

Interrogating theoretical models of neural computation with deep inference

Sean R. Bittner¹, Agostina Palmigiano¹, Alex T. Piet², Chunyu A. Duan³, Carlos D. Brody², Kenneth D. Miller¹, and John P. Cunningham⁴.

¹Department of Neuroscience, Columbia University,

²Princeton Neuroscience Institute,

³Institute of Neuroscience, Chinese Academy of Sciences,

⁴Department of Statistics, Columbia University

¹ 1 Abstract

² A cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures
³ a hypothesized neural mechanism. Such models are valuable when they give rise to an experimen-
⁴ tally observed phenomenon – whether behavioral or in terms of neural activity – and thus can offer
⁵ insights into neural computation. The operation of these circuits, like all models, critically depends
⁶ on the choices of model parameters. Historically, the gold standard has been to analytically derive
⁷ the relationship between model parameters and computational properties. However, this enterprise
⁸ quickly becomes infeasible as biologically realistic constraints are included into the model increas-
⁹ ing its complexity, 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
¹³ offer a principled means to understand the implications of model parameter choices on compu-
¹⁴ tational properties of interest. We motivate this methodology with a worked example analyzing
¹⁵ sensitivity in the stomatogastric ganglion. We then use it to generate insights into neuron-type
¹⁶ input-responsivity in a model of primary visual cortex, a new understanding of rapid task switch-
¹⁷ ing in superior colliculus models, and attribution of error in recurrent neural networks solving a
¹⁸ simple mathematical task. More generally, this work suggests a departure from realism vs tractabil-
¹⁹ ity considerations, towards the use of modern machine learning for sophisticated interrogation of
²⁰ biologically relevant 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]. A neural computation is systematized with a set of equations – the model – and
25 these equations are motivated by biophysics, neurophysiology, and other conceptual considerations.
26 The function of this system is governed by the choice of model parameters, which when configured
27 in a particular way, give rise to a measurable signature of a computation. The work of analyzing a
28 model then requires solving the inverse problem: given a computation of interest, how can we reason
29 about these particular parameter configurations? The inverse problem is crucial for reasoning about
30 likely parameter values, uniquenesses and degeneracies, 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 [5]. Unfortunately, as circuit models include more
37 biological realism, theory via analytical derivation becomes intractable. This creates an unfavor-
38 able tradeoff. On the one hand, one may tractably analyze systems of equations with unrealistic
39 assumptions (for example symmetry or gaussianity), producing accurate inferences about param-
40 eters 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

model, given data. Indeed, these tools have been used to great success in neuroscience as well, in particular for interrogating parameters (sometimes treated as hidden states) in models of both cortical population activity [8, 9, 10, 11] and animal behavior [12, 13, 14]. These works have used deep neural networks to expand the expressivity and accuracy of statistical models of neural data [15].

However, these inference tools have not significantly influenced the study of theoretical neuroscience 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 can not 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 system 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, these results point to the value of deep inference models for the 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, 30]. 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 [31]). 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

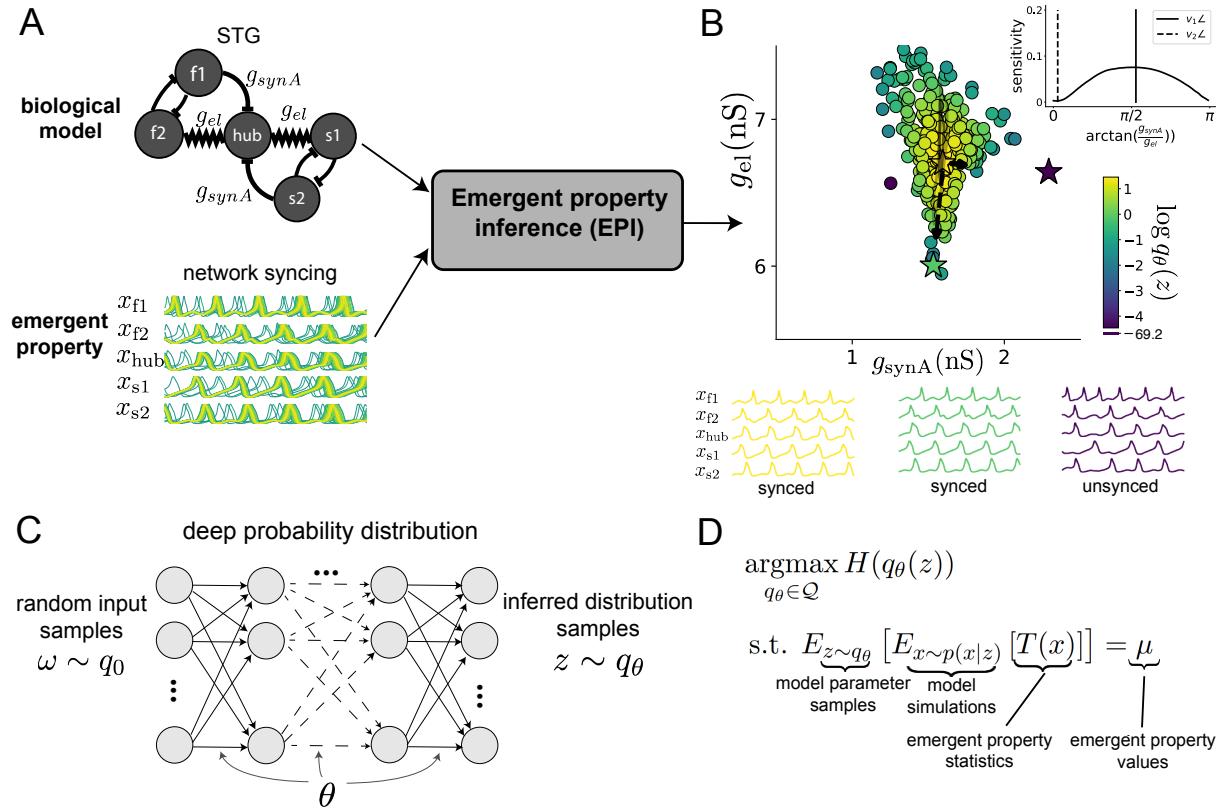


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. A. For a choice of model (STG) and emergent property (network syncing), emergent property inference (EPI, gray box) learns a distribution of the model parameters $z = [g_{el}, g_{synA}]$ producing network syncing. In the STG 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 (stars) in the EPI-inferred distribution. B. The EPI distribution of STG model parameters producing network syncing. Samples are colored by log probability density. Distribution contours of emergent property value error are shown at levels of 5×10^{-7} and 1×10^{-6} (dark and light gray). 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 B.2.1). C. Deep probability distributions map a latent random variable w through a deep neural network with weights and biases θ 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 μ .

