

Interrogating theoretical models of neural computation with deep inference
Sean R. Bittner¹, Agostina Palmigiano¹, Alex T. Piet^{2,3,4}, Chunyu A. Duan⁵, Carlos D. Brody^{2,3,6},
Kenneth D. Miller¹, and John P. Cunningham⁷.

¹Department of Neuroscience, Columbia University,

²Princeton Neuroscience Institute,

³Princeton University,

⁴Allen Institute for Brain Science,

⁵Institute of Neuroscience, Chinese Academy of Sciences,

⁶Howard Hughes Medical Institute,

⁷Department of Statistics, Columbia University

¹ 1 Abstract

² A cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures a
³ hypothesized neural mechanism. Such models are valuable when they give rise to an experimentally
⁴ 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. When analytic derivation of the relationship between model
⁷ parameters and computational properties is intractable, approximate inference and simulation-
⁸ based techniques are relied upon for scientific insight. We bring the use of deep generative models
⁹ for probabilistic inference to bear on this problem, learning complex distributions of parameters
¹⁰ that produce the specified properties of computation. Our novel method solves the inverse problem
¹¹ by identifying the full space of parameters producing the emergent property. We motivate this
¹² methodology with a worked example analyzing sensitivity in the stomatogastric ganglion. We then
¹³ use it to reveal the key factors of variability in a model of primary visual cortex, gain a mechanistic
¹⁴ understanding of rapid task switching in superior colliculus models, and scale inference of large
¹⁵ low-rank RNN’s exhibiting stable amplification. This work illustrates how we can further leverage
¹⁶ the power of deep learning towards solving inverse problems in theoretical neuroscience.

₁₇ **2 Introduction**

₁₈ The fundamental practice of theoretical neuroscience is to use a mathematical model to understand
₁₉ neural computation, whether that computation enables perception, action, or some intermediate
₂₀ processing. A neural computation is systematized with a set of equations – the model – and
₂₁ these equations are motivated by biophysics, neurophysiology, and other conceptual considerations
₂₂ [1, 2, 3, 4]. The function of this system is governed by the choice of model *parameters*, which when
₂₃ configured in a particular way, give rise to a measurable signature of a computation. The work
₂₄ of analyzing a model then requires solving the inverse problem: given a computation of interest,
₂₅ how can we reason about particular parameter configurations? The inverse problem is crucial for
₂₆ reasoning about likely parameter values, uniquenesses and degeneracies, and predictions made by
₂₇ the model [5, 6].

₂₈ Consider the idealized practice: one carefully designs a model and analytically derives how com-
₂₉ putational properties determine model parameters. Seminal examples of this gold standard (which
₃₀ often adopt approaches from statistical physics) include our field’s understanding of memory ca-
₃₁ pacity in associative neural networks [7], chaos and autocorrelation timescales in random neural
₃₂ networks [8], the paradoxical effect [9], and decision making [10]. Unfortunately, as circuit models
₃₃ include more biological realism, theory via analytical derivation becomes intractable. Alternatively,
₃₄ we can gain insight into these complex models by identifying the full distribution of parameters con-
₃₅ sistent with specified emergent phenomena. By solving the inverse problem in this way, scientists
₃₆ can reason about the sensitivity and robustness of the model with respect to different parameter
₃₇ combinations [11, 12, 13, 6, 14].

₃₈ The preferred formalism for parameter identification in science is statistical inference, which has
₃₉ been used to great success in neuroscience through the stipulation of statistical generative models
₄₀ [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29] (see review, [30]). However, most neural
₄₁ circuit models in theoretical neuroscience stipulate a noisy system of differential equations that can
₄₂ only be sampled or realized through forward simulation; they lack the explicit likelihood central to
₄₃ the probabilistic modeling toolkit. Therefore, the most popular approaches to the inverse problem
₄₄ have been likelihood-free methods such as approximate Bayesian computation (ABC) [31, 32], in
₄₅ which reasonable parameters are obtained via simulation and rejection.

₄₆ Of course, the challenge of doing inference in complex models has arisen in many scientific fields.
₄₇ In response, the machine learning community has made remarkable progress in recent years, via

48 the use of deep neural networks as powerful inference engines: a flexible function family that can
49 map observations back to probability distributions quantifying the likely parameter configurations.
50 One celebrated example of this approach from machine learning, of which we draw key inspiration
51 for this work, is the variational autoencoder (VAE) [33, 34], which uses a deep neural network
52 to induce an (approximate) posterior distribution on hidden variables in a latent variable model,
53 given data. Indeed, these tools have been used to great success in neuroscience as well, in particular
54 for interrogating hidden states in models of both cortical population activity [35, 36, 37, 38] and
55 animal behavior [39, 40, 41]. These works have used deep neural networks to expand the domain
56 of neural data sets amenable to statistical modeling [30].

57 Existing approaches to the inverse problem in theoretical neuroscience fall short in three key ways.
58 First, theoretical models of neural computation aim to reflect a complex biological reality, and as
59 a result, such models lack tractable likelihoods. Without an efficient calculation of the probability
60 of model properties given model parameters, neuroscientists resort to approximate Bayesian com-
61 putation [42, 43, 31], which requires a rejection heuristic, scales poorly, and only produces sets of
62 accepted parameters lacking probabilities. Second, there is an undesirable trade-off between the
63 flexibility and sampling speed of approximated posterior distributions. Sampling-based inference
64 approaches (e.g. ABC and Markov chain Monte Carlo (MCMC) [44, 45]) confer flexible approxima-
65 tions, yet scale poorly in number of parameters. While variational inference (VI) [46] often results
66 in fast posterior sampling, existing practice relies heavily on simplified classes of distributions [47].
67 Third, such parameter inference methods are designed to operate on experimentally collected data-
68 sets. Ultimately, the objects of interest in theoretical neuroscience are phenomena or features of
69 the model rather than singular data-sets.

70 To address these three challenges, we developed an inference methodology – ‘emergent property
71 inference’ – which learns a distribution over parameter configurations in a theoretical model. This
72 distribution has two critical properties: *(i)* it is chosen such that draws from the distribution (pa-
73 rameter configurations) correspond to systems of equations that give rise to a specified emergent
74 property (a set of constraints); and *(ii)* it is chosen to have maximum entropy given those con-
75 straints, such that we identify all likely parameters and can use the distribution to reason about
76 parametric sensitivity and degeneracies [48]. First, we use stochastic gradient techniques in the
77 spirit of likelihood-free variational inference [49] to enable inference in likelihood-free models of neu-
78 ral computation. Second, we stipulate a bijective deep neural network that induces a flexible family
79 of probability distributions over model parameterizations with a probability density we can calcu-

80 late [47, 50, 51], which confers fast sampling and sensitivity measurements. Third, we quantify the
81 notion of emergent properties as a set of moment constraints on datasets generated by the model.
82 Thus, an emergent property is not a single data realization, but a phenomenon or a feature of the
83 model. Conditioning on an emergent property requires a variant of deep probabilistic inference
84 methods, which we have previously introduced [52]. Taken together, emergent property inference
85 (EPI) provides a methodology for inferring parameter configurations consistent with a particular
86 emergent phenomena in theoretical models. We use a classic example of parametric degeneracy in
87 a biological system, the stomatogastric ganglion [53], to motivate and clarify the technical details
88 of EPI.

89 Equipped with this methodology, we then investigated three models of current importance in the-
90 oretical neuroscience. These models were chosen to demonstrate generality through ranges of bi-
91 ological realism (from conductance-based biophysics to recurrent neural networks), neural system
92 function (from pattern generation to decision making), and network scale (from four to hundreds of
93 neurons). First, we use EPI to understand the characteristics of noise across multiple neuron-type
94 populations that govern variability in a model of primary visual cortex. Then, we use EPI to infer
95 multiple regimes of superior colliculus connectivity that perform rapid task switching. The novel
96 scientific insights offered by EPI contextualize and clarify the previous studies exploring these mod-
97 els [54, 55]. Finally, we emphasize the scalability of EPI by inferring high-dimensional distributions
98 of RNNs exhibiting stable amplification. These results point to the value of deep inference for the
99 interrogation of biologically relevant models.

100 3 Results

101 3.1 Motivating emergent property inference of theoretical models

102 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-
103 erty inference. First, one designs or chooses an existing model that, it is hypothesized, captures
104 the computation of interest. To ground this process in a well-known example, consider the stom-
105 atogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic
106 muscle activation patterns for digestion [56]. Despite full knowledge of STG connectivity and a
107 precise characterization of its rhythmic pattern generation, biophysical models of the STG have
108 complicated relationships between circuit parameters and neural activity [53, 12]. A subcircuit
109 model of the STG [57] is shown schematically in Figure 1A, and note that the behavior of this

model will be critically dependent on its parameterization – the choices of conductance parameters $\mathbf{z} = [g_{el}, g_{synA}]$. Specifically, the two fast neurons (f_1 and f_2) mutually inhibit one another, and oscillate at a faster frequency than the mutually inhibiting slow neurons (s_1 and s_2). The hub neuron (hub) couples with either the fast or slow population or both.

Second, once the model is selected, one defines the emergent phenomena of scientific interest. In the STG example, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model 1B. An interesting emergent property of this stochastic model is when the hub neuron fires at an intermediate frequency between the intrinsic spiking rates of the fast and slow populations. This emergent property is shown in Figure 1C at an average frequency of 0.55Hz.

Third, parameter analyses ensue: brute-force parameter sweeps, ABC sampling, and sensitivity analyses are all routinely used to reason about what parameter configurations lead to an emergent property. In this last step lies the opportunity for a precise quantification of the emergent property as a statistical feature of the model. Once we have such a methodology, we can infer a probability distribution over parameter configurations that produce this emergent property.

Before presenting technical details (in the following section), let us understand emergent property inference schematically: EPI (Fig. 1D) takes, as input, the model and the specified emergent property, and as its output, produces the parameter distribution EPI (Fig. 1E). This distribution – represented for clarity as samples from the distribution – is then a scientifically meaningful and mathematically tractable object. In the STG model, this distribution can be specifically queried to reveal the prototypical parameter configuration for network syncing (the mode; Figure 1E yellow star), and how network syncing decays based on changes away from the mode. The eigenvectors (of the Hessian of the distribution at the mode) quantitatively formalize the robustness of intermediate hub frequency (Fig. 1E solid (v_1) and dashed (v_2) black arrows). Indeed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity (v_1) and degeneracy (v_2) agree with error contours (Fig. 1E contours) and have diminished or preserved hub frequency, respectively (Fig. 1F activity traces) (see Section 5.2.1).

3.2 A deep generative modeling approach to emergent property inference

Emergent property inference (EPI) systematizes the three-step procedure of the previous section. First, we consider the model as a coupled set of differential equations [57]. In the running STG example, the model activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$ is the membrane potential for each neuron,

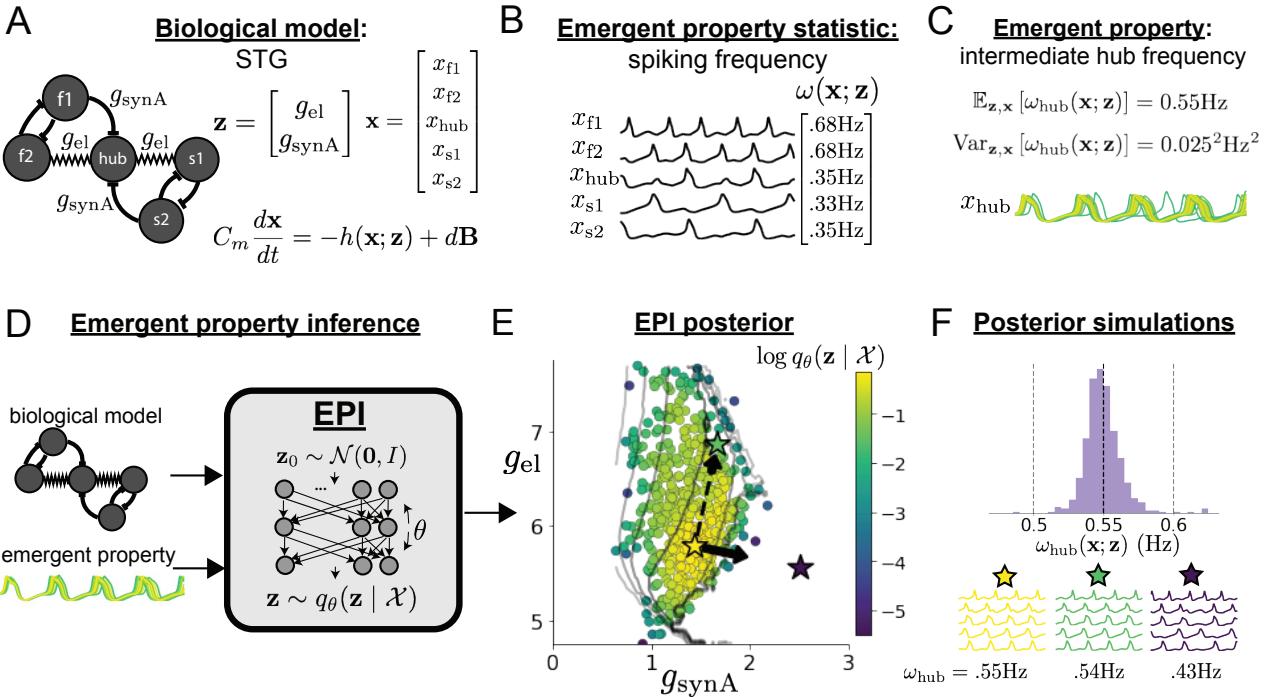


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based biophysical model of the STG subcircuit. In the STG model, jagged connections indicate electrical coupling having electrical conductance g_{el} . Other connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto the hub neuron, and $g_{synB} = 5\text{nS}$ for mutual inhibitory connections. Parameters are represented by the vector \mathbf{z} and membrane potentials by the vector \mathbf{x} . The evolution of this model's activity $\mathbf{x}(t)$ is predicated by differential equations. **B.** Spiking frequency $\omega(\mathbf{x}; \mathbf{z})$ is an emergent property statistic. In this example, spiking frequency is measured from simulated activity of the STG model at parameter choices of $g_{el} = 4.5\text{nS}$ and $g_{synA} = 3\text{nS}$. **C.** The emergent property of intermediate hub frequency, in which the hub neuron fires at a rate between the fast and slow frequencies. This emergent property is defined by a mean and variance on the emergent property statistic. Simulated activity traces are colored by log probability density of their generating parameters in the EPI-inferred distribution (Panel E). **D.** For a choice of model and emergent property, emergent property inference (EPI) learns a deep probability distribution of parameters \mathbf{z} . Deep probability distributions map a simple random variable $\mathbf{z}_0 \sim \mathcal{N}(0, I)$ through a deep neural network with weights and biases $\boldsymbol{\theta}$ to parameters $\mathbf{z} = q_{\boldsymbol{\theta}}(\mathbf{z}_0)$. In EPI optimization, stochastic gradient steps in $\boldsymbol{\theta}$ are taken such that entropy is maximized, and the emergent property \mathcal{X} is produced. The EPI posterior distribution is denoted $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X})$. **E.** The EPI posterior producing intermediate hub frequency. Samples are colored by log probability density. Distribution contours of average hub neuron frequency from mean of .55 Hz are shown at levels of .525, .53,575 Hz (dark to light gray away from mean). Eigenvectors of the Hessian at the mode of the inferred distribution are indicated as \mathbf{v}_1 (solid) and \mathbf{v}_2 (dashed) with lengths scaled by the square root of the absolute value of their eigenvalues. **F** Simulations from parameters in E. (Top) The predictive distribution of the posterior obeys the emergent property. The black and gray dashed lines show the mean and two standard deviations according the emergent property, respectively. (Bottom) Simulations at the starred parameter values.

140 which evolves according to the biophysical conductance-based equation:

$$C_m \frac{d\mathbf{x}(t)}{dt} = -h(\mathbf{x}(t); \mathbf{z}) + d\mathbf{B} \quad (1)$$

141 where $C_m = 1\text{nF}$, and \mathbf{h} is a sum of the leak, calcium, potassium, hyperpolarization, electrical, and
142 synaptic currents, all of which have their own complicated dependence on \mathbf{x} and $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$,
143 and $d\mathbf{B}$ is white gaussian noise (see Section 5.2.1).

