

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 that govern variability in a
94 stochastic four neuron-type model of primary visual cortex. Second, we discover connectivity pat-
95 terns in superior colliculus resilient to optogenetic perturbation by using EPI to condition on rapid
96 task switching. The novel scientific insights offered by EPI contextualize and clarify the previous
97 studies exploring these models [54, 55]. Third, we emphasize the scalability of EPI by inferring
98 high-dimensional distributions of RNNs exhibiting stable amplification. These results point to the
99 value of deep inference for the 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 unified intermediacy (Fig. 1B 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. 1B contours) and have diminished or preserved network syncing, respectively (Fig. 1F activity traces, Fig. S TODO) (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,

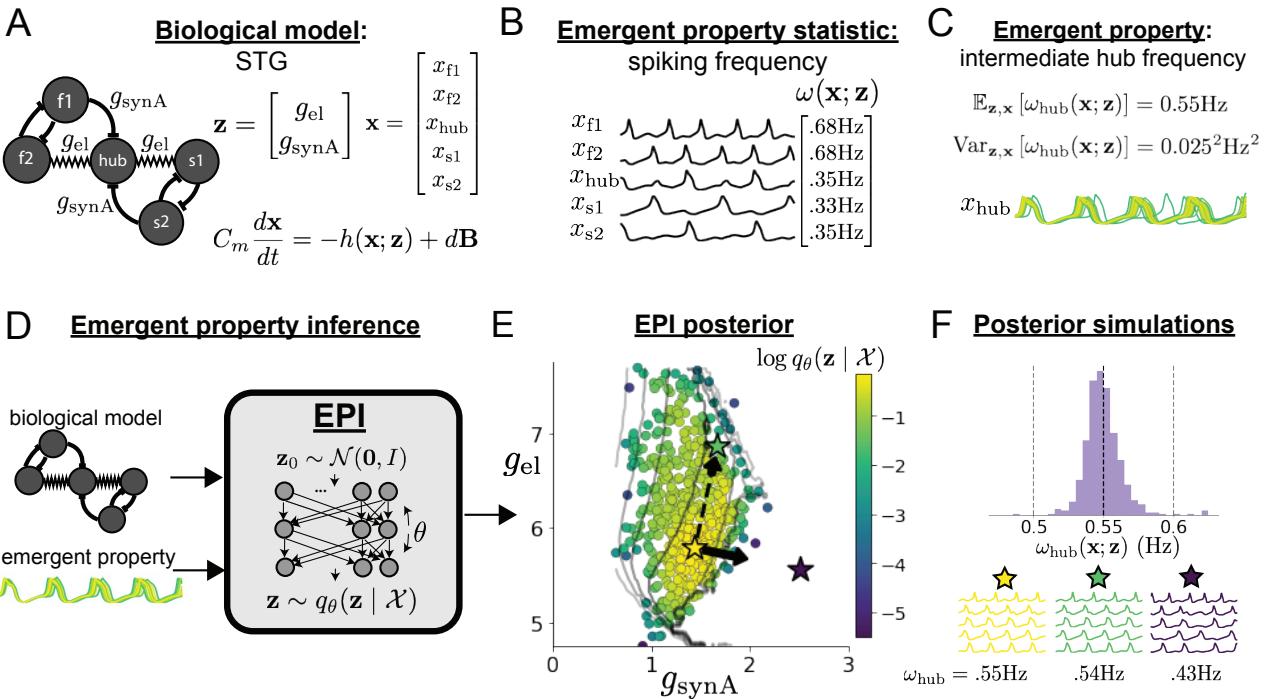


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based biophysical model of the STG subcircuit. In the STG model, jagged connections indicate electrical coupling having electrical conductance g_{el} . Other connections in the diagram are inhibitory synaptic 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. 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. 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 distribution of the model parameters $\mathbf{z} = [g_{el}, g_{synA}]$ producing intermediate hub frequency. Deep probability distributions map a simple random variable $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$ through a deep neural network with weights and biases $\boldsymbol{\theta}$ to parameters $\mathbf{z} = q_{\boldsymbol{\theta}}(\mathbf{z}_0)$ distributed as $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X})$. In EPI optimization, stochastic gradient steps in $\boldsymbol{\theta}$ are taken such that entropy is maximized, and the emergent property \mathcal{X} is produced. **E.** The EPI distribution of STG model parameters producing intermediate hub frequency. Samples are colored by log probability density. Distribution contours of hub neuron frequency from mean of .55 Hz are shown at levels of .525, .53, ..., .575 Hz (dark to light gray away from mean). Frequencies are averages over the stochasticity of the model. 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. Simulated activity is shown for three samples (stars). **F** Simulations from parameters in E. (Top) The predictive distribution of the posterior obeys the constraints stipulated by 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, ?, 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:

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

158 where $f(\mathbf{x}, \mathbf{z})$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ are defined as in Equation 10. 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 neural 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 noise across neural population types governs Fano factor in a stochastic inhibition stabilized 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) – composing