114 muscle activation patterns for digestion [32]. Despite full knowledge of STG connectivity and a
 115 precise characterization of its rhythmic pattern generation, biophysical models of the STG have
 116 complicated relationships between circuit parameters and neural activity [22, 33]. A model of the
 117 STG [23] is shown schematically in Figure 1A, and note that the behavior of this model will be crit-
 118 ically dependent on its parameterization – the choices of conductance parameters $z = [g_{el}, g_{synA}]$.
 119 Specifically, the two fast neurons (f_1 and f_2) mutually inhibit one another, and oscillate at a faster
 120 frequency than the mutually inhibiting slow neurons (s_1 and s_2), and the hub neuron (hub) couples
 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 specif-
 138 ically queried to reveal the prototypical parameter configuration for network syncing (the mode;
 139 Figure 1B yellow star), and how network syncing decays based on changes away from the mode.
 140 The eigenvectors (of the Hessian of the distribution at the mode) can be queried to quantitatively
 141 formalize the robustness of network syncing (Fig. 1B v_1 and v_2). Indeed, samples equidistant from
 142 the mode along these EPI-identified dimensions of sensitivity (v_1) and degeneracy (v_2) agree with
 143 error contours (Fig. 1B, contours) and have diminished or preserved network syncing, respectively
 144 (Figure 1B inset and activity traces). Further validation of EPI is available in the supplemen-
 145 tary materials, where we analyze a simpler model for which ground-truth statements can be made

¹⁴⁶ (Section B.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 $x = [x_{f1}, x_{f2}, x_{\text{hub}}, x_{s1}, x_{s2}]$ is the membrane potential
¹⁵¹ for each neuron, which evolves according to the biophysical conductance-based equation:

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

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 B.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 0.542Hz. 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:

$$\mathbb{E}[T(x)] \triangleq \mathbb{E} \begin{bmatrix} \omega_{f1}(x) \\ \vdots \\ (\omega_{f1}(x) - 0.542)^2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0.542 \\ \vdots \\ 0.025^2 \\ \vdots \end{bmatrix} \triangleq \mu, \quad (2)$$

¹⁶¹ which completes the quantification of the emergent property.

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

169 program:

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

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,
 173 we recognize that many distributions in \mathcal{Q} will respect the emergent property constraints, so we
 174 require a normative principle to select amongst them. This principle is captured in Equation 3 by
 175 the primal objective H . Here we chose Shannon entropy as a means to find parameter distributions
 176 with minimal assumptions beyond some chosen structure [34, 35, 20, 36], but we emphasize that
 177 the EPI method is unaffected by this choice (but the results of course will depend on the primal
 178 objective chosen).

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

189 3.3 Comprehensive input-responsivity in a nonlinear sensory system

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



Figure 2: Hypothesis generation through EPI in a V1 model. A. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons. Some neuron-types largely do not form synaptic projections to others (excitatory and inhibitory projections filled and unfilled, respectively). B. V1 model simulations for input (solid) $h = b$ and (dashed) $h = b + dh$. Stars indicate the linear response prediction. C. EPI distributions on differential input dh conditioned on differential response $B(\alpha, y)$. 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).

of GABAergic interneurons in V1 [40, 41, 42], and that these inhibitory cell types follow specific connectivity patterns (Fig. 2A) [43]. Recent theoretical advances [24, 44, 45], have only started to address the consequences of this multiplicity in the dynamics of V1, strongly relying on linear theoretical tools. Here, we go beyond linear theory by systematically generating and evaluating hypotheses of circuit model function using EPI distributions of neuron-type inputs producing various neuron-type population responses.

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

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

The effective connectivity weights W were obtained from experimental recordings of publicly available datasets of mouse V1 [46, 47] (see Section B.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_α (the change in x_α when receiving input $h = b + dh$ with respect to the baseline $h = b$) of value y with some small,

arbitrarily chosen amount of variance 0.01². 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 inferred distributions immediately suggest four hypotheses:

232

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

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

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

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

We evaluate these hypotheses by taking steps in individual neuron-type input δh_α away from the modes of the inferred distributions at $y = 0.1$

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

Here δx_α is the change in steady state response to the system with input $h = b + dh^* + \delta h_\alpha \hat{u}_\alpha$ compared to $h = b + dh^*$, where \hat{u}_α is a unit vector in the dimension of α . The EPI-generated hypotheses are confirmed:

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

H2: the E- and P-populations are not affected by δh_V (Fig. 3A green, 3B green);

H3: the S-population is not affected by δh_P (Fig. 3C blue);

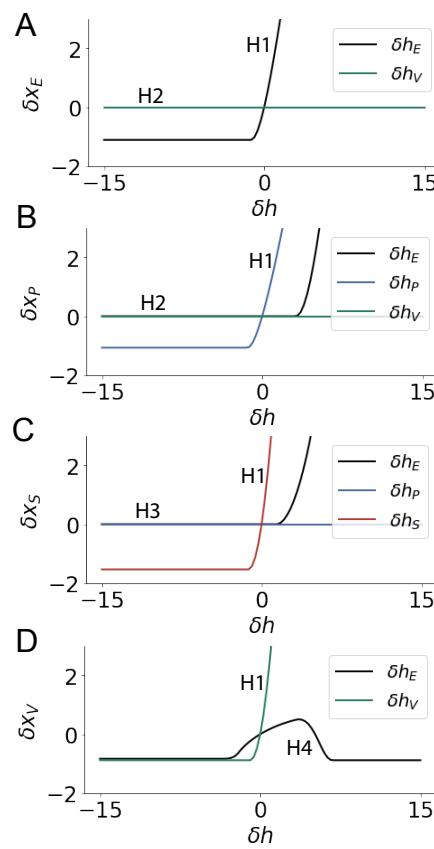


Figure 3: Confirming EPI generated hypotheses in V1. A. Differential responses by the E-population to changes in individual input $\delta h_\alpha \hat{u}_\alpha$ away from the mode of the EPI distribution dh^* . B-D Same plots for the P-, S-, and V-populations. Labels H1, H2, H3, and H4 indicate which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

252 H4: the V-population exhibits a nonmonotonic response to δh_E (Fig. 3D black), and is in
 253 fact the only population to do so (Fig. 3A-C black).

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

259 3.4 Identifying neural mechanisms of flexible task switching

260 In a rapid task switching experiment [48], rats were explicitly cued on each trial to either orient
 261 towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A)
 262 task (Fig. 4a). Neural recordings in the midbrain superior colliculus (SC) exhibited two population
 263 of neurons that simultaneously represented both task context (Pro or Anti) and motor response
 264 (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons [25].
 265 Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is

266 a four-population dynamical system. We analyzed this model, where the neuron-type populations
 267 are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right
 268 (R)). The Pro- or Anti-populations receive an input determined by the cue, and then the left and
 269 right populations receive an input based on the side of the light stimulus. Activities were bounded
 270 between 0 and 1, so that a high output of the Pro population in a given hemisphere corresponds
 271 to the contralateral response. An additional stipulation is that when one Pro population responds
 272 with a high-output, the opposite Pro population must respond with a low output. Finally, this
 273 circuit operates in the presence of Gaussian noise resulting in trial-to-trial variability (see Section
 274 B.2.3). The connectivity matrix is parameterized by the geometry of the population arrangement
 275 (Fig. 4B).