144 Second, we define the emergent property, which as above is “intermediate hub frequency” (Figure
145 1C). Quantifying this phenomenon is straightforward: we stipulate that the hub neuron’s spiking
146 frequency – denoted $\omega_{\text{hub}}(\mathbf{x})$ is close to an intermediate frequency of 0.55Hz. Mathematically, we
147 achieve this via constraints on the mean and variance of the hub neuron spiking frequency.

$$\begin{aligned} \mathcal{X} &: \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] \triangleq \mathbb{E}_{\mathbf{z}, \mathbf{x}} [\omega_{\text{hub}}(\mathbf{x}; \mathbf{z})] = [0.55] \triangleq \boldsymbol{\mu} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] &\triangleq \text{Var}_{\mathbf{z}, \mathbf{x}} [\omega_{\text{hub}}(\mathbf{x}; \mathbf{z})] = [0.025^2] \triangleq \boldsymbol{\sigma}^2. \end{aligned} \quad (2)$$

148 The emergent property statistic $f(\mathbf{x}; \mathbf{z}) = \omega_{\text{hub}}(\mathbf{x}; \mathbf{z})$ along with its constrained mean $\boldsymbol{\mu}$ and variance
149 $\boldsymbol{\sigma}^2$ define the emergent property denoted \mathcal{X} .

150 Third, we perform emergent property inference: we find a distribution over parameter configura-
151 tions \mathbf{z} , and insist that samples from this distribution produce the emergent property; in other
152 words, they obey the constraints introduced in Equation 2. This distribution will be chosen from a
153 family of probability distributions $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$, defined by a deep generative distribution
154 of the normalizing flow class [47, 50, 51] – neural networks which transform a simple distribution
155 into a suitably complicated distribution (as is needed here). This deep distribution is represented
156 in Figure 1C (see Section 5.1). Then, mathematically, we must solve the following optimization
157 program:

$$\begin{aligned} q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X}) &= \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\text{argmax}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) \\ \text{s.t. } \mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] &= \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2 \end{aligned} \quad (3)$$

158 where $f(\mathbf{x}, \mathbf{z})$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ are defined as in Equation ???. According to the emergent property of
159 interest, $f(\mathbf{x}, \mathbf{z})$ may contain multiple statistics, in which case the mean and variance vectors $\boldsymbol{\mu}$
160 and $\boldsymbol{\sigma}^2$ match this dimension. Finally, we recognize that many distributions in \mathcal{Q} will respect
161 the emergent property constraints, so we select that which has maximum entropy. This principle,
162 captured in Equation 3 by the primal objective H , identifies parameter distributions with minimal

assumptions beyond some chosen structure [58, 59, 52, 60]. Such a normative principle of maximum entropy, which is also that of Bayesian inference, naturally fits with our scientific objective of reasoning about parametric sensitivity and robustness. The recovered distribution of EPI is as variable as possible along each parametric manifold such that it produces the emergent property.

EPI optimizes the weights and biases θ of the deep network (which induces the probability distribution) by iteratively solving Equation 3. The optimization is complete when the sampled models with parameters $\mathbf{z} \sim q_\theta(z | \mathcal{X})$ produce activity consistent with the specified emergent property (Fig. S4). Such convergence is evaluated with a hypothesis test that the means and variances of each emergent property statistic are not different than their constrained values (see Section 5.1.3). Further validation of EPI is available in the supplementary materials, where we analyze a simpler model for which ground-truth statements can be made (Section 5.1.4).

In relation to broader methodology, inspection of the EPI objective reveals a natural relationship to posterior inference. Specifically, EPI executes a novel variant of Bayesian inference with a uniform prior and a gaussian likelihood on the emergent property statistic (see Section 5.1.5). A key advantage of EPI over established Bayesian inference is that the predictions made by the inferred distribution are constrained to produce the specified emergent property. Equipped with this method, we may examine structure in posterior distributions or make comparisons between posteriors conditioned at different levels of the same emergent property statistic. In Sections 3.3 and 3.4, we prove out the value of EPI by using it to investigate and produce novel insights into two prominent models in neuroscience. Subsequently in Section 3.5, we show EPI’s superiority in parameter scalability and fidelity of the posterior predictive distribution by conditioning on stable amplification in low-rank RNNs.

3.3 EPI reveals how neuron-type specific noise governs variability in the stochastic stabilized supralinear network

Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output function have succeeded in explaining a host of experimentally documented phenomena. In a regime characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to paradoxical responses [9], selective amplification [61, 62], surround suppression [63] and normalization [64]. Despite their strong predictive power, E-I circuit models rely on the assumption that inhibition can be studied as an indivisible unit. However, experimental evidence shows that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) –

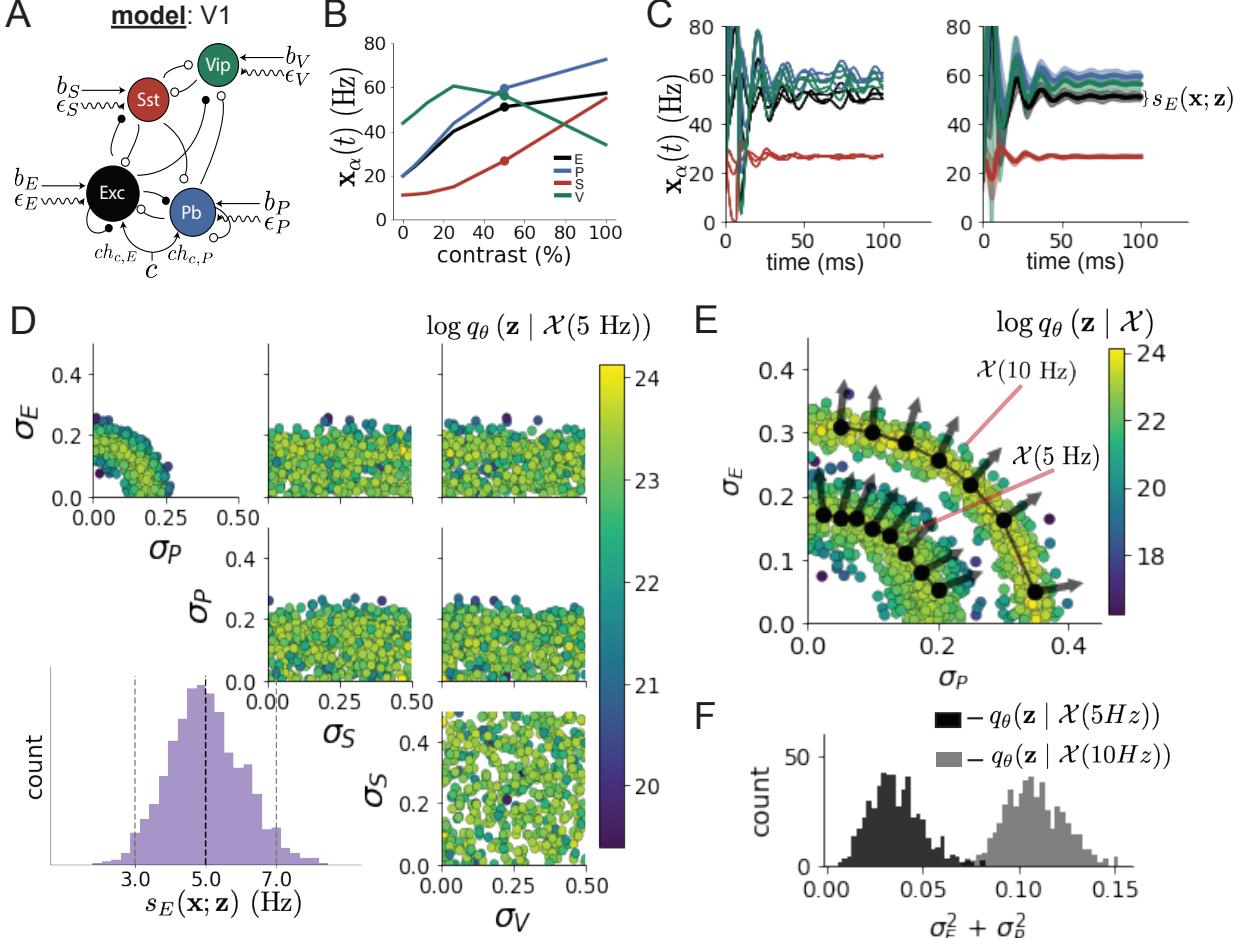


Figure 2: Emergent property inference in the stochastic stabilized supralinear network (SSSN) **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). Some neuron-types largely do not form synaptic projections to others ($|W_{\alpha_1, \alpha_2}| < 0.025$). Each neural population receives a baseline input \mathbf{h}_b , and the E- and P-populations also receive a contrast-dependent input \mathbf{h}_c . Additionally, each neural population receives a slow noisy input ϵ . **B.** Steady-state responses of the SSN model (deterministic, $\sigma = \mathbf{0}$) to varying contrasts. The response at 50% contrast (dots) is the focus of our analysis. **C.** Transient network responses of the SSSN model at 50 % contrast. (Left) Traces are independent trials with varying initialization $\mathbf{x}(0)$ and noise realization. (Right) Mean (solid line) and standard deviation (shading) of responses. **D.** EPI posterior of noise parameters \mathbf{z} conditioned on E-population variability. The posterior predictive distribution of $s_E(\mathbf{x}; \mathbf{z})$ is show on the bottom-left. **E.** (Top) Enlarged visualization of the σ_E - σ_P marginal distribution of the posteriors $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ and $q_\theta(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$. Each black dot shows the mode at each σ_P . The arrows show the most sensitive dimensions of the Hessian evaluated at these modes. **F.** The predictive distributions of $\sigma_E^2 + \sigma_P^2$ of each posterior $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ and $q_\theta(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$.

194 composing 80% of GABAergic interneurons in V1 [65, 66, 67], and that these inhibitory cell types
 195 follow specific connectivity patterns (Fig. 2A) [68]. Recent theoretical advances [54, 69, 70], have
 196 only started to address the consequences of this multiplicity in the dynamics of V1, strongly relying
 197 on linear theoretical tools. Here, we use EPI to analyze V1 models of greater complexity in order
 198 to characterize properties of slow noise governing circuit variability.

199 We considered the response properties of a nonlinear dynamical V1 circuit model (Fig. 2A) with
 200 a state comprised of each neuron-type population's rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population
 201 receives recurrent input $W\mathbf{x}$ from synaptic projections of effective connectivity W and an external
 202 input \mathbf{h} , which determine the population rate via supralinear nonlinearity $\phi = \|\cdot\|_+^2$. The input is
 203 also comprised of a slow noise component $\epsilon \sim OU(\tau_{\text{noise}}, \sigma)$ of time scale $\tau_{\text{noise}} > \tau$ and variance
 204 parameters σ (see Section 5.2.2)

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \epsilon). \quad (4)$$

205 This model is the stochastic stabilized supralinear network (SSSN) [71] generalized to have in-
 206 hibitory multiplicity, and introduces stochasticity to previous four neuron-type models of V1 [54].
 207 Stochasticity and inhibitory multiplicity introduce substantial complexity to mathematical deriva-
 208 tions (see Section 5.2.3) motivating the treatment of this model with EPI. Here, we consider fixed
 209 weights W and input \mathbf{h} according to a fit of the deterministic model to contrast responses [72] (Fig.
 210 2B), and study the effect of noise parameterization $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on fluctuations at 50%
 211 contrast.

212 For this SSSN, we are interested in how noise variability across neural populations governs stochastic
 213 fluctuations in the E-population. Here, we quantify different levels y of E-population variability
 214 with the emergent property

$$\begin{aligned} \mathcal{X}(y) : \mathbb{E}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= y \\ \text{Var}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= 1\text{Hz}^2, \end{aligned} \quad (5)$$

215 where $s_E(\mathbf{x}; \mathbf{z})$ is the standard deviation of the stochastic E-population response about its steady
 216 state (Fig. 2C).

217 We ran EPI to obtain a posterior distribution $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz})$ producing E-population variability
 218 around 5 Hz (Fig. 2D). From the marginal distribution of σ_E and σ_P (Fig. 2D, top-left), we can
 219 see that $s_E(\mathbf{x}; \mathbf{z})$ is sensitive to various combinations of σ_E and σ_P . Alternatively, both σ_S and σ_V
 220 are degenerate with respect to $s_E(\mathbf{x}; \mathbf{z})$ evidenced by the high variability in those dimensions of the
 221 posterior (Fig. 2D, bottom-right). Together, these observations imply a parametric manifold of

222 degeneracy with respect to $s_E(\mathbf{x}; \mathbf{z})$ of 5 Hz, which is indicated by the modes along σ_P in the σ_E - σ_P
 223 marginal (Fig. 2E). The dimensions of sensitivity conferred by EPI and this plain visual structure
 224 suggest a quadratic relationship in the emergent property statistic $s_E(\mathbf{x}; \mathbf{z})$ and parameters \mathbf{z} , which
 225 is preserved at a greater level of variability $\mathcal{X}(10 \text{ Hz})$ (Fig. 2E). Indeed, the sum of squares of σ_E
 226 and σ_P is larger in $q_{\theta}(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$ than $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ (Fig 2F, $p = 0$), while the sum of squares
 227 of σ_S and σ_V are not significantly different in the two posteriors (Fig. 11, $p = .402$).

228 While a quadratic relationship in $s_E(\mathbf{x}; \mathbf{z})$ and \mathbf{z} is potentially derivable by extending the derivation
 229 in Section 5.2.2 to the case of $\tau \neq \tau_{\text{noise}}$, the coefficients in front of each quadratic term would be
 230 unruly, and likely escape comprehensible analysis. This makes EPI an attractive tool for revealing
 231 the characteristics of noise governing variability and for answering other questions in this complex
 232 model. Intriguingly, this circuit exhibited a paradoxical effect in the P-population, and no other
 233 inhibitory types at 50% contrast (Fig. 11) implying that the E-population is P-stabilized. Future
 234 work motivated by our analysis here, may uncover a relationship between the neuron-type mediating
 235 stability and the factors governing circuit variability.

236 3.4 EPI identifies multiple regimes of rapid task switching

237 In a rapid task switching experiment [73], rats were explicitly cued on each trial to either orient
 238 towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti
 239 (A) task (Fig. 3A). Neural recordings in the midbrain superior colliculus (SC) exhibited two
 240 populations of neurons that simultaneously represented both task context (Pro or Anti) and motor
 241 response (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons
 242 [55]. Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is
 243 a four-population dynamical system. We analyzed this model, where the neuron-type populations
 244 are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right
 245 (R)), their connectivity is parameterized geometrically (Fig. 3B). The input-output function of
 246 this model is chosen such that the population responses $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^{\top}$ are bounded
 247 from 0 to 1 as a function ϕ of a dynamically evolving internal variable \mathbf{u} . The model responds to
 248 the side with greater Pro neuron activation; e.g. the response is left if $x_{LP} > x_{RP}$ at the end of
 249 the trial. The dynamics evolve with timescale $\tau = 90\text{ms}$ governed by connectivity weights W

$$\begin{aligned}
 \tau \frac{d\mathbf{u}}{dt} &= -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B} \\
 \mathbf{x} &= \phi(\mathbf{u})
 \end{aligned} \tag{6}$$

250 with white noise of variance 0.2^2 . The input \mathbf{h} is comprised of a cue-dependent input to the Pro
 251 or Anti populations, a stimulus orientation input to either the Left or Right populations, and a
 252 choice-period input to the entire network (see Section 5.2.4). Here, we use EPI to determine the
 253 network connectivity $\mathbf{z} = [sW, vW, dW, hW]^\top$ that produces rapid task switching behavior.

254 We define rapid task switching behavior as accurate execution of each task. Inferred models should
 255 not exhibit fully random responses (50%), or perfect performance (100%), since perfection is never
 256 attained by even the best trained rats. We formulate rapid task switching as an emergent property
 257 by stipulating that the average accuracy in the Pro task $p_P(\mathbf{x}; \mathbf{z})$ and Anti task $p_A(\mathbf{x}; \mathbf{z})$ be 75%
 258 with variance $7.5\%^2$.

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 75\% \\ 75\% \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 7.5\%^2 \\ 7.5\%^2 \end{bmatrix} \end{aligned} \quad (7)$$

259 A variance of $7.5\%^2$ in each task will confer a posterior producing performances ranging from about
 260 60% – 90%, allowing us to examine the properties of connectivity that yield better performance in
 261 each task. Notably, this is our first example using EPI to condition on multiple emergent property
 262 statistics ($|f(\mathbf{x}; \mathbf{z})| = 2$).

263 The EPI inferred parameters (Fig. 3C) generate a distribution of task accuracies (Fig. 3C, middle-
 264 left) according to our mathematical definition of rapid task switching (Equation 7). The nonlinear
 265 patterns of connectivity that govern each task accuracy (Fig. 12A-B) are not fully captured by
 266 linear prediction (Fig. 12C). For example, the patterns in connectivity increasing Pro accuracy
 267 change dramatically after crossing a threshold of sW (Fig. 12A sW - hW marginal). Not only has
 268 EPI captured this complex nonlinear posterior, it offers probabilistic tools for understanding the
 269 different regimes of model behavior.