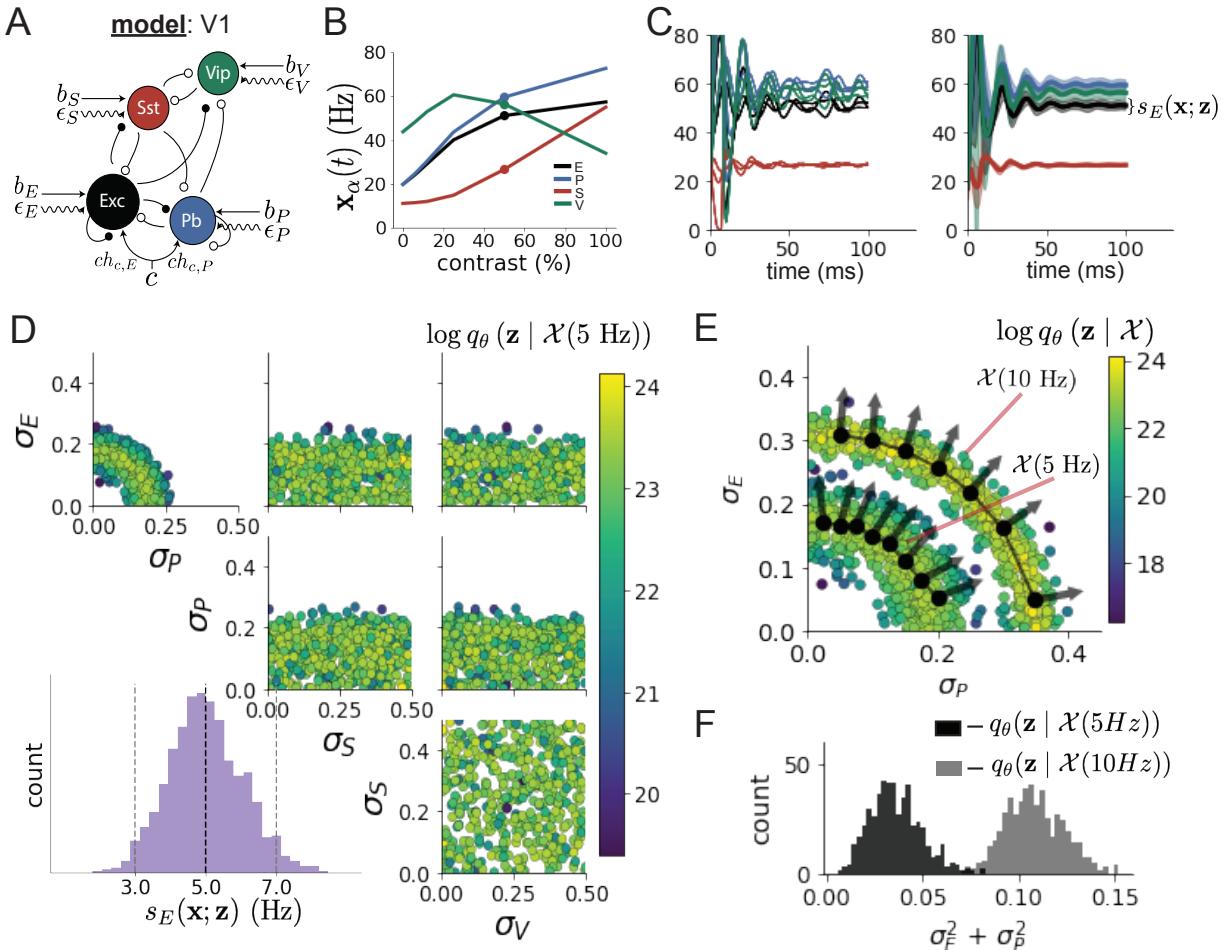


Figure 2: Emergent property inference of a stochastic stabilized supralinear network. **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). 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}_b . Additionally, each neural population receives a slow noisy input ϵ . **B.** Responses of the deterministic model ($\epsilon = \mathbf{0}$) to varying contrasts. The response at 50% contrast (dots) is the focus of our analysis. **C.** Paradoxical response of the stochastic model to a small increase in input to the P-population. **D.** EPI posterior of noise parameters \mathbf{z} conditioned on realistic E-population Fano factors. The posterior predictive distribution is shown on the bottom-left. and the mode of the distribution is starred. **E.** (Top) Enlarged visualization of the σ_E - σ_P marginal distribution of the posterior. Each gray dot is a choice of σ_P , for which a constrained mode $z^*(\sigma_P, P)$ is chosen. The arrows show the most sensitive dimensions of the Hessian evaluated at these modes. (Bottom) Such sensitive dimensions of the Hessian (dots) are significantly more sensitive than randomly chosen dimensions (box and whiskers). **F.** The Fano factor of the E-population is strongly correlated with each other neuron-type population. **G.** Mean and standard deviation (across EPI posterior) of Fano factor of each neuron-type population at each level of contrast.

194 80% of GABAergic interneurons in V1 [65, 66, 67], and that these inhibitory cell types follow
 195 specific connectivity patterns (Fig. 2A) [68]. Recent theoretical advances [54, 69, 70], have only