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

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

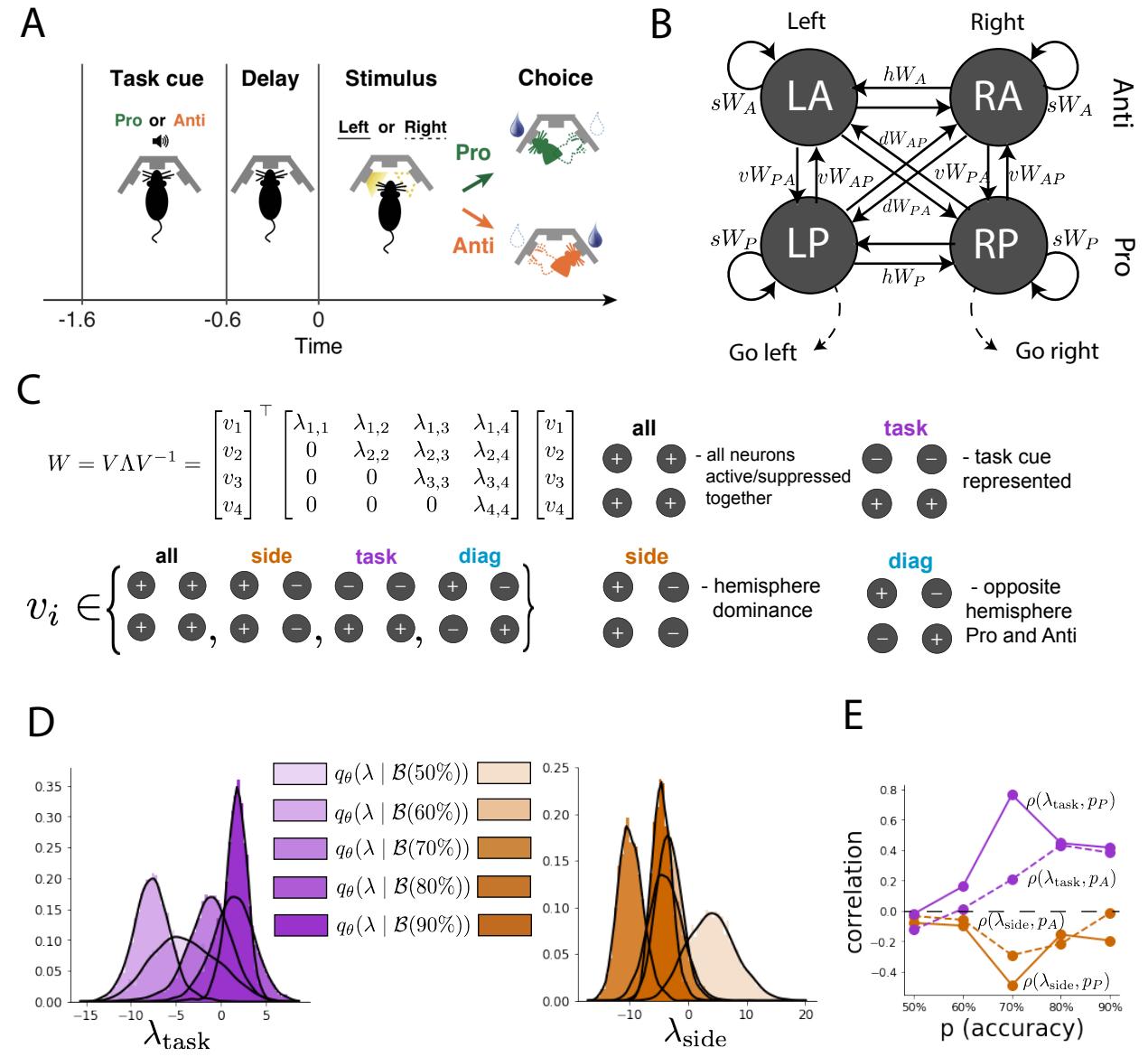


Figure 4: EPI reveals changes in SC [25] connectivity that control task accuracy. A. Rapid task switching behavioral paradigm (see text). B. Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, 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 Λ . Schur modes: v_{all} , v_{task} , v_{side} , and v_{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.

²⁹⁷ **3.5 Linking RNN connectivity to error**

²⁹⁸ So far, each model we have studied was designed from fundamental biophysical principles, genetically-
²⁹⁹ or functionally-defined neuron types. At a more abstract level of modeling, recurrent neural net-
³⁰⁰ works (RNNs) are high-dimensional dynamical models of computation that are becoming increas-
³⁰¹ ingly popular in neuroscience research [49]. In theoretical neuroscience, RNN dynamics usually
³⁰² follow the equation

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

³⁰³ where $x(t)$ is the network activity, W is the network connectivity, $\phi(\cdot) = \tanh(\cdot)$, and $h(t)$ is the
³⁰⁴ input to the system. Such RNNs are trained to do a task from a systems neuroscience experiment,
³⁰⁵ and then the unit activations of the trained RNN are compared to recorded neural activity. Fully-
³⁰⁶ connected RNNs with tens of thousands of parameters are challenging to characterize [50], especially
³⁰⁷ making statistical inferences about their parameterization. Alternatively, we considered a rank-1,
³⁰⁸ N -neuron RNN with connectivity

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

³⁰⁹ 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
³¹⁰ distributions $m_i \sim \mathcal{N}(M_m, 1)$ and $n_i \sim \mathcal{N}(M_n, 1)$. We used EPI to infer the parameterizations of
³¹¹ rank-1 RNNs solving an example task, enabling discovery of properties of connectivity that result
³¹² in different types of error in the computation.

³¹³ The task we consider is Gaussian posterior conditioning: calculate the parameters of a posterior
³¹⁴ distribution induced by a prior $p(\mu_y) = \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 1)$ and a likelihood $p(y|\mu_y) = \mathcal{N}(\mu_y, \sigma_y^2 =$
³¹⁵ 1), given a single observation y . Conjugacy offers the result analytically; $p(\mu_y|y) = \mathcal{N}(\mu_{post}, \sigma_{post}^2)$,
³¹⁶ where:

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

³¹⁷ The RNN is trained to solve this task by producing readout activity that is on average the posterior
³¹⁸ mean μ_{post} , and activity whose variability is the posterior variance σ_{post}^2 (Fig. 5A, a setup inspired
³¹⁹ by [51]). To solve this Gaussian posterior conditioning task, the RNN response to a constant input
³²⁰ $h(t) = yw + (n - M_n)$ must equal the posterior mean along readout vector r , where

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

³²¹ Additionally, the amount of chaotic variance Δ_T must equal the posterior variance. Theory for
³²² low-rank RNNs allows us to express κ_r and Δ_T in terms of each other through a solvable system

of nonlinear equations (see Section B.2.4) [26]. This allows us to mathematically formalize the execution of this task into an emergent property, where the emergent property statistics of the RNN activity are κ_r and Δ_T and the emergent property values are the ground truth posterior mean μ_{post} and variance σ_{post}^2 :

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

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

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

We tested this prediction by taking parameters z_1 and z_2 as representative samples from the positive and negative M_m - M_n quadrants, respectively. Instead of using the theoretical predictions shown in Figure 5B, we simulated finite-size realizations of these networks with 2,000 neurons (e.g. Fig. 5C). We perturbed these parameter choices by the product $M_m M_n$ clarifying that the posterior mean can be directly controlled in this way (Fig. 5D). Thus, EPI confers a clear picture of error in this computation: the product of the low rank vector means M_m and M_n modulates the estimated posterior mean while the random strength g modulates the chaotic variance. This novel procedure of inference on reduced parameterizations of RNNs conditioned on the emergent property of task execution is generalizable to other settings modeled in [26] like noisy integration

351 and context-dependent decision making (Fig. S4).