270 To establish these two regimes of connectivity, we took gradient steps along $q_{\theta}(\mathbf{z} | \mathcal{X})$ to produce
 271 modes \mathbf{z}_1 and \mathbf{z}_2 (Fig. 3C red and purple stars, Section 5.2.4). Simulations from these two regimes
 272 reveal different responses in each task (Fig. 3D). We characterized these regimes by identifying
 273 the dimensions of connectivity that rapid task switching is most sensitive to. The sensitivity
 274 dimensions \mathbf{v}_1 and \mathbf{v}_2 (Fig. 3C, red and purple arrows) point in different directions, resulting
 275 in different changes to task accuracy (Fig. 3D, bottom-left, 13). In regime 1, Anti accuracy
 276 diminishes in either direction of sensitivity away from the mode, while in regime 2, Anti accuracy
 277 tracks monotonic increases in Pro accuracy. These responses make intuitive sense, recognizing that

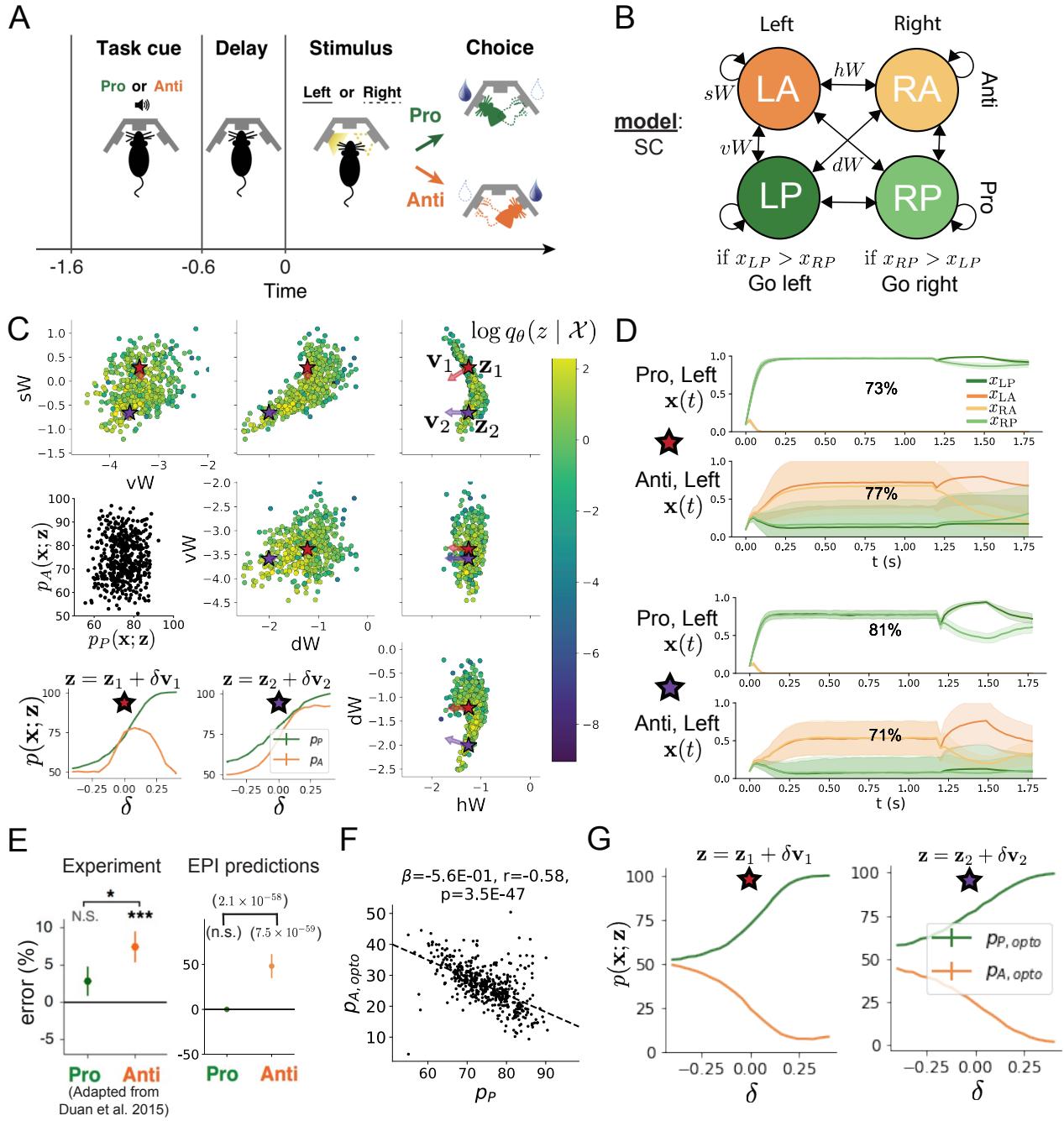


Figure 3: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. **C.** The EPI posterior distribution of rapid task switching networks. Red and purple stars (\mathbf{z}_1 and \mathbf{z}_2) indicate different connectivity regimes with different sensitivity vectors \mathbf{v}_1 and \mathbf{v}_2 . (Middle-left) Posterior predictive distribution of task accuracies. (Bottom-left) Task accuracy along dimensions of sensitivity in each connectivity regime. **D.** Means (solid) and standard deviations (shaded) of each population across random simulated trials. Top plots show Pro (top) and Anti (bottom) responses for connectivity \mathbf{z}_1 . Bottom rows show the same \mathbf{z}_2 . **E.** The EPI posterior predicts experimental results (left) showing no change in the Pro task, but larger error in the Anti task (right). **F.** Accuracy in the Anti task during delay period optogenetic inactivation $p_{A,\text{opto}}$ is strongly anticorrelated with accuracy in the Pro task. **G.** Accuracy with delay period inactivation along each connectivity regime's dimension of sensitivity.

278 \mathbf{v}_1 (unlike \mathbf{v}_2) points strongly in the direction of connectivity eigenvalue λ_{diag} , which is strongly
 279 anticorrelated with p_A (Fig. 14, 15, see Section 5.2.4).

280 In agreement with experimental results from Duan et al., we found optogenetic inactivation during
 281 the delay period consistently decreased performance in the Anti task, but had no effect on the
 282 Pro task (Fig. 3E)). This difference in resiliency across tasks to delay perturbation is a prediction
 283 made by the inferred EPI distribution, rather than an emergent property that was conditioned
 284 upon. Similarities across Pro and Anti trials in choice period responses following delay period
 285 inactivation (Fig. 17A) suggested that connectivity patterns inducing greater Pro task accuracy
 286 increase error in delay period inactivated Anti trials (Fig. 3F). The strong anticorrelation between
 287 p_P and $p_{A,\text{opto}}$ across posterior connectivities led to the following hypothesis about each connectivity
 288 regime: the sensitivity dimension of each regime decreases $p_{A,\text{opto}}$ irrespective of its effect on p_A ,
 289 since both \mathbf{v}_1 and \mathbf{v}_2 increase p_P . Indeed, in regimes 1 and 2 where sensitivity dimensions elicit
 290 different responses in p_A , $p_{A,\text{opto}}$ decreases since the connectivity changes enhancing p_P exacerbate
 291 Anti trial error (Fig. 3F).

292 In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task
 293 switching. This posterior revealed multiple regimes of rapid task switching, which we characterized
 294 using the probabilistic toolkit EPI seemlessly affords. EPI allowed us to conclude that since *all*
 295 parameters of this model producing rapid task switching make an experimentally verified prediction,
 296 the model is well-chosen in that regard. Finally, we used our knowledge about how \mathbf{z} governs $p_{A,\text{opto}}$
 297 to make accurate predictions about each identified regime of connectivity.

298 3.5 EPI scales well to high-dimensional parameter spaces

299 Here, we study the scalability of EPI in number of parameters $|\mathbf{z}|$ by inferring the connectivities
 300 of recurrent neural networks (RNNs, Fig. 4A). We consider a rank-2 RNN with N neurons having
 301 connectivity

$$W = UV^\top + g\chi \quad (8)$$

302 and dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x} \quad (9)$$

303 where $U = [\mathbf{u}_1 \ \mathbf{u}_2]$, $V = [\mathbf{v}_1 \ \mathbf{v}_2]$ and $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2 \in [-1, 1]^N$. The random component has
 304 strength $g = 0.01$ and $\chi_{i,j} \sim \mathcal{N}(0, 1)$. We infer connectivity distributions $\mathbf{z} = [\mathbf{u}_1^\top, \mathbf{u}_2^\top, \mathbf{v}_1^\top, \mathbf{v}_2^\top]^\top$
 305 producing stable amplification. RNN's exhibiting stable amplification amplify responses to input

306 along some dimensions, and are stable across all dimensions. Two conditions are both necessary
 307 and sufficient for RNNs to exhibit stable amplification [74]: $\text{real}(\lambda_1) < 1$ and $\lambda_1^s > 1$, where λ_1 is
 308 the eigenvalue of W with greatest real part and λ^s is the maximum eigenvalue of $W^s = \frac{W+W^\top}{2}$.

309 In our analysis, we seek to condition rank-2 networks of increasing size on a regime of stable ampli-
 310 fication. Networks with $\text{real}(\lambda_1) = 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$ will yield moderate amplification.
 311 EPI can naturally condition on this emergent property

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}. \end{aligned} \quad (10)$$

312 For comparison, we infer rank-2 RNN connectivities with alternative approaches to likelihood free-
 313 inference. ABC methods define a tolerance ϵ from observed data x_0 for which we keep sampled
 314 parameters. To make this ABC approach as similar as possible to the EPI program defined by
 315 Equation 10, we chose $\epsilon = 0.5$, an l_2 -distance metric, and

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \quad (11)$$

316 located at the mean of our desired emergent property. We use sequential Monte Carlo ABC (SMC-
 317 ABC), to compare efficiency, since it is considered the most efficient ABC approach. SNPE [75] is
 318 another deep likelihood-free inference method that emerged along with this work. In contrast to
 319 EPI, SNPE cannot condition on the variance of the posterior predictive distribution. Also, there
 320 is no tolerance parameter for SNPE like ϵ in ABC, so the comparative SNPE approach simply
 321 conditions on observation x_0 .

322 As we scale the number of neurons N in the RNN, and thus the dimensionality of the parameter
 323 space $\mathbf{z} \in [-1, 1]^{4N}$, we see that EPI has superior scaling properties (Fig. 4B). SMC-ABC and
 324 SNPE become intractable around 25 and 90 dimensions respectively, while EPI can infer 1000-
 325 dimensional distributions in about 1 day. No matter the number of neurons, EPI always produces
 326 connectivity distributions with mean and variance of $\text{real}(\lambda_1)$ and λ_1^s of \mathcal{X} (Fig. 4C, blue), and high
 327 variation in response profiles 4D, blue). For the dimensionalities in which SMC-ABC is tractable,
 328 the inferred parameters always exhibit stable amplification, are less varied 4C, red) and largely
 329 produce similar responses 4D, red). When using SNPE the inferred parameters are widely varied
 330 4C, orange), but often produce non-amplified or unstable responses 4D, orange). In conclusion, we

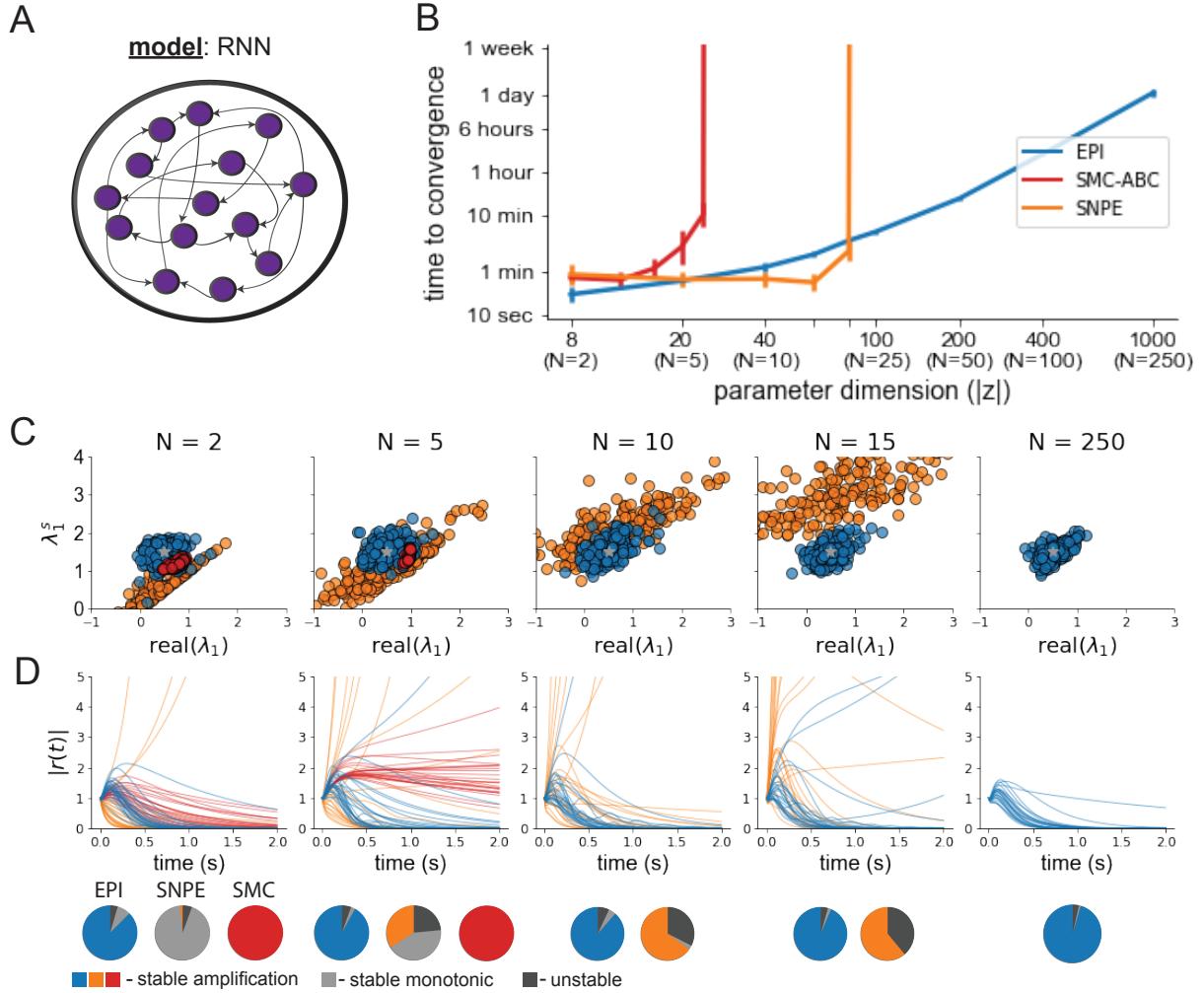


Figure 4: **A.** Recurrent neural network. **B.** EPI scales with z to high dimensions. Convergence definitions: EPI (blue) - satisfies all moment constraints, SNPE (orange)- produces at least $2/n_{\text{train}}$ parameter samples are in the bounds of emergent property (mean ± 0.5), and SMC-ABC (red) - 100 particles with $\epsilon < 0.5$ are produced. **C.** Posterior predictive distributions of EPI (blue), SNPE (orange), and SMC-ABC (red). Gray star indicates emergent property mean, and gray dashed lines indicate two standard deviations corresponding to the variance constraint. For $N \leq 6$ where SMC-ABC converges, samples are not diverse (path degeneracies). For $N \geq 25$, SNPE does not produce a posterior approximation yielding parameters with simulations near x_0 . **D.** Simulations of network parameters resulting from each method ($\tau = 100ms$). Each trace corresponds to simulation of one z . (Below) Ratio of obtained samples producing stable amplification.

331 found that deep likelihood-free inference techniques are capable of scaling to higher dimensional
332 inference than SMC-ABC. However, only EPI can scale to high dimensions while reproducing the
333 emergent property.

334 4 Discussion

335 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets
336 [30]. Such careful inference procedures are developed for these statistical models allowing precise,
337 quantitative reasoning, which clarifies the way data informs beliefs about the model parameters.
338 However, these statistical models lack resemblance to the underlying biology, making it unclear
339 how to go from the structure revealed by these methods, to the neural mechanisms giving rise
340 to it. In contrast, theoretical neuroscience has focused on careful mechanistic modeling and the
341 production of emergent properties of computation. The careful steps of *i.*) model design and
342 *ii.*) emergent property definition, are followed by *iii.*) practical inference methods resulting in an
343 opaque characterization of the way model parameters govern computation. In this work, we improve
344 upon parameter inference techniques in theoretical neuroscience with emergent property inference,
345 harnessing deep learning towards careful inference in careful models of neural computation (see
346 Section 5.1.1).