 196 started to address the consequences of this multiplicity in the dynamics of V1, strongly relying on
 197 linear theoretical tools. Here, we use EPI to characterize the properties of slow noise in a stochastic
 198 version of this model, which result in biologically realistic responses.

199 We considered the contrast response 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 nonlinearity $\phi = \|\cdot\|_+^2$ (see Section 5.2.2). The
 203 circuit model evolves from an initial condition $\mathbf{x}(0) \sim \mathcal{U}([10, 25])$ with time constant $\tau = 1\text{ms}$
 204 according to a contrast-dependent input \mathbf{h} and slow noise ϵ of time constant $\tau_{\text{noise}} = 5\text{ms}$. This
 205 model is the stochastic stabilized supralinear network (SSSN) [71] generalized to have inhibitory
 206 multiplicity

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

207 As contrast increases, input to the E- and P-populations increases relative to a baseline input \mathbf{h}_b
 208 via \mathbf{h}_c

$$\mathbf{h} = \mathbf{h}_b + c\mathbf{h}_c, \quad (5)$$

209 where $h_{c,E}, h_{c,P} > 0$ and $h_{c,S}, h_{c,V} = 0$. In this analysis, we fixed W, \mathbf{h}_b , and \mathbf{h}_c to values fit to
 210 mean contrast responses in mice with the deterministic model [72] ($\epsilon = \mathbf{0}$, Fig. 2B, see Section
 211 5.2.2). At all contrasts, the E-population of this SSSN is unstable without recurrent inhibitory
 212 feedback. At 50% contrast, only the P-population exhibits the paradoxical effect (2C, Fig. 9), so
 213 the network is P-stabilized.