352 4 Discussion

353 4.1 EPI is a general tool for theoretical neuroscience

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

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

373 There are additional technical considerations when assessing the suitability of EPI for a particu-
374 lar modeling question. First and foremost, as in any optimization problem, the defined emergent
375 property should always be appropriately conditioned (constraints should not have wildly different
376 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution
377 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-
378 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,
379 one should consider the computational cost of the gradient calculations. In the best circumstance,
380 there is a simple, closed form expression (e.g. Section B.1.1) for the emergent property statistic

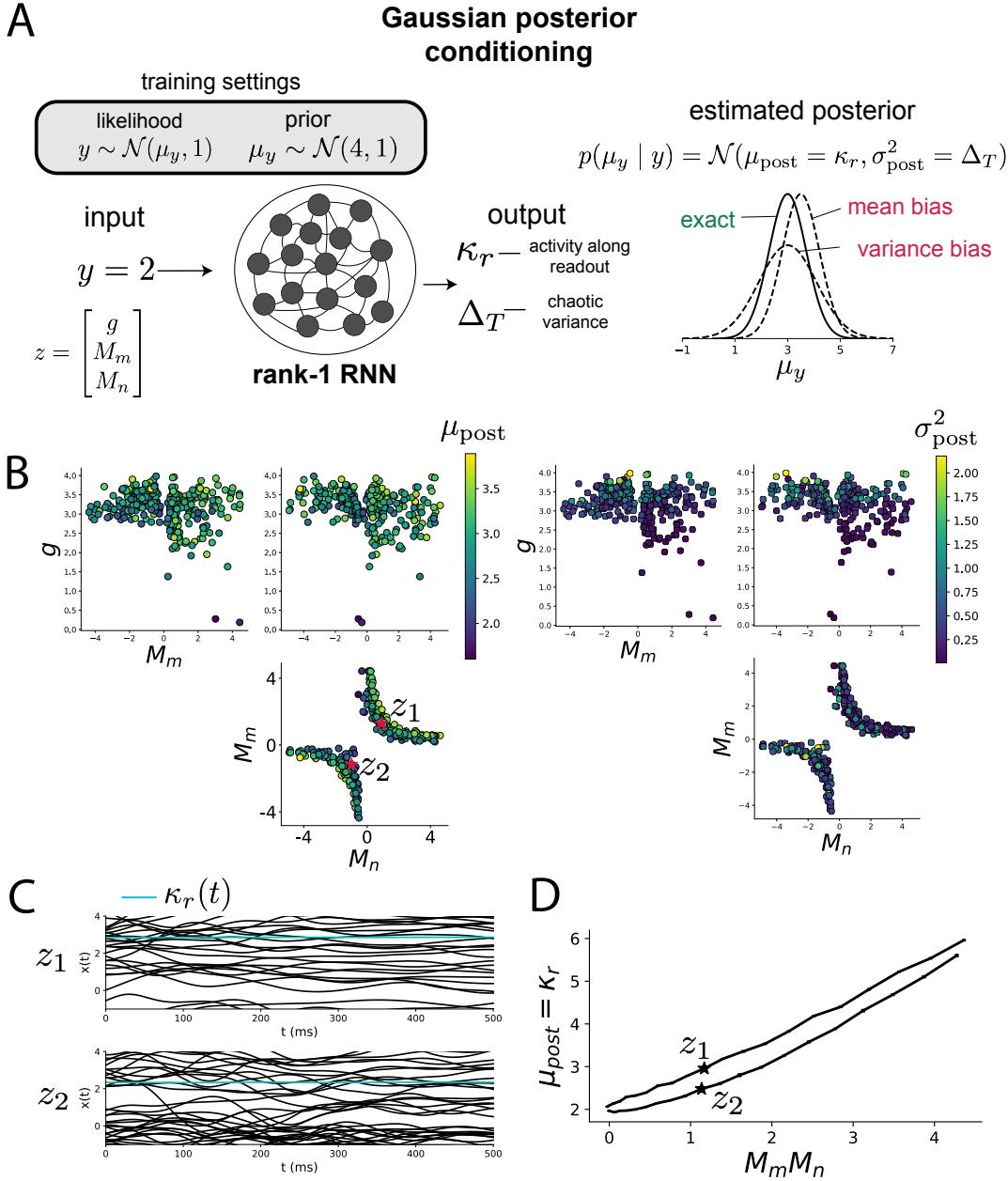


Figure 5: Sources of error in an RNN solving a simple task. A. (left) A rank-1 RNN executing a Gaussian posterior conditioning computation on μ_y . (right) Error in this computation can come from over- or under-estimating the posterior mean or variance. B. EPI distribution of rank-1 RNNs executing Gaussian 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 network simulations of 2,000 neurons with parameters z_1 and z_2 sampled from the inferred distribution. Activity along readout κ_r (cyan) is stable despite chaotic fluctuations. D. The posterior mean computed by RNNs parameterized by z_1 and z_2 pertrubed in the dimension of the product of M_m and M_n . Means and standard errors are shown across 10 realizations of 2,000-neuron networks.

381 given the model parameters. On the other end of the spectrum, many forward simulation iterations
382 may be required before a high quality measurement of the emergent property statistic is available
383 (e.g. Section B.2.1). In such cases, optimization will be expensive.

384 **4.2 Novel hypotheses from EPI**

385 In neuroscience, machine learning has primarily been used to revealed structure in large-scale neural
386 datasets [52, 53, 54, 55, 56, 57] (see review, [15]). Such careful inference procedures are developed
387 for these statistical models allowing precise, quantitative reasoning, which clarifies the way data
388 informs knowledge of the model parameters. However, these inferable statistical models lack re-
389 semblance to the underlying biology, making it unclear how to go from the structure revealed by
390 these methods, to the neural mechanisms giving rise to it. In contrast, theoretical neuroscience has
391 focused on careful mechanistic modeling and the production of emergent properties of computation.
392 The careful steps of 1.) model design and 2.) emergent property definition, are followed by 3.)
393 practical inference methods resulting in an opaque characterization of the way model parameters
394 govern computation. In this work, we replaced this opaque procedure of parameter identification
395 in theoretical neuroscience with emergent property inference, opening the door to careful inference
396 in careful models of neural computation.

397 Biologically realistic models of neural circuits often prove formidable to analyze. For example,
398 consider the fact that we do not fully understand the (only) four-dimensional models of V1 [24]
399 and SC [25]. Because analytical approaches to studying nonlinear dynamical systems become
400 increasingly complicated when stepping from two-dimensional to three- or four-dimensional systems
401 in the absence of restrictive simplifying assumptions [58], it is unsurprising that these models pose a
402 challenge. In Section 3.3, we showed that EPI was far more informative about neuron-type input-
403 responsivity than the predictions afforded through the available linear analytical methods. By
404 flexibly conditioning this V1 model on different emergent properties, we performed an exploratory
405 analysis of a *model* rather than a dataset, which generated a set of testable hypotheses, which
406 were proved out. Of course, exploratory analyses can be directed towards formulating hypotheses
407 of a specific form. For example, when interested in model parameter changes with behavioral
408 performance, one can use EPI to condition on various levels of task accuracy as we did in Section
409 3.4. This analysis identified experimentally testable predictions (proved out *in-silico*) of patterns
410 of effective connectivity in SC that should be correlated with increased performance.

411 In our final analysis, we presented a novel procedure for doing statistical inference on interpretable

parameterizations of RNNs executing simple tasks. Specifically, we analyzed RNNs solving a posterior conditioning problem in the spirit of [51]. This methodology relies on recently extended theory of responses in random neural networks with minimal structure [26]. While we focused on rank-1 RNNs, which were sufficient for solving this task, we can more generally use this approach to analyze rank-2 and greater RNNs. The ability to apply the probabilistic model selection toolkit to such black box models should prove invaluable as their use in neuroscience increases.