347 Specifically, approximate Bayesian computation [42, 43, 31] has been the standard approach to
348 parameter inference in neural circuit models lacking tractable likelihoods. ABC methods do not
349 confer probabilities on accepted parameters, require an acceptance threshold chosen to trade-off
350 inference quality with tractability, do not scale efficiently to high-dimensional parameter spaces, and
351 require independent techniques to analyze sensitivity for local parameter choices [76]. In contrast,
352 EPI allows probability evaluations at any point in parameter space, conditions posteriors on the
353 natural quantification of emergent properties, scales to high dimensional parameter spaces, and
354 naturally admits sensitivity quantification via fast evaluations of the posterior Hessian.

355 Technically, EPI is a maximum entropy method, which learns parameter distributions that are
356 as random as possible given that they produce the emergent property. Conceptually, maximally
357 random distributions given some constraints are useful for understanding parametric sensitivity.
358 This is well understood in Bayesian inference, where maximum entropy is the chosen normative
359 principle. This is emphasized by an innovative formalism unifying top-down maximum entropy
360 normative models with bottom-up statistical models [77]. Indeed, EPI is an adaptive variational

361 inference program, and may be considered to have a Bayesian uniform prior (see Section 5.1.6).
362 Biologically realistic models of neural circuits often prove formidable to analyze for two main rea-
363 sons. A primary challenge is that the number of parameters scales dramatically with the number of
364 neurons, limiting analysis of its parameter space. We see in Section 3.5 that EPI scales seemlessly
365 to high dimensional parameter spaces of RNN connectivities, while maintaining the production
366 of the specified emergent property. EPI strongly outperforms the standard likelihood-free infer-
367 ence technique (SMC-ABC [31]), and a recently developed deep likelihood-free inference technique
368 (SNPE [75]), most likely because of it’s ability to leverage the gradient information of the emer-
369 gent property statistics and to adapt it’s paramter sampling distribution at every step of gradient
370 descent.

371 A secondary challenge is that the structure of the parametric regimes governing emergent properties
372 is intricate. For example, even in low dimensional circuits, models can support more than one steady
373 state [78] and non-trivial dynamics on strange attractors [79]. With EPI, we use deep probabillity
374 distributions to capture the complex nonlinear parameter distributions governing model behavior.
375 In Section 3.3, we used EPI to reveal a curved parametric manifolds governing curcuit variability
376 in the stochastic stabilized supralinear network, and used hypothesis testing techniques to validate
377 our findings. In Section 3.4, we identified two regimes of SC connectivity resulting in rapid task
378 switching, and found that the full distribution of rapid task switching networks reproduced an
379 experimental result.

380 EPI leverages deep learning technology for neuroscientific inquiry in a categorically different way
381 than approaches focused on training neural networks to execute behavioral tasks [80]. These works
382 focus on examining optimized deep neural networks while considering the objective function, learn-
383 ing rule, and architecture used. This endeavor efficiently obtains sets of parameters that can be
384 reasoned about with respect to such considerations, but lacks the careful probabilistic treatment of
385 parameter inference in EPI. All of these approaches can be used complementarily to enhance the
386 practice of theoretical neuroscience.

387 **Acknowledgements:**

388 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
389 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
390 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
391 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
392 Fellowship. Helpful conversations were had with Francesca Mastrogiovanni, Srdjan Ostojic, James

393 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

394 **Data availability statement:**

395 The datasets generated during and/or analyzed during the current study are available from the
396 corresponding author upon reasonable request.

397 **Code availability statement:**

398 All software written for the current study is available at <https://github.com/cunningham-lab/epi>.

399 **References**

- 400 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
401 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 402 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
403 21(1):25–45, 1998.
- 404 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 405 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in
406 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 407 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
408 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
409 models. *PLoS Comput Biol*, 3(10):e189, 2007.
- 410 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
411 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
412 sion model. *Neuron*, 82(4):809–821, 2014.
- 413 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-
414 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 415 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
416 networks. *Physical review letters*, 61(3):259, 1988.
- 417 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
418 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
419 17(11):4382–4388, 1997.

- 420 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration
421 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 422 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
423 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 424 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
425 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 426 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
427 model. *PLoS computational biology*, 2(7):e94, 2006.
- 428 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
429 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 430 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
431 13(8):1713–1720, 2001.
- 432 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
433 A statistical paradigm for neural spike train decoding applied to position prediction from
434 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
435 7425, 1998.
- 436 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
437 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 438 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.
439 A point process framework for relating neural spiking activity to spiking history, neural
440 ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 441 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations
442 imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–
443 1012, 2006.
- 444 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan
445 Segev. A novel multiple objective optimization framework for constraining conductance-based
446 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 447 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature
448 analysis. *Neural computation*, 19(4):1022–1038, 2007.

- 449 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
450 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of
451 neural population activity. In *Advances in neural information processing systems*, pages
452 1881–1888, 2009.
- 453 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and
454 Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural*
455 *information processing systems*, 24:1350–1358, 2011.
- 456 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In
457 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 458 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-
459 dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,
460 9(3):e1002922, 2013.
- 461 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
462 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
463 *Science*, 349(6244):184–187, 2015.
- 464 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq
465 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in
466 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 467 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
468 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th In-*
469 *ternational Conference on Machine Learning*, 2019.
- 470 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.
471 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.
472 *Nature Communications*, 10(4933), 2019.
- 473 [30] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
474 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
475 2018.
- 476 [31] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
477 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

- 478 [32] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
479 A framework for parameter estimation and model selection from experimental data in systems
480 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 481 [33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Con-*
482 *ference on Learning Representations*, 2014.
- 483 [34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropaga-
484 tion and variational inference in deep latent gaussian models. *International Conference on*
485 *Machine Learning*, 2014.
- 486 [35] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
487 neural population models through nonlinear embeddings. In *Advances in neural information*
488 *processing systems*, pages 163–171, 2016.
- 489 [36] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-
490 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 491 [37] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-
492 encoders. *bioRxiv*, page 399246, 2018.
- 493 [38] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
494 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R
495 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-
496 encoders. *Nature methods*, page 1, 2018.
- 497 [39] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
498 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
499 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 500 [40] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
501 Datta. Composing graphical models with neural networks for structured representations and
502 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 503 [41] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
504 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
505 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural*
506 *Information Processing Systems*, 2019.

- 507 [42] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-
508 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 509 [43] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
510 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
511 2003.
- 512 [44] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
513 1970.
- 514 [45] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
515 Edward Teller. Equation of state calculations by fast computing machines. *The journal of
516 chemical physics*, 21(6):1087–1092, 1953.
- 517 [46] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
518 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 519 [47] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
520 *International Conference on Machine Learning*, 2015.
- 521 [48] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R
522 Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics,
523 biology, and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.
- 524 [49] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
525 free variational inference. In *Advances in Neural Information Processing Systems*, pages
526 5523–5533, 2017.
- 527 [50] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
528 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 529 [51] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for
530 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,
531 2017.
- 532 [52] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
533 networks. *International Conference on Learning Representations*, 2017.

- 534 [53] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-
535 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,
536 2001.
- 537 [54] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and
538 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysi-
539 ology*, 115(3):1399–1409, 2016.
- 540 [55] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
541 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
542 routing. *bioRxiv*, page 245613, 2018.
- 543 [56] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
544 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 545 [57] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
546 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
547 *Neuron*, 77(5):845–858, 2013.
- 548 [58] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
549 1957.
- 550 [59] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
551 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 552 [60] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
553 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 554 [61] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
555 2009.
- 556 [62] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
557 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 558 [63] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
559 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
560 592, 2009.

- 561 [64] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
562 network: a unifying circuit motif underlying multi-input integration in sensory cortex.
563 *Neuron*, 85(2):402–417, 2015.
- 564 [65] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
565 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
566 5(10):793, 2004.
- 567 [66] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
568 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-*
569 *robiology*, 71(1):45–61, 2011.
- 570 [67] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-
571 tex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 572 [68] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
573 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
574 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 575 [69] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing
576 Wang. Paradoxical response reversal of top- down modulation in cortical circuits with three
577 interneuron types. *Elife*, 6:1–15, 2017.
- 578 [70] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying
579 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 580 [71] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
581 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
582 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 583 [72] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
584 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
585 operating regime of cortex. *bioRxiv*, 2020.
- 586 [73] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and
587 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,
588 2015.

- 589 [74] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural
590 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 591 [75] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
592 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P
593 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
594 dynamics. *bioRxiv*, page 838383, 2019.
- 595 [76] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian com-*
596 *putation*. CRC Press, 2018.
- 597 [77] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical
598 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.
- 599 [78] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give
600 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of*
601 *Sciences*, 115(13):3464–3469, 2018.
- 602 [79] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emergent
603 dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint arXiv:1605.04463*, 2016.
- 604 [80] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,
605 2019.
- 606 [81] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
607 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
608 73(2):123–214, 2011.
- 609 [82] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic biochemical
610 network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
611 820, 2011.
- 612 [83] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference.
613 *Proceedings of the National Academy of Sciences*, 2020.
- 614 [84] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degenerate
615 solution networks for theoretical neuroscience. *Computational and Systems Neuroscience
616 Meeting (COSYNE), Lisbon, Portugal*, 2019.

- 618 [85] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
619 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience
620 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 621 [86] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
622 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
623 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 624 [87] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-
625 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,*
626 *Germany*, 2019.
- 627 [88] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
628 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
629 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 630 [89] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
631 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on*
632 *Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 633 [90] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
634 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
635 4248. PMLR, 2020.
- 636 [91] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
637 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 638 [92] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
639 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 640 [93] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
641 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
642 2012.
- 643 [94] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
644 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
645 2018.

- 646 [95] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
647 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 648 [96] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
649 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
650 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
651 1929, 2009.
- 652 [97] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
653 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 654 [98] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
655 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 656 [99] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-
657 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv
658 preprint arXiv:1912.02762*, 2019.
- 659 [100] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
660 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 661 [101] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
662 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-
663 tion processing systems*, 29:4743–4751, 2016.
- 664 [102] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Inter-
665 national Conference on Learning Representations*, 2015.
- 666 [103] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for
667 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

668 **5 Methods**

669 **5.1 Emergent property inference (EPI)**

670 Determining the combinations of model parameters that can produce observed data or a desired
671 output is a key part of scientific practice. Solving inverse problems is especially important in
672 neuroscience, since we require complex models to describe the complex phenomena of neural com-
673 putations. While much machine learning research has focused on how to find latent structure
674 in large-scale neural datasets, less has focused on inverting theoretical circuit models conditioned
675 upon the emergent phenomena they produce. Here, we introduce a novel method for statistical
676 inference, which finds distributions of parameter solutions that only produce the desired emer-
677 gent property. This method seamlessly handles neural circuit models with stochastic nonlinear
678 dynamical generative processes, which are predominant in theoretical neuroscience.

679 Consider model parameterization \mathbf{z} , which is a collection of scientifically interesting variables that
680 govern the complex simulation of data \mathbf{x} . For example (see Section 3.1), \mathbf{z} may be the electrical
681 conductance parameters of an STG subcircuit, and \mathbf{x} the evolving membrane potentials of the five
682 neurons. In terms of statistical modeling, this circuit model has an intractable likelihood $p(\mathbf{x} | \mathbf{z})$,
683 which is predicated by the stochastic differential equations that define the model. Even so, we do
684 not scientifically reason about how \mathbf{z} governs all of \mathbf{x} , but rather specific phenomena that are a
685 function of the data $f(\mathbf{x}; \mathbf{z})$. In the STG example, $f(\mathbf{x}; \mathbf{z})$ measures hub neuron frequency from the
686 evolution of \mathbf{x} governed by \mathbf{z} . With EPI, we learn distributions of \mathbf{z} that results in an average and
687 variance of $f(\mathbf{x}; \mathbf{z})$, denoted $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. We refer to the collection of these statistical moments as an
688 emergent property. Such emergent properties \mathcal{X} are defined through choice of $f(\mathbf{x}; \mathbf{z})$ (which may
689 be one or multiple statistics), $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}^2$

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (12)$$

690 Precisely, the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ must have means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ over the
691 EPI distribution of parameters and stochasticity of the data given the parameters.

692 In EPI, deep probability distributions are used as posterior approximations $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X})$. In deep
693 probability distributions, a simple random variable $\mathbf{z}_0 \sim q_0(\mathbf{z}_0)$ is mapped deterministically via a
694 sequence of deep neural network layers (g_1, \dots, g_l) parameterized by weights and biases $\boldsymbol{\theta}$ to the
695 support of the distribution of interest:

$$\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0) = g_l(\dots g_1(\mathbf{z}_0)) \sim q_{\boldsymbol{\theta}}(\mathbf{z}). \quad (13)$$

696 Such deep probability distributions embed the posterior distribution in a deep network. Once
697 optimized, this deep network representation has remarkably useful properties: immediate posterior
698 sampling, and immediate probability, gradient, and Hessian evaluation at any parameter choice.

699 Given a choice of model $p(\mathbf{x} \mid \mathbf{z})$ and emergent property of interest \mathcal{X} , $q_{\theta}(\mathbf{z})$ is optimized via
700 the neural network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep
701 variational family \mathcal{Q} producing the emergent property \mathcal{X} :

$$q_{\theta}(\mathbf{z} \mid \mathcal{X}) = q_{\theta}^*(\mathbf{z}) = \operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}(\mathbf{z})) \quad (14)$$

s.t. $\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2$.

702 Entropy is chosen as the normative selection principle, since we want the posterior to only contain
703 structure predicated by the emergent property [58, 59]. This choice of selection principle is also
704 that of standard Bayesian inference, and we derive an exact relation between EPI and variational
705 inference (see Section 5.1.5). However, a key difference is that variational inference and other
706 Bayesian methods do not constrain the predictions of their inferred posteriors. This optimization
707 is executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [52].

708 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.
709 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-
710 ond, we describe the special class of deep probability distributions used in EPI called normalizing
711 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-
712 tion 14 (Section 5.1.3). Then, we demonstrate the details of this optimization in a toy example
713 (Section 5.1.4). Finally, we establish the known relationship between maximum entropy distribu-
714 tions and exponential families (Section 5.1.5), which is used to explain the relation between EPI
715 and variational inference (Section 5.1.6).

716 5.1.1 Related approaches

717 When Bayesian inference problems lack conjugacy, scientists use approximate inference methods
718 like variational inference (VI) [46] and Markov chain Monte Carlo (MCMC) [45, 44]. After opti-
719 mization, variational methods return a parameterized posterior distribution, which we can analyze.
720 Also, the variational approximating distribution class is often chosen such that it permits fast
721 sampling. In contrast MCMC methods only produce samples from the approximated posterior dis-
722 tribution. No parameterized distribution is estimated, and additional samples are always generated
723 with the same sampling complexity. Inference in models defined by systems of differential has been

724 demonstrated with MCMC [81], although this approach requires tractable likelihoods. Advances
725 have leveraged structure in stochastic differential equation models to improve likelihood
726 approximations, thus expanding the domain of applicable models [82].

727 Likelihood-free (or “simulation-based”) inference (LFI) [83] is model parameter inference in the
728 absence of a tractable likelihood function. The most prevalent approach to LFI is approximate
729 Bayesian computation [42], in which satisfactory parameter samples are kept from random prior
730 sampling according to a rejection heuristic. The obtained set of parameters do not have a prob-
731 abilities, and further insight about the model must be gained from examination of the parameter
732 set and their generated activity. Methodological advances to ABC methods have come through
733 the use of Markov chain Monte Carlo (MCMC-ABC) [43] and sequential Monte Carlo (SMC-ABC)
734 [31] sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still
735 struggles to scale in dimensionality (cf. Fig. 4). Furthermore, once a parameter set has been
736 obtained by SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again
737 with a new population of initialized particles to obtain additional samples.

738 For scientific model analysis, we seek a posterior distribution exhibiting the properties of a well-
739 chosen variational approximation: a parametric form conferring analytic calculations, and trivial
740 sampling time. For this reason, ABC and MCMC techniques are unattractive, since they only
741 produce a set of parameter samples and have unchanging sampling rate. EPI executes likelihood-
742 free inference using the MEFN [52] algorithm using a deep variational posterior approximation.
743 The deep neural network of EPI defines the parametric form of the posterior approximation. Fur-
744 thermore, the EPI distribution is constrained to produce an emergent property. In other words,
745 the summary statistics of the posterior predictive distribution are fixed to have certain first and
746 second moments. EPI optimization is enabled using stochastic gradient techniques in the spirit
747 of likelihood-free variational inference [49]. The analytic relationship between EPI and variational
748 inference is explained in Secton 5.1.6.

