hypothesis testing criteria
⁷⁶¹ for convergence to the emergent property, we set a slack variable threshold for these technically
⁷⁶² infeasible constraints to 0.05.

⁷⁶³ Training DSNs to learn distributions of dynamical system parameterizations that produce Bernoulli
⁷⁶⁴ responses at a given rate (with small variance around that rate) was harder to do than expected.
⁷⁶⁵ There is a pathology in this optimization setup, where the learned distribution of weights is bimodal
⁷⁶⁶ attributing a fraction p of the samples to an expansive mode (which always sends x_{LP} to 1), and a
⁷⁶⁷ fraction $1 - p$ to a decaying mode (which always sends x_{LP} to 0). This pathology was avoided using
⁷⁶⁸ an inequality constraint prohibiting parameter samples that resulted in low variance of responses
⁷⁶⁹ across noise.

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

$$\mathcal{B}(p) \leftrightarrow \begin{bmatrix} \hat{p}_P \\ \hat{p}_A \\ (\hat{p}_P - p)^2 \\ (\hat{p}_A - p)^2 \\ \sigma_{P,err}^2 \\ \sigma_{A,err}^2 \\ d_P \\ d_A \end{bmatrix} = \begin{bmatrix} p \\ p \\ 0.15^2 \\ 0.15^2 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad (61)$$

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

773 **A.2.4 Rank-1 RNN**

774 The network dynamics of neuron i 's rate x evolve according to:

$$\dot{x}_i(t) = -x_i(t) + \sum_{j=1}^N J_{ij}\phi(x_j(t)) + I_i \quad (62)$$

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

$$J_{ij} = g\chi_{ij} + P_{ij} \quad (63)$$

776 The random bulk component has elements drawn from $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$, and the structured compo-
 777 nent is a sum of r unit rank terms:

$$P_{ij} = \sum_{k=1}^r \frac{m_i^{(k)} n_j^{(k)}}{N} \quad (64)$$

778 Rank-1 vectors m and n have elements drawn

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

779

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

780 The current has the following statistics:

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

781 where x_1 , x_2 , and h are standard normal random variables following the rank-1 input-driven ex-
 782 ample from [21].

⁷⁸³ We followed their prescription for deriving the consistency equations in the presence of chaos. The
⁷⁸⁴ $\ddot{\Delta}$ equation is broken into the equation for Δ_0 and Δ_∞ by the autocorrelation dynamics assertions.

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

⁷⁸⁵

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

⁷⁸⁶ We can write out the potential function by integrating the negated RHS.

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

⁷⁸⁷

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

⁷⁸⁸ We assume that as time goes to infinity, the potential relaxes to a steady state.

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

⁷⁸⁹

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

⁷⁹⁰ This can be written more explicitly in terms of the gaussian integrals which are relatively (with
⁷⁹¹ respect to nongaussian distributions) cheap to evaluate.

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

⁷⁹² Also, we assume that the energy of the system is preserved throughout the entirety of its evolution.

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

⁷⁹³

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

⁷⁹⁴ We can arrange the terms into a difference of squares in Δ_0 and Δ_∞ .

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

⁷⁹⁵ Similarly, we write out the resulting equation explicitly in terms of the gaussian integrals present.

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

797 This results in a set of consistency equations for the dynamic mean field variables μ , κ , Δ_0 , and
 798 Δ_∞ . In order to obtain the values of these variables for a given parameterization, we must solve
 799 the following system of equations.

$$\begin{aligned} \mu &= F(\mu, \kappa, \Delta_0, \Delta_\infty) = M_m \kappa + M_I \\ \kappa &= G(\mu, \kappa, \Delta_0, \Delta_\infty) = M_n \langle [\phi_i] \rangle + \Sigma_{nI} \langle [\phi'_i] \rangle \\ \frac{\Delta_0^2 - \Delta_\infty^2}{2} &= H(\mu, \kappa, \Delta_0, \Delta_\infty) = g^2 \left(\int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0} z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right) \\ &\quad + (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI} \kappa + \Sigma_I^2)(\Delta_0 - \Delta_\infty) \\ \Delta_\infty &= L(\mu, \kappa, \Delta_0, \Delta_\infty) = g^2 \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 + \Sigma_m^2 \kappa^2 + 2\Sigma_{mI} \kappa + \Sigma_I^2 \end{aligned} \tag{65}$$