418 References

- 419 [1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 420 [2] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 422 [3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.
- 424 [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.
- 427 [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.
- 429 [6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.
- 431 [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.
- 434 [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.
- 437 [9] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.

- 439 [10] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders.
440 *bioRxiv*, page 399246, 2018.
- 441 [11] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
442 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg,
443 et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature
444 methods*, page 1, 2018.
- 445 [12] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
446 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
447 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 448 [13] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
449 Datta. Composing graphical models with neural networks for structured representations and
450 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 451 [14] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
452 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
453 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural
454 Information Processing Systems*, 2019.
- 455 [15] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
456 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
457 2018.
- 458 [16] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R
459 Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology,
460 and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.
- 461 [17] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
462 *International Conference on Machine Learning*, 2015.
- 463 [18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
464 *arXiv preprint arXiv:1605.08803*, 2016.
- 465 [19] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density
466 estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

- [20] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow networks. *International Conference on Learning Representations*, 2017.
- [21] 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.
- [22] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] Francesca Mastrogiovanni and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- [27] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degenerate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon, Portugal*, 2019.
- [28] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller, Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- [29] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In *Bernstein Conference 2018, Berlin, Germany*, 2018.

- 494 [30] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical inference for
495 analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin, Germany*,
496 2019.
- 497 [31] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
498 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
499 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 500 [32] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
501 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 502 [33] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
503 parameters. *Nature neuroscience*, 7(12):1345, 2004.
- 504 [34] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
505 1957.
- 506 [35] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
507 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 508 [36] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
509 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 510 [37] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
511 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 512 [38] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
513 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
514 592, 2009.
- 515 [39] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
516 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,
517 85(2):402–417, 2015.
- 518 [40] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
519 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
520 5(10):793, 2004.

- 521 [41] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
522 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-*
523 *biology*, 71(1):45–61, 2011.
- 524 [42] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex:
525 From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 526 [43] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
527 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
528 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 529 [44] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing Wang.
530 Paradoxical response reversal of top- down modulation in cortical circuits with three interneu-
531 ron types. *Elife*, 6:1–15, 2017.
- 532 [45] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying
533 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 534 [46] (2018) Allen Institute for Brain Science. Layer 4 model of v1. available from:
535 <https://portal.brain-map.org/explore/models/l4-mv1>.
- 536 [47] Yazan N Billeh, Binghuang Cai, Sergey L Gratiy, Kael Dai, Ramakrishnan Iyer, Nathan W
537 Gouwens, Reza Abbasi-Asl, Xiaoxuan Jia, Joshua H Siegle, Shawn R Olsen, et al. Systematic
538 integration of structural and functional data into multi-scale models of mouse primary visual
539 cortex. *bioRxiv*, page 662189, 2019.
- 540 [48] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain
541 regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.
- 542 [49] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*
543 *opinion in neurobiology*, 46:1–6, 2017.
- 544 [50] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-
545 dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.
- 546 [51] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like
547 dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*,
548 page 696088, 2019.

- 549 [52] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
550 13(8):1713–1720, 2001.
- 551 [53] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
552 A statistical paradigm for neural spike train decoding applied to position prediction from
553 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
554 7425, 1998.
- 555 [54] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
556 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 557 [55] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
558 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis
559 of neural population activity. In *Advances in neural information processing systems*, pages
560 1881–1888, 2009.
- 561 [56] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
562 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
563 *Science*, 349(6244):184–187, 2015.
- 564 [57] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
565 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th Inter-*
566 *national Conference on Machine Learning*, 2019.
- 567 [58] Steven H Strogatz. Nonlinear dynamics and chaos: with applications to physics. *Biology,*
568 *Chemistry, and Engineering (Studies in Nonlinearity)*, Perseus, Cambridge, UK, 1994.
- 569 [59] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial*
570 *Intelligence and Statistics*, pages 814–822, 2014.
- 571 [60] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
572 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 573 [61] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
574 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 575 [62] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for
576 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

577 **A Acknowledgements**

578 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
 579 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award DBI-
 580 1707398, and The Gatsby Charitable Foundation. Helpful conversations were had with Francesca
 581 Mastrogiovanni, Srdjan Ostojic, James Fitzgerald, Stephen Baccus, Dhruva Raman, Mehrdad
 582 Jazayeri, Liam Paninski, and Larry Abbott.

583 **B Methods**

584 **B.1 Emergent property inference (EPI)**

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

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

591 Precisely, over the EPI distribution of parameters $q_\theta(z)$ and distribution of simulated activity
 592 $p(x | z)$, the emergent property statistics $T(x)$ must equal the emergent property values μ on
 593 average. This is a viable way to represent emergent properties in theoretical models, as we have
 594 demonstrated in the main text, and enables the EPI optimization.

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

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

599 Given a simulator defined by a theoretical model $x \sim p(x | z)$ and some emergent property of
 600 interest \mathcal{B} , $q_\theta(z)$ is optimized via the neural network parameters θ to find an optimally entropic

601 distribution q_θ^* within the deep variational family \mathcal{Q} producing the emergent property:

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

602 Since we are optimizing parameters θ of our deep probability distribution with respect to the entropy
 603 $H(q_\theta(z))$, we will need to take gradients with respect to the log probability density of samples from
 604 the deep probability distribution.

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

605

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

606 This optimization is done using the approach of MEFN [20], using architectures for deep proba-
 607 bility distributions, called normalizing flows (see Section B.1.3), conferring a tractable calculation
 608 of sample probability. In EPI, this methodology for learning maximum entropy distributions is
 609 repurposed toward variational learning of model parameter distributions. Similar to LFVI [21], we
 610 are motivated to do variational learning in models with intractable likelihood functions, in which
 611 standard methods like stochastic gradient variational Bayes [6] or black box variational inference[59]
 612 are not tractable. Furthermore, EPI focuses on setting mathematically defined emergent property
 613 statistics to emergent property values of interest, whereas LFVI is focused on learning directly from
 614 datasets. Optimizing this objective is a technological challenge, the details of which we elaborate
 615 in Section B.1.2. Before going through those details, we ground this optimization in a toy example.

616 B.1.1 Example: 2D LDS

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

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

618 with

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

619 To do EPI with the dynamics matrix elements as the free parameters $z = [a_1 \ a_2 \ a_3 \ a_4]$ (fixing
 620 $\tau = 1$), the emergent property statistics $T(x)$ were chosen to contain the first- and second-moments
 621 of the oscillatory frequency ω and the growth/decay factor d of the oscillating system. To learn the
 622 distribution of real entries of A that yield a distribution of d with mean zero with variance 0.25²,

and oscillation frequency ω 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. λ_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 λ_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 λ_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_θ^* directly. This is due to the formally hard problem of the backward mapping: finding the natural parameters η from the mean parameters μ of an exponential family distribution [60]. 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.