749 We note that, during our preparation and early presentation of this work [84, 85], another work
750 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
751 circuits ([86, 87, 75]). We are encouraged by this general problem being recognized by others in the
752 community, and we emphasize that these works offer complementary neuroscientific contributions
753 (different theoretical models of focus) and use different technical methodologies (ours is built on
754 our prior work [52], theirs similarly [88]).

755 The method EPI differs from SNPE in some key ways. SNPE belongs to a “sequential” class of

756 recently developed LFI methods in which two neural networks are used for posterior inference.
757 This first neural network is a normalizing flow used to estimate the posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE)
758 or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural likelihood (SNL [89])). A recent advance uses an
759 unconstrained neural network to estimate the likelihood ratio (sequential neural ratio estimation
760 (SNRE [90])). In SNL and SNRE, MCMC sampling techniques are used to obtain samples from
761 the approximated posterior. This contrasts with EPI and SNPE, which afford a normalizing flow
762 approximation to the posterior, which facilitates immediate measurements of sample probability,
763 gradient, or Hessian for system analysis. The second neural network in this sequential class of
764 methods is the amortizer. This network maps data \mathbf{x} (or statistics $f(\mathbf{x}; \mathbf{z})$ or model parameters \mathbf{z})
765 to the weights and biases of the first neural network. These methods are optimized on a conditional
766 density (or ratio) estimation objective on a sequentially adapting finite sample-based approximation
767 to the posterior.

768 The approximating fidelity of the first neural network in sequential approaches is optimized to
769 generalize across the entire distribution it is conditioned upon. This optimization towards gen-
770 eralization of sequential methods can reduce the accuracy at the singular posterior of interest.
771 Whereas in EPI, the entire expressivity of the normalizing flow is dedicated to learning a single
772 distribution as well as possible. While amortization is not possible in EPI parameterized by the
773 mean parameter μ (due to the inverse mapping problem [91]), we have shown this two-network
774 amortization approach to be effective in exponential family distributions defined by their natural
775 parameterization [92].

776 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in
777 natural models. Around a point, one can measure the Jacobian. One approach that scales well is
778 EAR [93]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [94] and
779 its stochastic adaptation [95]. Casting identifiability as a statistical estimation problem, the profile
780 likelihood can assess via iterated optimization while holding parameters fixed [96]. An exciting
781 recent method is capable of recovering the functional form of such unidentifiabilities away from a
782 point by following degenerate dimensions of the fisher information matrix [97]. Global structural
783 non-identifiabilities can be found for models with polynomial or rational dynamics equations using
784 DAISY [98]. With EPI, we have all the benefits given by a statistical inference method plus the
785 ability to query the gradient or Hessian of the inferred distribution at any chosen parameter value.

786 **5.1.2 Normalizing flows**

787 Deep probability distributions are comprised of multiple layers of fully connected neural networks
 788 (Equation). When each neural network layer is restricted to be a bijective function, the sample
 789 density can be calculated using the change of variables formula at each layer of the network. For
 790 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

$$p(\mathbf{z}_i) = p(g_i^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial g_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} \right| = p(\mathbf{z}_{i-1}) \left| \det \frac{\partial g_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right|^{-1}. \quad (15)$$

791 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 792 restricting our layers to normalizing flows [47, 99] – bijective functions with fast log determinant
 793 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log
 794 probability calculation confers efficient optimization of the maximum entropy objective (see Section
 795 5.1.3). We use the Real NVP [50] normalizing flow class, because its coupling architecture confers
 796 both fast sampling (forward) and fast log probability evaluation (backward). Fast probability
 797 evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout
 798 parameter space. Glow permutations were used in between coupling stages [100]. This is in contrast
 799 to autoregressive architectures [51, 101], in which only forward or backward passes are efficient. In
 800 this work, normalizing flows are used as flexible posterior approximations $q_{\boldsymbol{\theta}}(\mathbf{z})$ having weights and
 801 biases $\boldsymbol{\theta}$. We specify the architecture used in each application by the number of Real-NVP affine
 802 coupling stages, and the number of neural network layers and units per layer of the conditioning
 803 functions.

804 **5.1.3 Augmented Lagrangian optimization**

805 To optimize $q_{\boldsymbol{\theta}}(\mathbf{z})$ in Equation 14, the constrained maximum entropy optimization is executed using
 806 the augmented Lagrangian method. The following objective is minimized:

$$L(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{opt}}, c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top R(\boldsymbol{\theta}) + \frac{c}{2} \|R(\boldsymbol{\theta})\|^2 \quad (16)$$

807 where average constraint violations $R(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [T(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_{\text{opt}}]]$, $\boldsymbol{\eta}_{\text{opt}} \in \mathbb{R}^m$ are the
 808 Lagrange multipliers where $m = |\boldsymbol{\mu}_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})| = 2|f(\mathbf{x}; \mathbf{z})|$, and c is the penalty coefficient.
 809 The sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter $\boldsymbol{\mu}_{\text{opt}}$ are determined by the means $\boldsymbol{\mu}$ and
 810 variances $\boldsymbol{\sigma}^2$ of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 14. Specifically, $T(\mathbf{x}; \mathbf{z})$ is
 811 a concatenation of the first and second moments, $\boldsymbol{\mu}_{\text{opt}}$ is a concatenation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ (see section
 812 5.1.5), and the Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of exponential

813 families (see Section 5.1.6). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution are optimized
814 according to Equation 16 using the Adam optimizer with learning rate 10^{-3} [102].

815 To take gradients with respect to the entropy $H(q_{\boldsymbol{\theta}}(\mathbf{z}))$, it can be expressed using the reparam-
816 eterization trick as an expectation of the negative log density of parameter samples \mathbf{z} over the
817 randomness in the parameterless initial distribution $q_0(\mathbf{z}_0)$:

$$H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \int -q_{\boldsymbol{\theta}}(\mathbf{z}) \log(q_{\boldsymbol{\theta}}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [-\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (17)$$

818 Thus, the gradient of the entropy of the deep probability distribution can be estimated as an
819 average with respect to the base distribution \mathbf{z}_0 :

$$\nabla_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\nabla_{\boldsymbol{\theta}} \log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (18)$$

820 The lagrangian parameters $\boldsymbol{\eta}_{\text{opt}}$ are initialized to zero and adapted following each augmented
821 Lagrangian epoch, which is a period of optimization with fixed $(\boldsymbol{\eta}_{\text{opt}}, c)$ for a given number of
822 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
823 after each epoch based on constraint error reduction. The penalty coefficient is updated based
824 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
825 $\mathbb{E}[|R(\boldsymbol{\theta}_{k+1})|] > \gamma \mathbb{E}[|R(\boldsymbol{\theta}_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
826 other update rule is $\boldsymbol{\eta}_{\text{opt},k+1} = \boldsymbol{\eta}_{\text{opt},k} + c_k \frac{1}{n} \sum_{i=1}^n (T(\mathbf{x}^{(i)}) - \boldsymbol{\mu}_{\text{opt}})$ given a batch size n . Throughout
827 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
828 according to application.

829 The intention is that c and $\boldsymbol{\eta}_{\text{opt}}$ start at values encouraging entropic growth early in optimization.
830 With each training epoch in which the update rule for c is invoked by unsatisfactory constraint
831 error reduction, the constraint satisfaction terms are increasingly weighted, resulting in a decreased
832 entropy. This encourages the discovery of suitable regions of parameter space, and the subsequent
833 refinement of the distribution to produce the emergent property (see example in Section 5.1.4). The
834 momentum parameters of the Adam optimizer are reset at the end of each augmented Lagrangian
835 epoch.

836 Rather than starting optimization from some $\boldsymbol{\theta}$ drawn from a randomized distribution, we found
837 that initializing $q_{\boldsymbol{\theta}}(\mathbf{z})$ to approximate an isotropic Gaussian distribution conferred more stable, con-
838 sistent optimization. The parameters of the Gaussian initialization were chosen on an application-
839 specific basis. Throughout the study, we chose isotropic Gaussian initializations with mean $\boldsymbol{\mu}_{\text{init}}$
840 at the center of the distribution support and some standard deviation σ_{init} , except for one case,
841 where an initialization informed by random search was used (see Section 5.2.1).

842 To assess whether the EPI distribution $q_{\theta}(\mathbf{z})$ produces the emergent property, we assess whether
 843 each individual constraint on the means and variances of $f(\mathbf{x}; \mathbf{z})$ is satisfied. We consider the EPI
 844 to have converged when a null hypothesis test of constraint violations $R(\boldsymbol{\theta})_i$ being zero is accepted
 845 for all constraints $i \in \{1, \dots, m\}$ at a significance threshold $\alpha = 0.05$. This significance threshold is
 846 adjusted through Bonferroni correction according to the number of constraints m . The p-values for
 847 each constraint are calculated according to a two-tailed nonparametric test, where 200 estimations
 848 of the sample mean $R(\boldsymbol{\theta})^i$ are made using N_{test} samples of $\mathbf{z} \sim q_{\theta}(\mathbf{z})$ at the end of the augmented
 849 Lagrangian epoch.

850 When assessing the suitability of EPI for a particular modeling question, there are some important
 851 technical considerations. First and foremost, as in any optimization problem, the defined emergent
 852 property should always be appropriately conditioned (constraints should not have wildly different
 853 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution
 854 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-
 855 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,
 856 one should consider the computational cost of the gradient calculations. In the best circumstance,
 857 there is a simple, closed form expression (e.g. Section 5.2.5) for the emergent property statistic
 858 given the model parameters. On the other end of the spectrum, many forward simulation iterations
 859 may be required before a high quality measurement of the emergent property statistic is available
 860 (e.g. Section 5.2.1). In such cases, backpropagating gradients through the SDE evolution will be
 861 expensive.

862 5.1.4 Example: 2D LDS

863 To gain intuition for EPI, consider a two-dimensional linear dynamical system (2D LDS) model
 864 (Fig. S1A):

$$865 \quad \tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (19)$$

865 with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (20)$$

866 To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fix-
 867 ing $\tau = 1$), the emergent property statistics $T(\mathbf{x})$ were chosen to contain the first and second
 868 moments of the oscillatory frequency, $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscil-
 869 lating system. λ_1 is the eigenvalue of greatest real part when the imaginary component is zero, and

alternatively of positive imaginary component when the eigenvalues are complex conjugate pairs.
 To learn the distribution of real entries of A that produce a band of oscillating systems around
 1Hz, we formalized this emergent property as $\text{real}(\lambda_1)$ having mean zero with variance 0.25^2 , and
 the oscillation frequency $2\pi\text{imag}(\lambda_1)$ having mean $\omega = 1$ Hz with variance $(0.1\text{Hz})^2$:
 870
 871
 872
 873

$$\mathbb{E}[T(\mathbf{x})] \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}. \quad (21)$$

874

Unlike the models we presented in the main text, this model admits an analytical form for the
 mean emergent property statistics given parameter \mathbf{z} , since the eigenvalues can be calculated using
 the quadratic formula:
 875
 876
 877

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (22)$$

Importantly, even though $\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})}[T(\mathbf{x})]$ is calculable directly via a closed form function and
 does not require simulation, we cannot derive the distribution q_{θ}^* directly. This fact is due to the
 formally hard problem of the backward mapping: finding the natural parameters η from the mean
 parameters $\boldsymbol{\mu}$ of an exponential family distribution [91]. Instead, we used EPI to approximate this
 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two
 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped
 onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

Even this relatively simple system has nontrivial (though intuitively sensible) structure in the
 parameter distribution. To validate our method, we analytically derived the contours of the prob-
 ability density from the emergent property statistics and values. In the a_1 - a_4 plane, the black
 line at $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$,
 and the dotted gray 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. S2A). 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 bimodality of the EPI distribution, we examined the imaginary component of λ_1 .

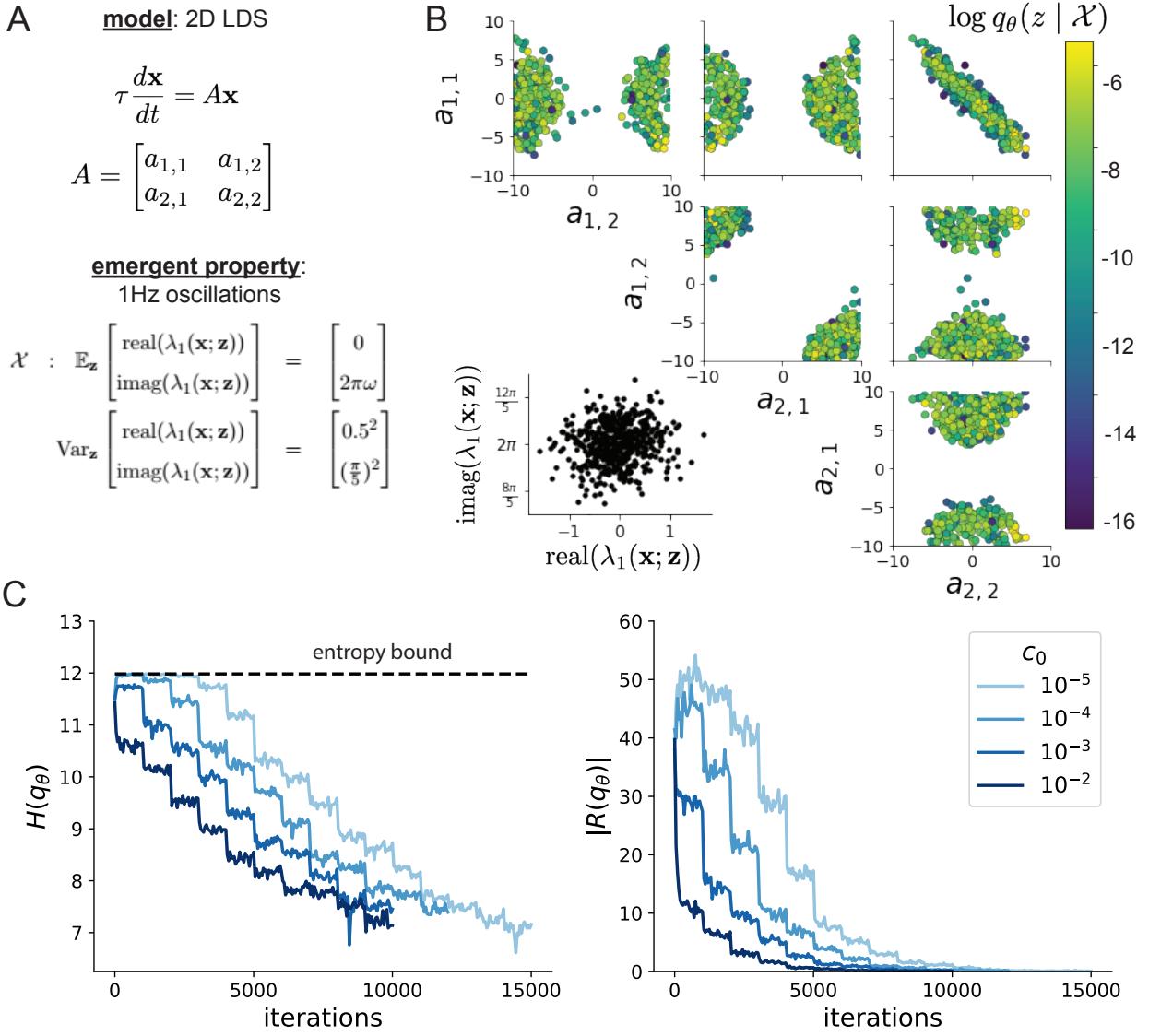


Figure 5: (LDS1): A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed lines indicate the parameter axes. C. Entropy throughout the optimization. At the beginning of each augmented Lagrangian epoch (2,000 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 adjust closer to their constraints.