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

$$\begin{aligned} x(t) &= \frac{\Delta_0(t)^2 - \Delta_\infty(t)^2}{2} \\ \Delta_0(t) &= \sqrt{2x(t) + \Delta_\infty(t)^2} \\ \dot{\mu}(t) &= -\mu(t) + F(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t)) \\ \dot{\kappa}(t) &= -\kappa + G(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t)) \\ \dot{x}(t) &= -x(t) + H(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t)) \\ \dot{\Delta_\infty}(t) &= -\Delta_\infty(t) + L(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t)) \end{aligned} \tag{66}$$

801 Then, the temporal variance, which is necessary for the gaussian posterior conditioning example, is
 802 simply calculated via

$$\Delta_T = \Delta_0 - \Delta_\infty \tag{67}$$

803 A.3 Supplementary Figures

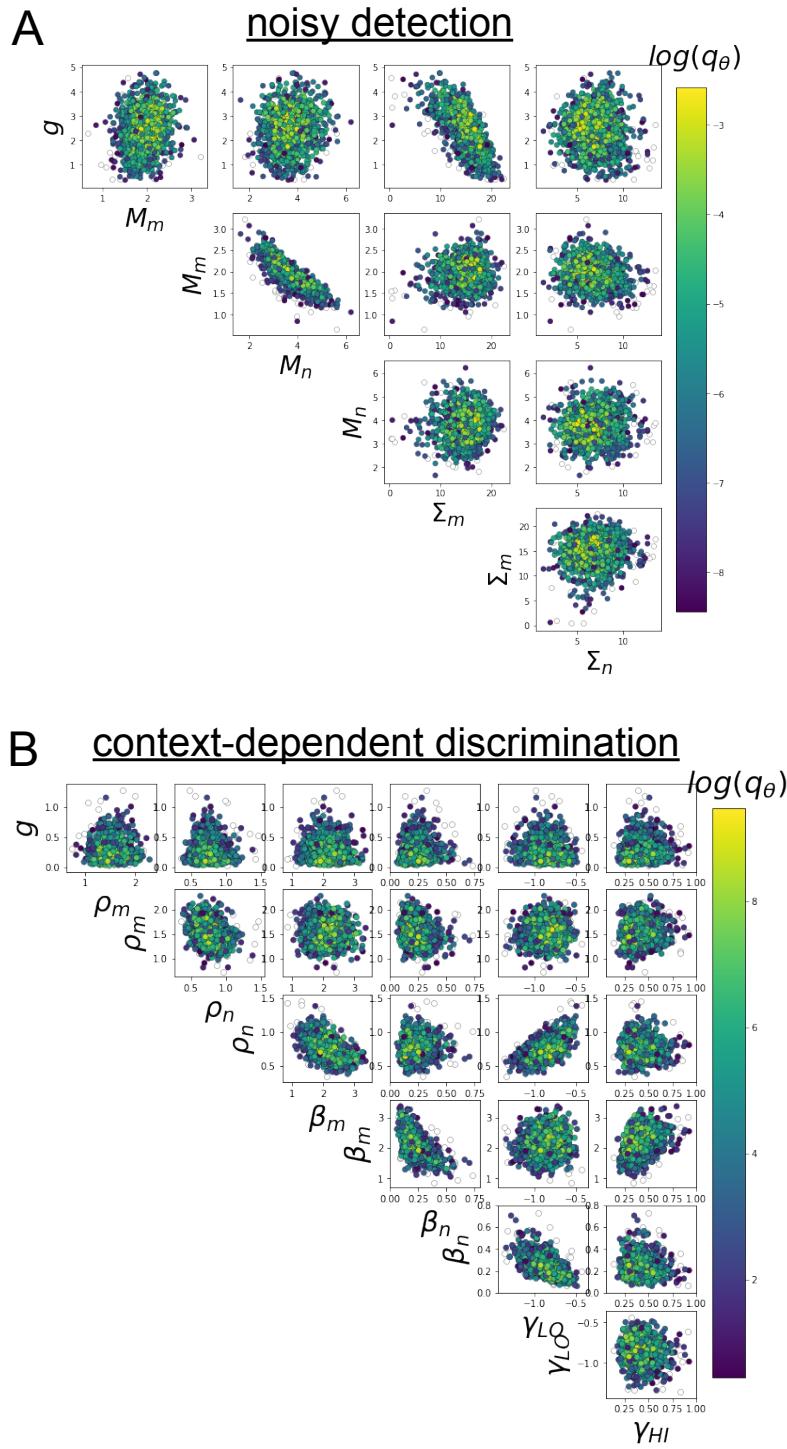


Fig. S1: A. EPI for rank-1 networks doing discrimination. B. EPI for rank-2 networks doing context-dependent discrimination.