To explain the structure in the bimodality of the EPI distribution, we examined the imaginary

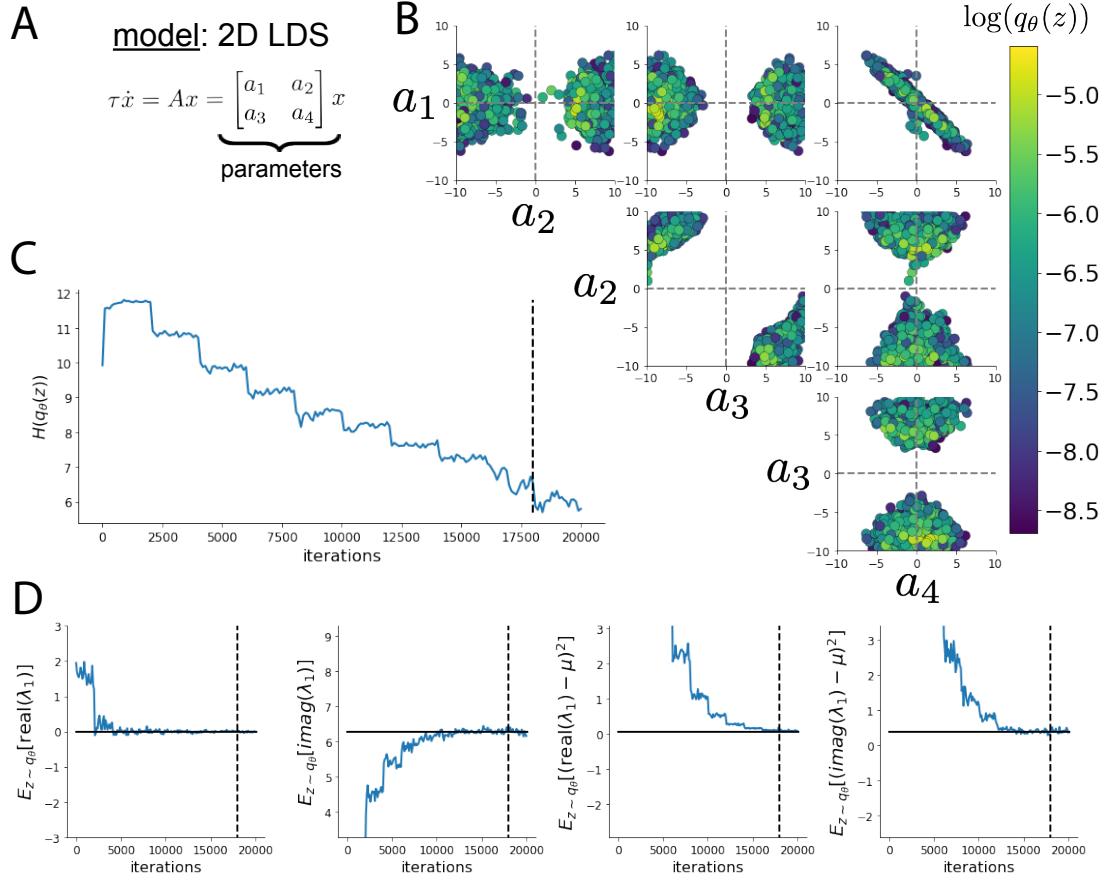


Fig. S1: A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The DSN distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. C. Entropy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

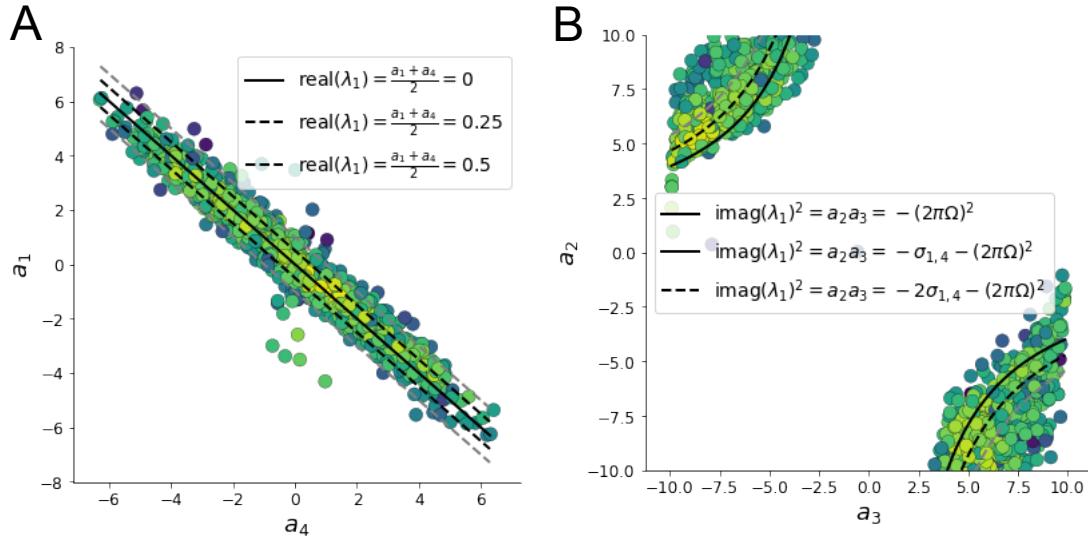


Fig. S2: A. Probability contours in the $a_1 - a_4$ plane can be derived from the relationship to emergent property statistic of growth/decay factor. B. Probability contours in the $a_2 - a_3$ plane can be derived from relationship to the emergent property statistic of oscillation frequency.

647 component of λ_1 . When $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, we have

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

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

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

650 Since we constrained $\mathbb{E}_{z \sim q_\theta} [\text{imag}(\lambda)] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation
651 $\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 = (2\pi)^2$ for various $a_1 a_4$ (Fig. S2A). If $\sigma_{1,4} = \mathbb{E}_{z \sim q_\theta} [|a_1 a_4 - E_{q_\theta}[a_1 a_4]|)$,
652 then we plot the contours as $a_1 a_4 = 0$ (black), $a_1 a_4 = -\sigma_{1,4}$ (black dotted), and $a_1 a_4 = -2\sigma_{1,4}$
653 (grey dotted) (Fig. S2B). This validates the curved structure of the inferred distribution learned
654 through EPI. We take steps in negative standard deviation of $a_1 a_4$ (dotted and gray lines), since
655 there are few positive values $a_1 a_4$ in the learned distribution. Subtler model-emergent property
656 combinations will have even more complexity, further motivating the use of EPI for understanding
657 these systems. As we expect, the distribution results in samples of two-dimensional linear systems
658 oscillating near 1Hz (Fig. S3).

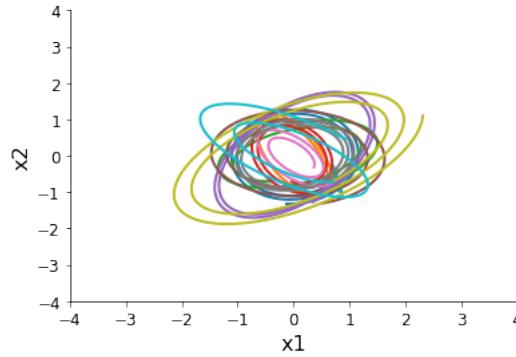


Fig. S3: Sampled dynamical system trajectories from the EPI distribution. Each trajectory is initialized at $x(0) = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$.