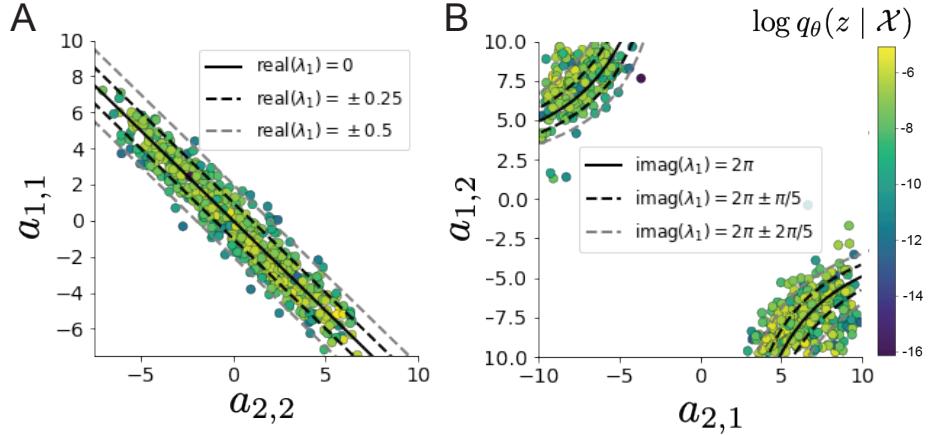


Figure 6: (LDS2): A. Probability contours in the a_1 - a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. B. Probability contours in the a_2 - a_3 plane were derived from the emergent property statistic of oscillation frequency $2\pi\text{imag}(\lambda_1)$.

894 When $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, we have

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4 - a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (23)$$

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

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \quad (24)$$

897 Since we constrained $\mathbb{E}_{\mathbf{z} \sim q_\theta} [\text{imag}(\lambda)] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation
898 $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_\theta} [|a_1a_4 - E_{q_\theta}[a_1a_4]|]$,
899 we show the contours as $a_1a_4 = 0$ (black), $a_1a_4 = -\sigma_{1,4}$ (black dotted), and $a_1a_4 = -2\sigma_{1,4}$ (grey
900 dotted). This validates the curved structure of the inferred distribution learned through EPI. We
901 took steps in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive
902 values a_1a_4 in the learned distribution. Subtler combinations of model and emergent property will
903 have more complexity, further motivating the use of EPI for understanding these systems. As we
904 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz
905 (Fig. S3).

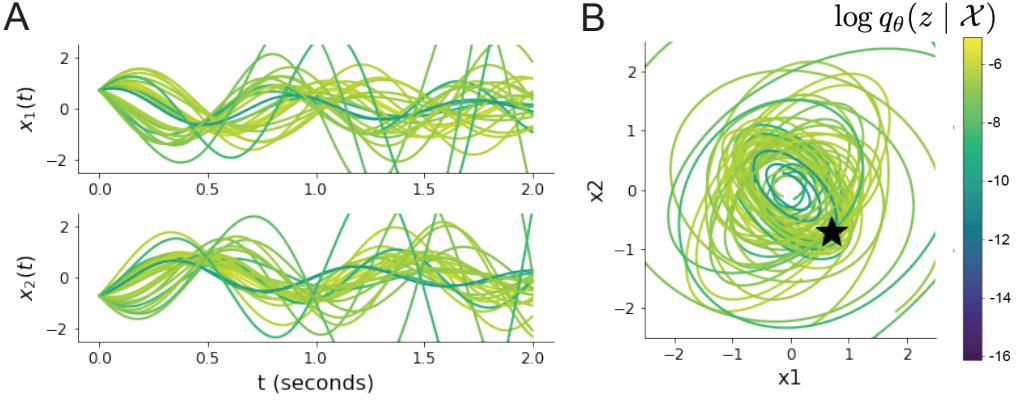


Figure 7: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\theta}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. A. Each dimension of the simulated trajectories throughout time. B The simulated trajectories in phase space.

906 5.1.5 Maximum entropy distributions and exponential families

907 Maximum entropy distributions have a fundamental link to exponential family distributions. A
 908 maximum entropy distribution of form:

$$p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \quad (25)$$

s.t. $\mathbb{E}_{\mathbf{z} \sim p}[T(\mathbf{z})] = \boldsymbol{\mu}_{\text{opt}}$.

909 will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \quad (26)$$

910 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
 911 formally hard to identify [91].

912 In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equation
 913 2

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (27)$$

914 The variance constraint is a second moment constraint on $f(\mathbf{x}; \mathbf{z})$

$$\operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \mathbb{E}_{\mathbf{z}, \mathbf{x}}[(f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2] \quad (28)$$

915 As a general maximum entropy distribution (Equation 25), the sufficient statistics vector contains

916 both first and second order moments of $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (29)$$

917 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (30)$$

918 5.1.6 EPI as variational inference

919 In Bayesian inference a prior belief about model parameters \mathbf{z} is stated in a prior distribution $p(\mathbf{z})$,
920 and the statistical model capturing the effect of \mathbf{z} on observed data points \mathbf{x} is formalized in the
921 likelihood distribution $p(\mathbf{x} | \mathbf{z})$. In Bayesian inference, we obtain a posterior distribution $p(z | \mathbf{x})$,
922 which captures how the data inform our knowledge of model parameters using Bayes' rule:

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \quad (31)$$

923 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
924 However, conjugacy is rare in practice, and alternative methods, such as variational inference [103],
925 are utilized.

926 In variational inference, a posterior approximation $q_{\boldsymbol{\theta}}^*$ is chosen from within some variational family
927 \mathcal{Q}

$$q_{\boldsymbol{\theta}}^*(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})). \quad (32)$$

928 The KL divergence can be written in terms of entropy of the variational approximation:

$$KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})) = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] - \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{z} | \mathbf{x}))] \quad (33)$$

$$= -H(q_{\boldsymbol{\theta}}) - \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z})) - \log(p(\mathbf{x}))] \quad (34)$$

930 Since the marginal distribution of the data $p(\mathbf{x})$ (or ‘evidence’) is independent of $\boldsymbol{\theta}$, variational
931 inference is executed by optimizing the remaining expression. This is usually framed as maximizing
932 the evidence lower bound (ELBO)

$$\underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}} || p(\mathbf{z} | \mathbf{x})) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmax}} H(q_{\boldsymbol{\theta}}) + \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z}))]. \quad (35)$$

933 Now, consider the setting where we have chosen a uniform prior, and stipulate a mean-field gaussian
934 likelihood on a chosen statistic of the data $f(\mathbf{x}; \mathbf{z})$

$$p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) | \boldsymbol{\mu}_f, \Sigma_f), \quad (36)$$

935 where $\Sigma_f = \text{diag}(\boldsymbol{\sigma}_f^2)$. The log likelihood is then proportional to a dot product of the natural
 936 parameter of this mean-field gaussian distribution and the first and second moment statistics.

$$\log p(\mathbf{x} | \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \quad (37)$$

937 where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\boldsymbol{\mu}_f}{\sigma_f^2} \\ \frac{-1}{2\sigma_f^2} \end{bmatrix}, \text{ and} \quad (38)$$

$$938 \quad T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \quad (39)$$

939 The variational objective is then

$$\underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta) + \boldsymbol{\eta}_f^\top \mathbb{E}_{\mathbf{z} \sim q_\theta} [T(\mathbf{x}; \mathbf{z})] \quad (40)$$

940 Comparing this to the Lagrangian objective (without augmentation) of EPI, we see they are the
 941 same

$$\begin{aligned} q_\theta^*(\mathbf{z}) &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top (\mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}_{\text{opt}}) \\ &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})]. \end{aligned} \quad (41)$$

942 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
 943 (Equation 29). Thus, EPI is implicitly executing variational inference with a uniform prior and a
 944 mean-field gaussian likelihood on the emergent property statistics. The data \mathbf{x} used by this implicit
 945 variational inference program would be that generated by the adapting variational approximation
 946 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{z})q_\theta(\mathbf{z})$, and the likelihood parameters $\boldsymbol{\eta}_f$ of EPI optimization epoch k are predicated
 947 by $\boldsymbol{\eta}_{\text{opt},k}$. However, in EPI we have not specified a prior distribution, or collected data, which can
 948 inform us about model parameters. Instead we have a mathematical specification of an emergent
 949 property, which the model must produce, and a maximum entropy selection principle. Accordingly,
 950 we replace the notation of $p(\mathbf{z} | \mathbf{x})$ with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys
 951 emergent property \mathcal{X} (see Section 5.1).

952 5.2 Theoretical models

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

955 **5.2.1 Stomatogastric ganglion**

956 We analyze how the parameters $\mathbf{z} = [g_{el}, g_{synA}]$ govern the emergent phenomena of intermediate
 957 hub frequency in a model of the stomatogastric ganglion (STG) [57] shown in Figure 1A with
 958 activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$, using the same hyperparameter choices as Gutierrez et al.
 959 Each neuron's membrane potential $x_\alpha(t)$ for $\alpha \in \{f1, f2, hub, s1, s2\}$ is the solution of the following
 960 stochastic differential equation:

$$C_m \frac{dx_\alpha}{dt} = -[h_{leak}(\mathbf{x}; \mathbf{z}) + h_{Ca}(\mathbf{x}; \mathbf{z}) + h_K(\mathbf{x}; \mathbf{z}) + h_{hyp}(\mathbf{x}; \mathbf{z}) + h_{elec}(\mathbf{x}; \mathbf{z}) + h_{syn}(\mathbf{x}; \mathbf{z})] + dB. \quad (42)$$

961 The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization,
 962 electrical and synaptic currents as well as gaussian noise dB . Each current component is a function
 963 of all membrane potentials and the conductance parameters \mathbf{z} .

964 The capacitance of the cell membrane was set to $C_m = 1nF$. Specifically, the currents are the
 965 difference in the neuron's membrane potential and that current type's reversal potential multiplied
 966 by a conductance:

$$h_{leak}(\mathbf{x}; \mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (43)$$

$$h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{post} - x_\alpha^{pre}) \quad (44)$$

$$h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (45)$$

$$h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (46)$$

$$h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (47)$$

$$h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (48)$$

972 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 973 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 974 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 975 conductances had values $g_{Ca} = 1.9 \times 10^{-2}$, $g_K = 3.9 \times 10^{-2}$, and $g_{hyp} = 2.5 \times 10^{-2}$. The intermediate
 976 conductances had values $g_{Ca} = 1.7 \times 10^{-2}$, $g_K = 1.9 \times 10^{-2}$, and $g_{hyp} = 8.0 \times 10^{-3}$. Finally, the
 977 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

978 Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating
 979 dynamics dependent on steady-state gating variables M_∞ , N_∞ and H_∞ , respectively:

$$M_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_1}{v_2} \right) \right) \quad (49)$$

$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (50)$$

$$N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_3}{v_4} \right) \right) \quad (51)$$

$$\lambda_N = \phi_N \cosh \left(\frac{x_\alpha - v_3}{2v_4} \right) \quad (52)$$

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (53)$$

$$H_\infty = \frac{1}{1 + \exp\left(\frac{x_\alpha + v_5}{v_6}\right)} \quad (54)$$

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp\left(\frac{-x_\alpha + v_7}{v_8}\right)} \right). \quad (55)$$

where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

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

$$S_\infty = \frac{1}{1 + \exp\left(\frac{v_{th} - x_\alpha}{v_9}\right)}. \quad (56)$$

When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear dynamical system. Gaussian noise $d\mathbf{B}$ of variance $(1 \times 10^{-12})^2$ A² makes the model stochastic, and introduces variability in frequency at each parameterization \mathbf{z} .

In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for $T = 300$ time steps of $dt = 25\text{ms}$. The chosen dt and T were the most computationally convenient choices yielding accurate frequency measurement. We used a basis of complex exponentials with frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

$$\Phi = [0.0, 0.01, \dots, 1.0]^\top .. \quad (57)$$

996 To measure spiking frequency, we processed simulated membrane potentials with a relu (spike
997 extraction) and low-pass filter with averaging window of size 20, then took the frequency with the
998 maximum absolute value of the complex exponential basis coefficients of the processed time-series.
999 The first 20 temporal samples of the simulation are ignored to account for initial transients.

To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in \mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in$

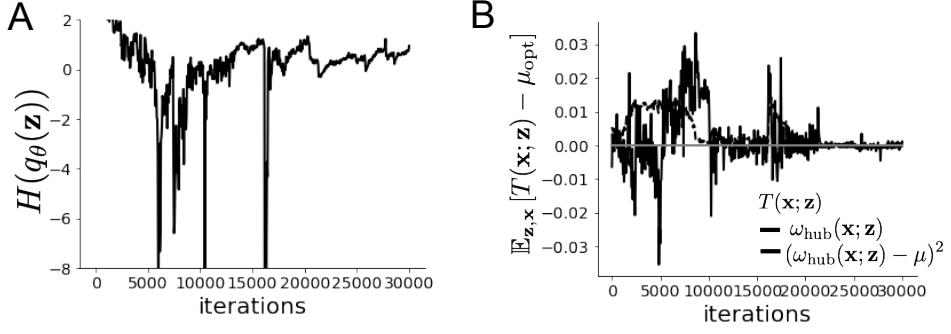


Figure 8: (STG1): EPI optimization of the STG model producing network syncing. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

1002 $\{f_1, f_2, \text{hub}, s_1, s_2\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta |X_\alpha| \odot i), \quad (58)$$

1003 where $i = [0, 1, \dots, 100]$. The frequency is then calculated as

$$\omega_\alpha = 0.01\psi_\alpha \text{Hz}. \quad (59)$$

1004 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
 1005 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
 1006 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be $(0.025\text{Hz})^2$ to
 1007 capture variation in frequency between 0.5Hz and 0.6Hz (Equation 2). As a maximum entropy dis-
 1008 tribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments of the hub neuron frequency
 1009 (as in Equations 29 and 30)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \quad (60)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \quad (61)$$

1010 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 1011 epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

1012 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-
 1013 layer neural networks of 25 units per layer. The normalizing flow architecture mapped $z_0 \sim \mathcal{N}(\mathbf{0}, I)$

1015 to a support of $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}] \in [4, 8] \times [0.01, 4]$, initialized to a gaussian approximation of samples
 1016 returned by a preliminary ABC search. We did not include $g_{\text{synA}} < 0.01$, for numerical stability.
 1017 EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an
 1018 augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The distribution
 1019 shown is that of the architecture converging with criteria $N_{\text{test}} = 100$ at greatest entropy across
 1020 random seeds.

1021 We calculated the Hessian at the mode of the inferred EPI distribution. The Hessian of a probability
 1022 model is the second order gradient of the log probability density $\log q_{\boldsymbol{\theta}}(\mathbf{z})$ with respect to the
 1023 parameters \mathbf{z} : $\frac{\partial^2 \log q_{\boldsymbol{\theta}}(\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^\top}$. With EPI, we can examine the Hessian, which is analytically available
 1024 throughout distribution, to indicate the dimensions of parameter space that are sensitive (strongly
 1025 negative eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the
 1026 emergent property produced. In Figure 1D, the eigenvectors of the Hessian v_1 (solid) and v_2
 1027 (dashed) are shown evaluated at the mode of the distribution. The length of the arrows is inversely
 1028 proportional to the square root of absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$.
 1029 Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D parameter
 1030 space are chosen arbitrarily.