⁶⁵⁹ **B.1.2 Augmented Lagrangian optimization**

⁶⁶⁰ To optimize $q_\theta(z)$ in Equation 14, the constrained optimization is performed using the augmented
⁶⁶¹ Lagrangian method. The following objective is minimized:

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

⁶⁶² 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
⁶⁶³ related to the natural parameters of exponential families (see Section B.1.4)) and c is the penalty
⁶⁶⁴ coefficient. For a fixed (η, c) , θ is optimized with stochastic gradient descent. A low value of
⁶⁶⁵ c is used initially, and increased during each augmented Lagrangian epoch, which is a period of
⁶⁶⁶ optimization with fixed η and c for a given number of stochastic optimization iterations. Similarly,
⁶⁶⁷ η is tuned each epoch based on the constraint violations. For the linear two-dimensional system
⁶⁶⁸ (Fig. S1C), optimization hyperparameters are initialized to $c_1 = 10^{-4}$ and $\eta_1 = \mathbf{0}$. The penalty
⁶⁶⁹ coefficient is updated based on the result of a hypothesis test regarding the reduction in constraint
⁶⁷⁰ 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 βc_k
⁶⁷¹ with probability $1 - p$. Throughout the study, $\beta = 4.0$ and $\gamma = 0.25$ were used. The other update
⁶⁷² 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
⁶⁷³ for 2,000 iterations. We consider the optimization to have converged when a null hypothesis test of
⁶⁷⁴ constraint violations being zero is accepted for all constraints at a significance threshold 0.05. This
⁶⁷⁵ is the dotted line on the plots below depicting the optimization cutoff of EPI for the 2-dimensional
⁶⁷⁶ linear system.

⁶⁷⁷ The intention is that c and η start at values encouraging entropic growth early in optimization.

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

681 **B.1.3 Normalizing flows**

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

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

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

691 **B.1.4 Emergent property inference as variational inference in an exponential family**

692 Consider the goal of doing variational inference with an exponential family posterior distribution
 693 $p(z | x)$. We use the following abbreviated notation to collect the base measure $b(z)$ and sufficient
 694 statistics $T(z)$ into $\tilde{T}(z)$ and likewise concatenate a 1 onto the end of the natural parameter $\tilde{\eta}(x)$.
 695 The log normalizing constant $A(\eta(x))$ remains unchanged.

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

696 Variational inference with an exponential family posterior distribution uses optimization to mini-
 697 mize the following divergence [62]:

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

698 $q_\theta(z)$ is the variational approximation to the posterior with variational parameters θ . We can write
 699 this KL divergence in terms of entropy of the variational approximation.

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

700

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

701 As far as the variational optimization is concerned, the log normalizing constant is independent of
 702 $q_\theta(z)$, so it can be dropped.

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

703 Further, we can write the objective in terms of the first moment of the sufficient statistics $\mu =$
 704 $\mathbb{E}_{z \sim p(z|x)} [T(z)]$.

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

705

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

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

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

707 The Lagrangian objective (without the augmentation) is

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

708 As the optimization proceeds, η_{opt}^\top should converge to the natural parameter $\tilde{\eta}(x)$ through its
 709 adaptations in each epoch (see Section B.1.2).

710 The derivation of the natural parameter $\tilde{\eta}(x)$ of an exponential family distribution from its mean
 711 parameter μ is referred to as the backward mapping and is formally hard to identify [60]. Since
 712 this backward mapping is deterministic, we can replace the notation of $p(z | x)$ with $p(z | \mathcal{B})$
 713 conceptualizing an inferred distribution that obeys emergent property \mathcal{B} (see Section B.1).

714 B.2 Theoretical models

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

717 **B.2.1 Stomatogastric ganglion**

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

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

719 The membrane potential of each neuron is affected by the leak, calcium, potassium, hyperpolariza-
 720 tion, electrical and synaptic currents, respectively. The capacitance of the cell membrane was set to
 721 $C_m = 1nF$. Each current is a function of the neuron's membrane potential x_m and the parameters
 722 of the circuit such as g_{el} and g_{syn} , whose effect on the circuit is considered in the motivational
 723 example of EPI in Fig. 1. Specifically, the currents are the difference in the neuron's membrane
 724 potential and that current type's reversal potential multiplied by a conductance:

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

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

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

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

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

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

730 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 731 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 732 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. Fast: $g_{Ca} =$
 733 1.9×10^{-2} , $g_K = 3.9 \times 10^{-2}$, and $g_{hyp} = 2.5 \times 10^{-2}$. Intermediate: $g_{Ca} = 1.7 \times 10^{-2}$, $g_K = 1.9 \times 10^{-2}$,
 734 and $g_{hyp} = 8.0 \times 10^{-3}$. Intermediate: $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

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

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

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

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

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

740

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

741

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

742

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

743 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 744 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$. These are the same parameter
 745 values used in [23].

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

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

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

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

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

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

762 Finally, to differentiate through the maximum frequency identification, we used a sum-of-powers
 763 normalization. Let $\mathcal{X}_i \in \mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal
 764 $x_i \in \mathbb{R}^N$, where $i \in \{f1, f2, \text{hub}, s1, s2\}$. The “frequency identification” vector is

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

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

766 Network syncing, like all other emergent properties in this work, are defined by the emergent
 767 property statistics and values. The emergent property statistics are the first- and second-moments
 768 of the firing frequencies. The first moments are set to 0.542Hz, while the second moments are set
 769 to 0.025Hz².

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

770 For EPI in Fig 2C, we used a real NVP architecture with two coupling layers. Each coupling layer
 771 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
 772 same considered in [23]). We have shown the EPI optimization that converged with maximum
 773 entropy across 5 random seeds and augmented Lagrangian coefficient initializations of $c_0 \in \{5, 10\}$.

774 B.2.2 Primary visual cortex

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

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

776 Some neuron-types largely lack synaptic projections to other neuron-types [43], and it is popular
 777 to only consider a subset of the effective connectivities [24, 44, 45].

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

778 By consolidating information from many experimental datasets, Billeh et al. [47] produce estimates
 779 of the synaptic strength (in mV)

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

780 and connection probability

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

781 Multiplying these connection probabilities and synaptic efficacies gives us an effective connectivity
 782 matrix:

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

783 We used the entries of this full effective connectivity matrix that are not considered to be ineffectual
 784 (Equation 52).

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

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

789 All simulations of this system had $T = 100$ time points, a time step $dt = 5\text{ms}$, and time constant
 790 $\tau = 20\text{ms}$. And the system was initialized to a random draw $x(0)_i \sim \mathcal{N}(1, 0.01)$.

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

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

792 where the input to each neuron is

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

⁷⁹³ Let $F_{ij} = \gamma_i \delta(i, j)$, where $\gamma_i = f'(u_i)$. Then, the linear response is

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

⁷⁹⁴ which is calculable by

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

⁷⁹⁵ This calculation is used to produce the magenta lines in Figure 2C, which show the linearly predicted
⁷⁹⁶ inputs that generate a response from two standard deviations (of \mathcal{B}) below and above y .

⁷⁹⁷ The emergent property we considered was the first and second moments of the change in steady
⁷⁹⁸ state rate dx_{ss} between the baseline input $h = b$ and $h = b + dh$. We use the following notation to
⁷⁹⁹ indicate that the emergent property statistics were set to the following values:

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

⁸⁰⁰ In the final analysis for this model, we sweep the input one neuron at a time away from the mode
⁸⁰¹ of each inferred distributions $dh^* = z^* = \text{argmax}_z \log q_\theta(z \mid \mathcal{B}(\alpha, 0.1))$. The differential responses
⁸⁰² $\delta x_{\alpha,ss}$ are examined at perturbed inputs $h = b + dh^* + \delta h_\alpha \hat{u}_\alpha$ where \hat{u}_α is a unit vector in the
⁸⁰³ dimension of α and $\delta h_\alpha \in [-15, 15]$.