1031 5.2.2 Primary visual cortex

1032 In the stochastic stabilized supralinear network [71], population rate responses \mathbf{x} to input \mathbf{h} , recur-
 1033 rent input $W\mathbf{x}$ and slow noise $\boldsymbol{\epsilon}$ are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \quad (62)$$

1034 where the noise is an Ornstein-Uhlenbeck process $\boldsymbol{\epsilon} \sim OU(\tau_{\text{noise}}, \boldsymbol{\sigma})$

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (63)$$

1035 with $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$. The noisy process is parameterized as

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (64)$$

1036 so that $\boldsymbol{\sigma}$ parameterizes the variance of the noisy input in the absence of recurrent connectivity
 1037 ($W = \mathbf{0}$). As contrast increases, input to the E- and P-populations increases relative to a baseline
 1038 input $\mathbf{h} = \mathbf{h}_b + c\mathbf{h}_c$. Connectivity (W_{fit}) and input ($\mathbf{h}_{b,\text{fit}}$ and $\mathbf{h}_{c,\text{fit}}$) parameters were fit using the
 1039 deterministic V1 circuit model [72]

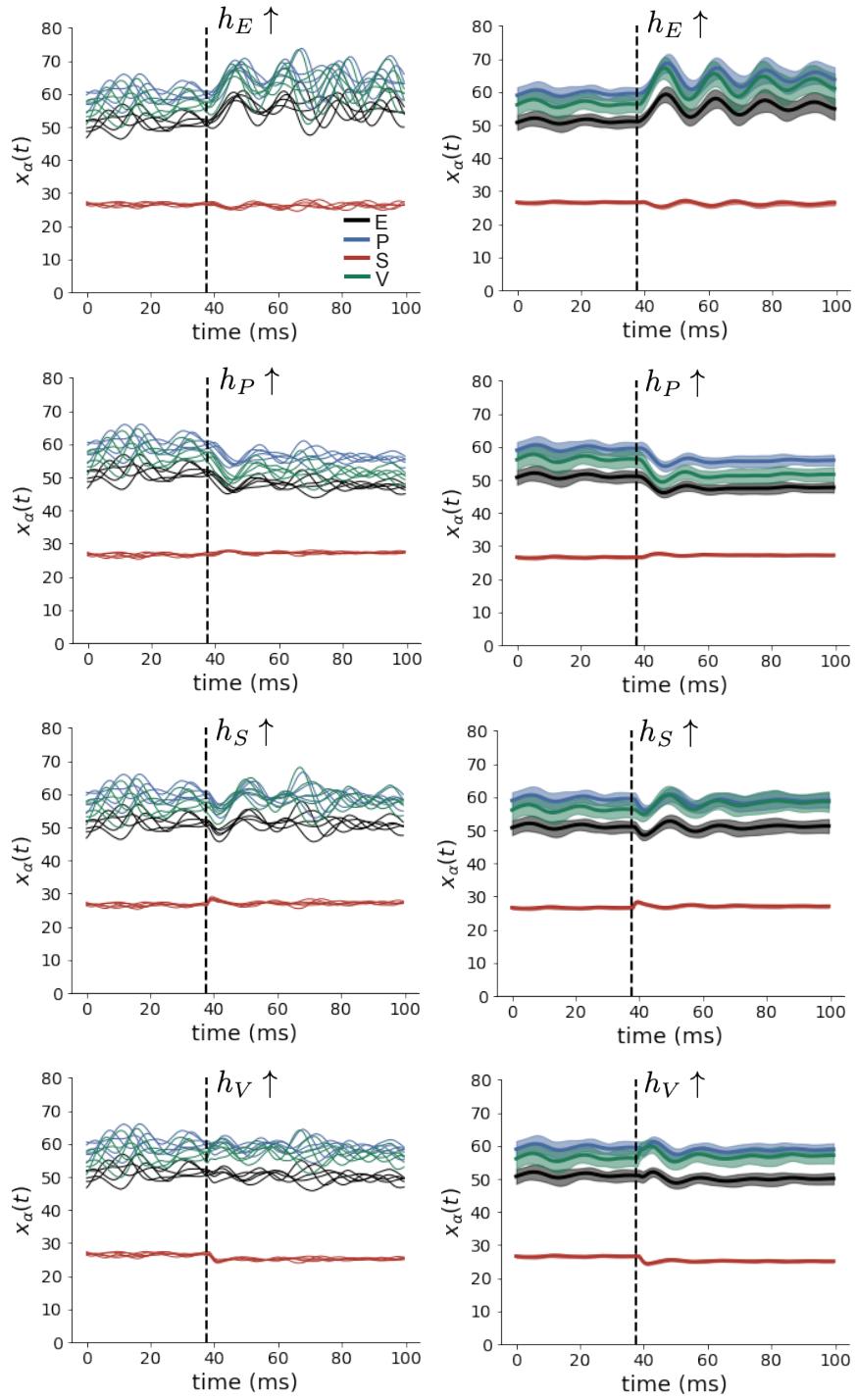


Figure 9: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

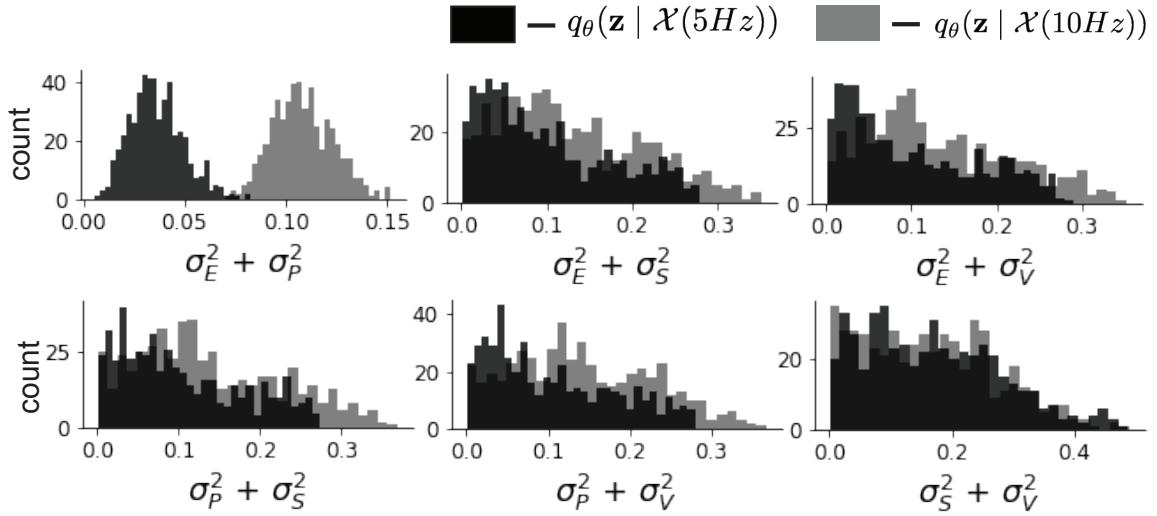


Figure 10: (V1 2) Posterior predictive distributions of the sum of squares of each pair of noise parameters.

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (65)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (66)$$

¹⁰⁴⁰ and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (67)$$

¹⁰⁴¹ To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equivalence class
¹⁰⁴²

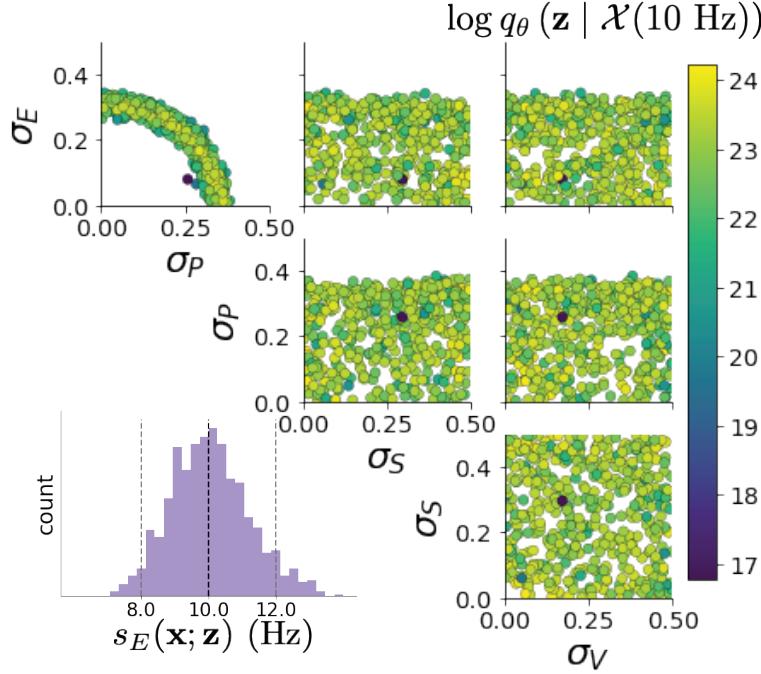


Figure 11: (V1 3) EPI posterior for $\mathcal{X}(10 \text{ Hz})$.

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (68)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (69)$$

¹⁰⁴³ and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (70)$$

¹⁰⁴⁴ Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5\text{ms}$ from an initial condition

1045 drawn from $\mathbf{x}(0) \sim U[10 \text{ Hz}, 25 \text{ Hz}]$. Standard deviation of the E-population $s_E(\mathbf{x}; \mathbf{z})$ is calculated
 1046 as the square root of the temporal variance from $t_{ss} = 75\text{ms}$ to $Tdt = 100\text{ms}$ averaged over 100
 1047 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[\sqrt{\mathbb{E}_{t > t_{ss}} \left[(x_E(t) - \mathbb{E}_{t > t_{ss}} [x_E(t)])^2 \right]} \right] \quad (71)$$

1048 For EPI in Fig 2D-E, we used a real NVP architecture with three Real NVP coupling layers
 1049 and two-layer neural networks of 50 units per layer. The normalizing flow architecture mapped
 1050 $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ to a support of $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V] \in [0.0, 0.5]^4$. EPI optimization was run using three
 1051 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an augmented Lagrangian coefficient of
 1052 $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the architectures
 1053 converging with criteria $N_{\text{test}} = 100$ at greatest entropy across random seeds.

1054 In Fig. 2E, we visualize the modes of $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ throughout the σ_E - σ_P marginal. Specifically, we
 1055 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (72)$$

1056 At each mode \mathbf{z}^* , we calculated the Hessian and visualized the sensitivity dimension in the direction
 1057 of positive σ_E .

1058 5.2.3 Primary visual cortex: challenges to analysis

1059 TODO Agostina and I are putting this together now.

1060 5.2.4 Superior colliculus

1061 In the model of Duan et al [55], there are four total units: two in each hemisphere corresponding to
 1062 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right
 1063 Pro (RP) and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related
 1064 by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh \left(\frac{u_\alpha - a}{b} \right) + \frac{1}{2} \right) \quad (73)$$

1065 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-
 1066 earity, respectively. During periods of optogenetic inactivation, activity was decreased proportional
 1067 to the optogenetic strength γ

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (74)$$

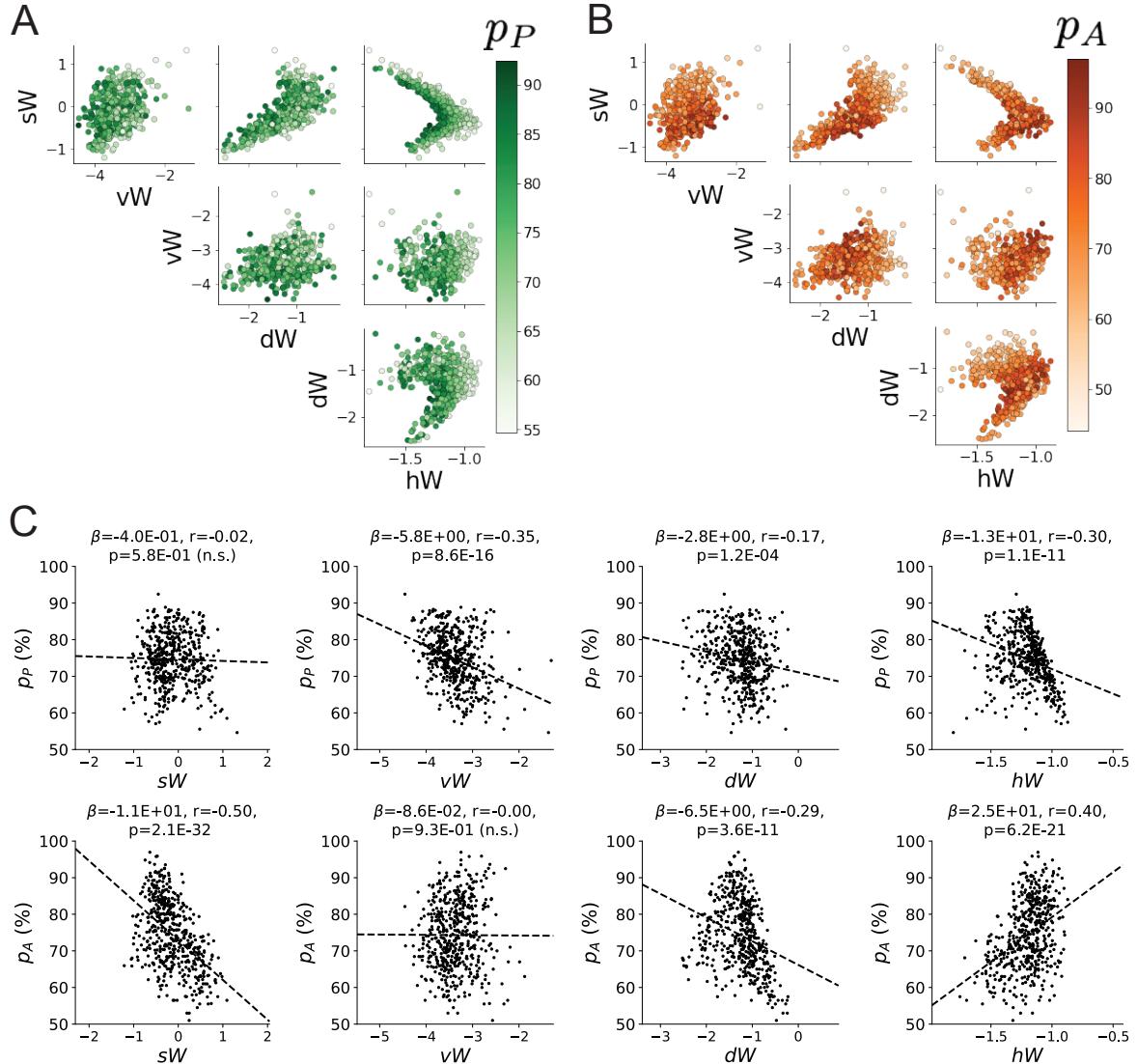


Figure 12: (SC1): **A.** Same pairplot as Fig. 3C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

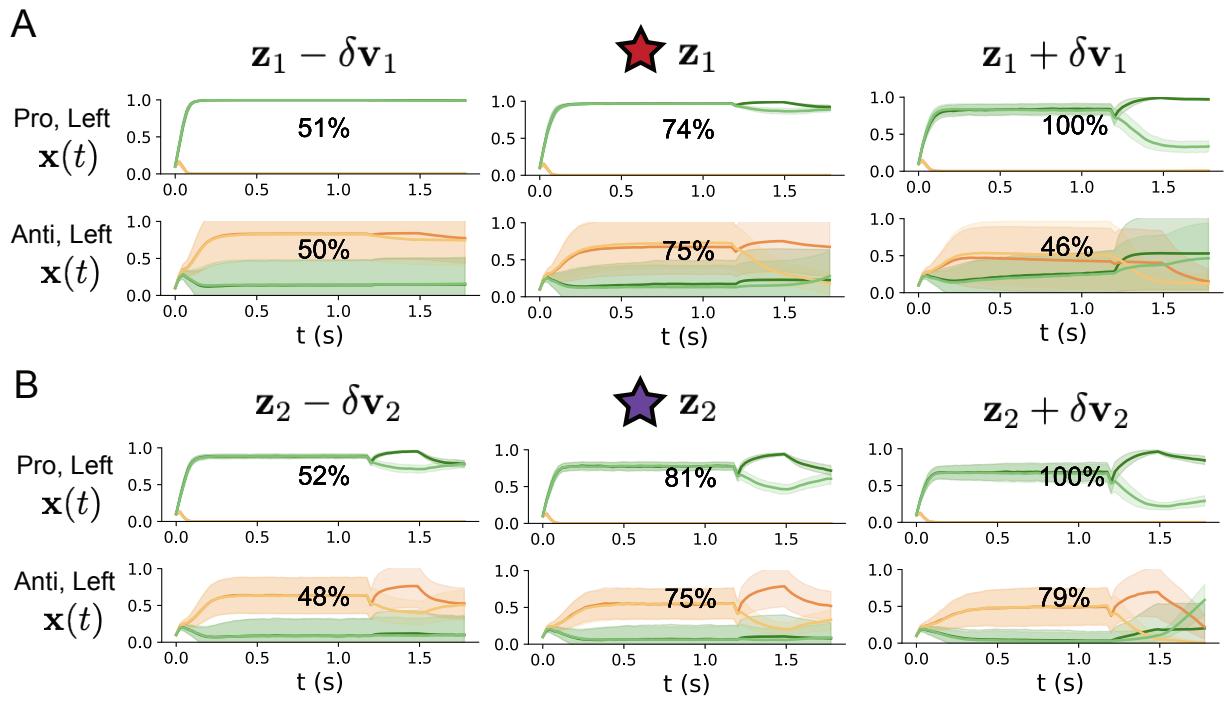


Figure 13: (SC2): **A.** Simulations in network regime \mathbf{z}_1 (center) with simulations given connectivity perturbations in the negative direction of the sensitivity vector \mathbf{v}_1 (left) and positive direction (right). **B.** Same as A for network regime \mathbf{z}_2 .