⁸⁰⁴ 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
⁸⁰⁵ initial seeds using an architecture of four coupling layers, each with two hidden layers of 10 units.
⁸⁰⁶ We set $c_0 = 10^5$. The support of the learned distribution was restricted to $z_i \in [-5, 5]$.

⁸⁰⁷ B.2.3 Superior colliculus

⁸⁰⁸ In the model of Duan et al [25], there are four total units: two in each hemisphere corresponding to
⁸⁰⁹ the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right
⁸¹⁰ Pro (RP) and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related
⁸¹¹ by

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

⁸¹² where $\alpha \in \{LP, LA, RA, RP\}$ $\epsilon = 0.05$ and $\zeta = 0.5$ control the position and shape of the nonlin-
⁸¹³ earity, repsectively.

814 We order the elements of x and u in the following manner

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

815 The internal variables follow dynamics:

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

816 with time constant $\tau = 0.09s$ and Gaussian noise σdB controlled by the magnitude of $\sigma = 1.0$. The
817 weight matrix has 8 parameters sW_P , sW_A , vW_{PA} , vW_{AP} , hW_P , hW_A , dW_{PA} , and dW_{AP} (Fig.
818 4B).

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

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

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

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

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

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

822 a choice-period input,

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

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

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

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

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

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

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

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

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

832

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

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

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

835

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

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

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

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

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

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

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

855 B.2.4 Rank-1 RNN

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

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

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

⁸⁶⁷ Δ_0 , structured variance Δ_∞ , and the chaotic variance Δ_T) are written as functions of one another
⁸⁶⁸ in terms of connectivity parameters. The values of these variables can be used obtained using a
⁸⁶⁹ nonlinear system of equations solver. These dynamic mean field variables are then cast as task-
⁸⁷⁰ relevant variables with respect to the context of the provided inputs. Mastrogiuseppe et al. designed
⁸⁷¹ low-rank RNN connectivities via minimalist connectivity parameters to solve canonical tasks from
⁸⁷² behavioral neuroscience.

⁸⁷³ We consider the DMFT equation solver as a black box that takes in a low-rank parameterization
⁸⁷⁴ z (e.g. $z = [g \quad M_m \quad M_n]$) and outputs the values of the dynamic mean field variables, of which
⁸⁷⁵ we cast κ_r and Δ_T as task-relevant variables μ_{post} and σ_{post}^2 in the Gaussian posterior conditioning
⁸⁷⁶ toy example. Importantly, the solution produced by the solver is differentiable with respect to the
⁸⁷⁷ input parameters, allowing us to use DMFT to calculate the emergent property statistics in EPI
⁸⁷⁸ to learn distributions on such connectivity parameters of RNNs that execute tasks.

⁸⁷⁹ Specifically, we solve for the mean field variables κ_r , κ_n , Δ_0 and Δ_∞ , where the readout is nominally
⁸⁸⁰ chosen to point in the unit orthant $r = [1 \quad \dots \quad 1]^\top$. The consistency equations for these variables
⁸⁸¹ in the presence of an constant input $h(t) = y - (n - M_n)$ can be derived following [26] are

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

⁸⁸² where z here is a gaussian integration variable. We can solve these equations by simulating the
⁸⁸³ following Langevin dynamical system to a steady state.

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

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

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

886 **B.3 Supplementary Figures**

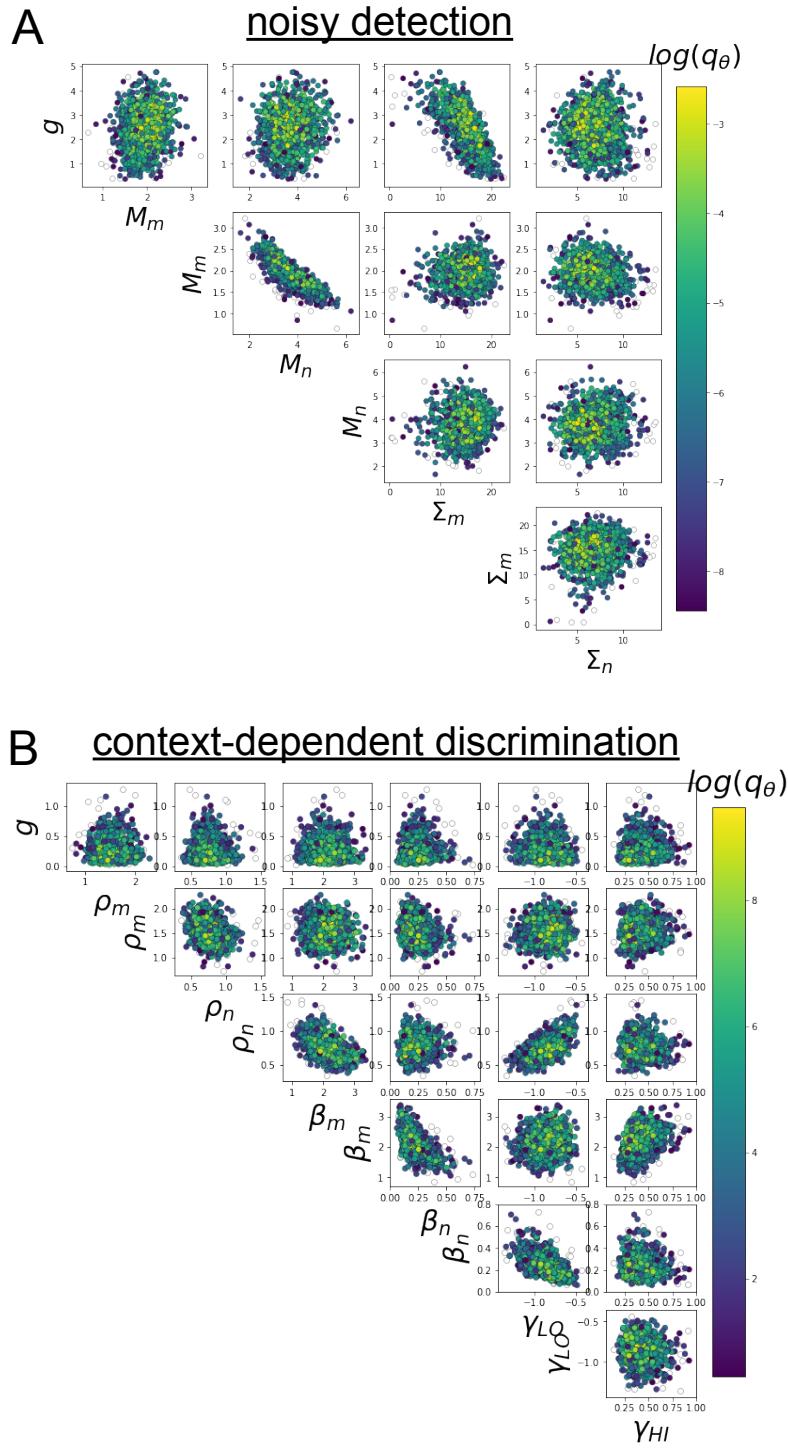


Fig. S4: A. EPI for rank-1 networks doing discrimination. B. EPI for rank-2 networks doing context-dependent discrimination. See [26].