1068 We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (75)$$

1069 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (76)$$

1070 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2 . The weight
1071 matrix has 4 parameters sW , vW , hW , and dW :

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (77)$$

1072 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (78)$$

1073 There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (79)$$

1074 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (80)$$

1075 rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (81)$$

1076

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (82)$$

1077 a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (83)$$

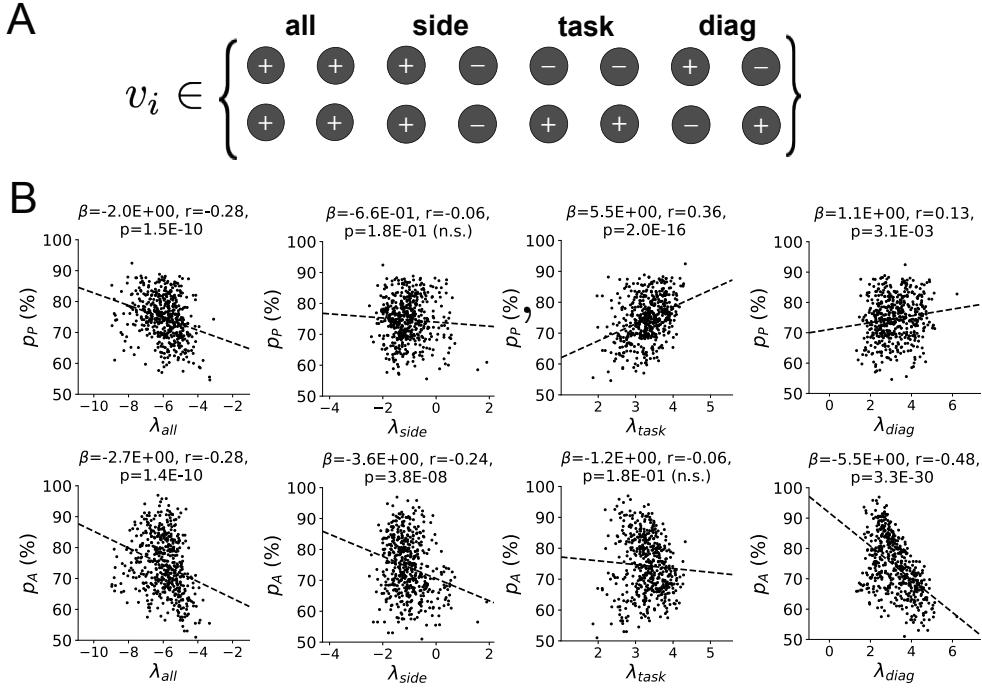


Figure 14: (SC3): **A.** Invariant eigenvectors of connectivity matrix W . **B.** Eigenvalues of connectivities of EPI distribution versus task accuracies.

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

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases} \quad (84)$$

The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{P,\text{bias}} = 0.5$, $I_{P,\text{rule}} = 0.6$, $I_{A,\text{rule}} = 0.6$, $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

The accuracies of p_P and p_A are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (85)$$

and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (86)$$

given that the stimulus is on the left side, where Θ is the Heaviside step function, and the accuracy is averaged over 200 independent trials. The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (87)$$

where $\beta = 100$.

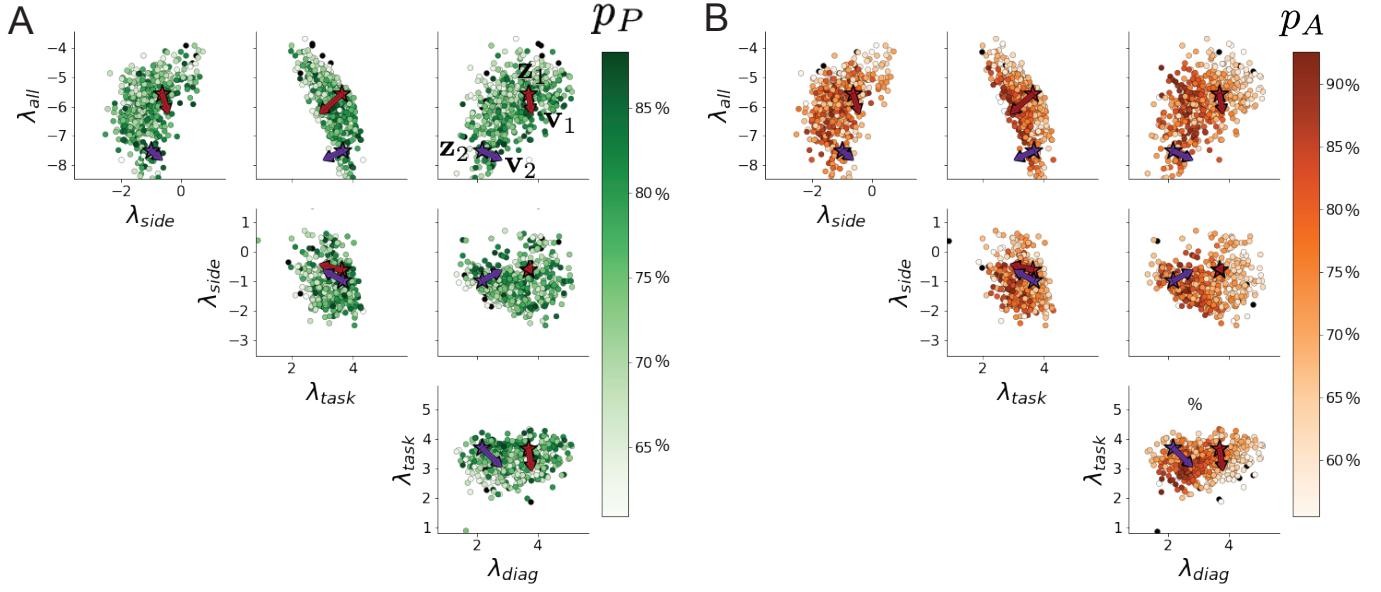


Figure 15: (SC4): **A.** Pairplots of eigenvalues of connectivity matrices in EPI distribution colored by Pro task accuracy. Red and purple stars and arrows correspond to eigenvalues and sensitivity directions \mathbf{z}_1 , \mathbf{z}_2 , \mathbf{v}_1 , and \mathbf{v}_2 . **B.** Same colored by Anti task accuracy.

1086 Writing the EPI posterior as a maximum entropy distribution, $T(\mathbf{x}; \mathbf{z})$ is comprised of both these
 1087 first and second moments of the accuracy in each task (as in Equations 29 and 30)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p(\mathbf{x}; \mathbf{z})_P \\ p(\mathbf{x}; \mathbf{z})_A \\ (p(\mathbf{x}; \mathbf{z})_P - 75\%)^2 \\ (p(\mathbf{x}; \mathbf{z})_A - 75\%)^2 \end{bmatrix}, \quad (88)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 7.5\%^2 \\ 7.5\%^2 \end{bmatrix}. \quad (89)$$

1088 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 1089 epoch of 2,000 iterations(see Section 5.1.3). The optimization converged after six epochs (Fig. 17).

1090 For EPI in Fig. 3C, we used a real NVP architecture with three coupling layers of affine transfor-
 1091 mations parameterized by two-layer neural networks of 50 units per layer. The initial distribution
 1092 was a standard isotropic gaussian $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to a support of $\mathbf{z}_i \in [-5, 5]$. We used an
 1093 augmented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution
 1094

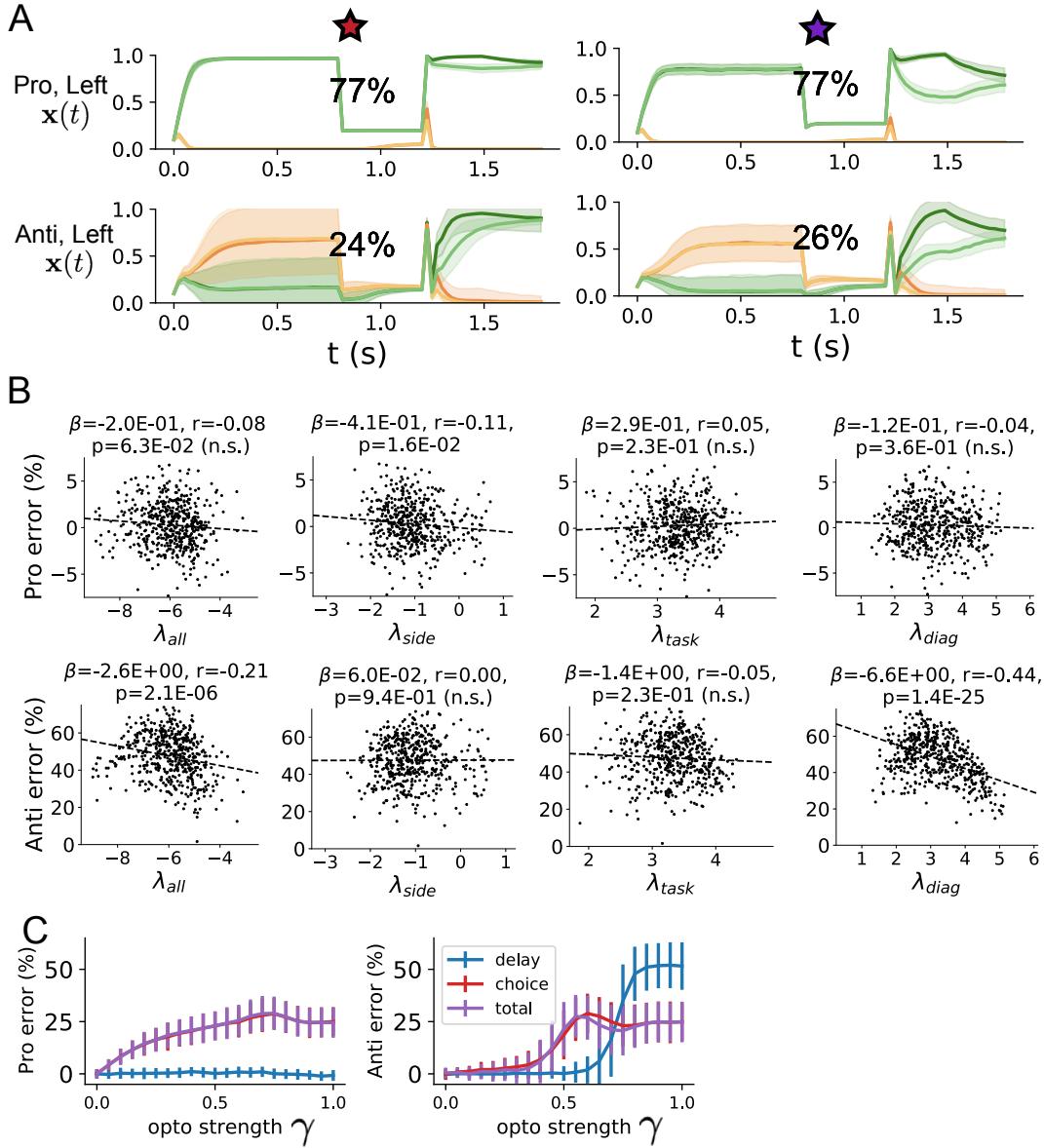


Figure 16: (SC5): **A.** Response of each parameter regime to optogenetic silencing during the delay period. **B.** Connectivity eigenvalues versus the task error induced by delay period inactivation. **C.** Error induced by delay period inactivation with increasing optogenetic strength. Means and standard deviations are calculated across the entire EPI posterior.

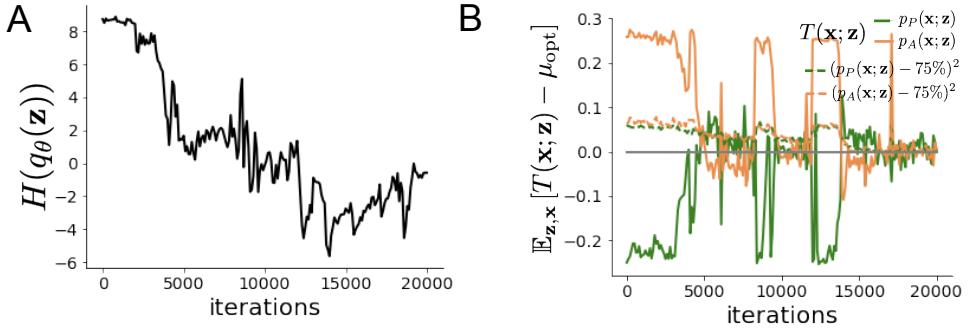


Figure 17: (SC6): A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

1095 shown is that of the architecture converging with criteria $N_{\text{test}} = 25$ at greatest entropy across
1096 random seeds.

1097 To make sense of this inferred distribution, we identified two modes used to represent the two
1098 regimes of connectivity in this posterior:

$$\begin{aligned} \mathbf{z}_1 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW > 0 \end{aligned} \tag{90}$$

1099 and

$$\begin{aligned} \mathbf{z}_2 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW < 0 \end{aligned} \tag{91}$$

1100 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-
1101 tion of the symmetric connectivity matrices $W = V\Lambda V^{-1}$, which results in the same basis vectors
1102 \mathbf{v}_i for all W parameterized by \mathbf{z} (Fig. 14A). These basis vectors have intuitive roles in processing
1103 for this task, and are accordingly named the *all* mode - all neurons co-fluctuate, *side* mode - one
1104 side dominates the other, *task* mode - the Pro or Anti populations dominate the other, and *diag*
1105 mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair. We found
1106 significant trends across the EPI posterior connectivities: the eigenvalues λ_{task} and λ_{diag} were cor-
1107 related with p_P , while λ_{all} was anticorrelated with p_P . λ_{all} , λ_{side} , and λ_{diag} were all significantly
1108 anticorrelated with p_A .

1109 Under this decomposition, we can re-visualize the posterior in eigenvalue space (Fig. 15). Fur-
1110 thermore, we can project the dimensions of sensitivity into eigenvalue space as well, giving us a
1111 more intuitive sense of how connectivity affects computation in each regime. We see that sensitivity

1112 dimensions \mathbf{v}_1 and \mathbf{v}_2 , which cause p_P to increase and a regime dependent change in p_A , both point
1113 in the direction of increasing λ_{side} and decreasing λ_{task} . These eigenvalue changes are evident in
1114 the simulations of connectivity perturbations away from the modes (Fig. 13). As the component
1115 of connectivity along \mathbf{v}_1 and \mathbf{v}_2 becomes stronger (left-to-right), there is less separation between
1116 Pro an Anti populations (lower λ_{task}) and greater separation between Left and Right populations
1117 following stimulus presentation (greater λ_{side}). A key differentiating factor is that \mathbf{v}_1 substantially
1118 increases λ_{diag} , while \mathbf{v}_2 does not.

1119 During optogenetic silencing simulations, activations $x_\alpha(t)$ were set to a fraction of their values ($1 -$
1120 γ), where γ is the optogenetic perturbation strength. We found that λ_{all} and λ_{diag} were significantly
1121 anticorrelated with Anti error during delay period inactivation. Delay period inactivation was from
1122 $0.8 < t < 1.2$, choice period inactivation was for $t > 1.2$ and total inactivation was for the entire
1123 trial.

1124 5.2.5 Rank-2 RNN

1125 Traditional approaches to likelihood-free inference – approximate Bayesian computation (ABC)
1126 methods – randomly sample parameters \mathbf{z} until a suitable set is obtained. State-of-the-art ABC
1127 methods leverage sequential Monte Carlo (SMC) sampling techniques to obtain parameter sets more
1128 efficiently. To obtain more parameter samples, SMC-ABC must be run from scratch again. ABC
1129 methods do not confer log probabilities of samples. Like EPI, sequential neural posterior estimation
1130 (SNPE) uses deep learning to produce flexible posterior approximations. Like traditional Bayesian
1131 inference methods, SNPE conditions directly on the statistics of data. This differs from EPI, where
1132 posteriors are conditioned on emergent properties (moment constraints on the posterior predictive
1133 distribution). Peculiarities of SNPE (density estimation approach, two deep networks) make scaling
1134 in \mathbf{z} prohibitive.

1135 SMC-ABC has many hyperparameters, of which pyABC selects automatically by running some ini-
1136 tial diagnostics upon initialization. In concurrence with the literature, SMC-ABC fails to converge
1137 around 25-30 dimensions, since it's proposal samples never get close enough to the target statis-
1138 tics. We searched over many SNPE hyperparameter choices: $n_{\text{train}} \in [2,000, 10,000, 100,000]$ is the
1139 number of simulations run per training epoch, and $n_{\text{mades}} \in [2, 3]$ is the number of masked autore-
1140 gressive density estimators in the deep parameter distribution architecture. The greater n_{train} , the
1141 longer each epoch will take, but the more likely SNPE may converge during that epoch. Greater
1142 n_{mades} increases the flexibility of the deep parameter distribution of SNPE, but slows optimization.

₁₁₄₃ For the timing plot, we show the fastest among all of these choices, and for the convergence plot,
₁₁₄₄ we show the best convergence among all of these choices. During optimization, we used $n_{\text{atom}}=100$
₁₁₄₅ atomic proposals as is recommended.