214 The slow noise of the SSSN is an Ornstein-Uhlenbeck process

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \sigma_\alpha dB, \quad (6)$$

215 parameterized by σ_α , which can be different for each neuron type,

$$\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top. \quad (7)$$

216 For this SSSN, we are interested in the parameters of slow noise that produce realistic stochastic
 217 fluctuations. Here, we quantify this emergent property as having an excitatory population Fano
 218 factor near 1:

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} [F_E(\mathbf{x}; \mathbf{z})] &= 1 \\ \text{Var}_{\mathbf{z}} [F_E(\mathbf{x}; \mathbf{z})] &= 0.125^2, \end{aligned} \quad (8)$$

219 where $F_\alpha(\mathbf{x}; \mathbf{z})$ is the Fano factor of the α -population.

220 We ran EPI to obtain a posterior $q_{\theta}(\mathbf{z} | \mathcal{X})$, where each parameter \mathbf{z} produces biologically realistic
221 levels of E-population variability (Fig. 2D). From the marginal distribution of σ_E and σ_P (Fig.
222 2D, top-left), we can see that $F_E(\mathbf{x}; \mathbf{z})$ is sensitive to the combination of σ_E and σ_P . In fact, the
223 posterior obtained through EPI offers exactly how this sensitivity changes along this ridge of the
224 posterior (Fig. 2E). σ_S and σ_V are degenerate with respect to $F_E(\mathbf{x}; \mathbf{z})$ evidenced by the uniform
225 distribution in those dimensions of the posterior (Fig. 2D, bottom-right). Together, this posterior
226 indicates a parametric manifold of degeneracy with respect to Fano factor: the ridge visualized in
227 the σ_E - σ_P marginal (Fig. 10) and the dimensions of σ_S and σ_V .

228 Greater σ_E and σ_P confer greater Fano factor, and the Fano factors of each neuron-type are
229 strongly correlated across the posterior (Fig 2F), showing that Fano factor of each neuron-type
230 can be modulated globally via σ_E and σ_P . Furthermore, across the entire posterior distribution of
231 noise parameterizations, we find that when contrast is increased above 50%, variability is quenched
232 for all neuron types (Fig 2G). In summary, we used EPI to obtain a posterior of SSSNs producing
233 realistic Fano factors, which allowed degenerate manifold identification via sample visualization,
234 fast sensitivity measurements via Hessian evaluation, and predictions of variability quenching.

235 3.4 EPI identifies neural mechanisms of flexible task switching

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

248 the trial. The dynamics evolve with timescale $\tau = 0.09$ governed by connectivity weights W

$$\begin{aligned} \tau \frac{d\mathbf{u}}{dt} &= -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B} \\ \mathbf{x} &= \phi(\mathbf{u}) \end{aligned} \quad (9)$$

249 with white noise of variance 0.2^2 . The input \mathbf{h} is comprised of a cue-dependent input to the Pro
250 or Anti populations, a stimulus orientation input to either the Left or Right populations, and
251 a choice-period input to the entire network (see Section 5.2.3). Here, we use EPI to determine
252 the changes in network connectivity $\mathbf{z} = [sW, vW, dW, hW]^\top$ resulting in execution of rapid task
253 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 $5\%^2$.

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 75\% \\ 75\% \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 5\%^2 \\ 5\%^2 \end{bmatrix} \end{aligned} \quad (10)$$

259 A variance of $5\%^2$ performance in each task will confer a posterior producing performances ranging
260 from about 65% – 85%, allowing us to examine the properties of connectivity that yield better
261 performance.

262 We ran EPI to obtain SC model connectivity parameters \mathbf{z} producing rapid task switching (Fig.
263 3C). Some parameters were predictive of accuracy while others were not (Fig. 11), and often
264 had different effects on p_P and p_A . To make sense of this inferred distribution, we took the
265 eigendecomposition of the symmetric connectivity matrices $W = V\Lambda V^{-1}$, which results in the
266 same basis vectors \mathbf{v}_i for all W parameterized by \mathbf{z} (Fig. 12A). These basis vectors have intuitive
267 roles in processing for this task, and are accordingly named the *all* mode - all neurons co-fluctuate,
268 *side* mode - one side dominates the other, *task* mode - the Pro or Anti populations dominate the
269 other, and *diag* mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite
270 pair.

271 Greater λ_{task} , λ_{side} , and λ_{diag} all produce greater Pro accuracy. This shows that strong task
272 representations and hemispheric

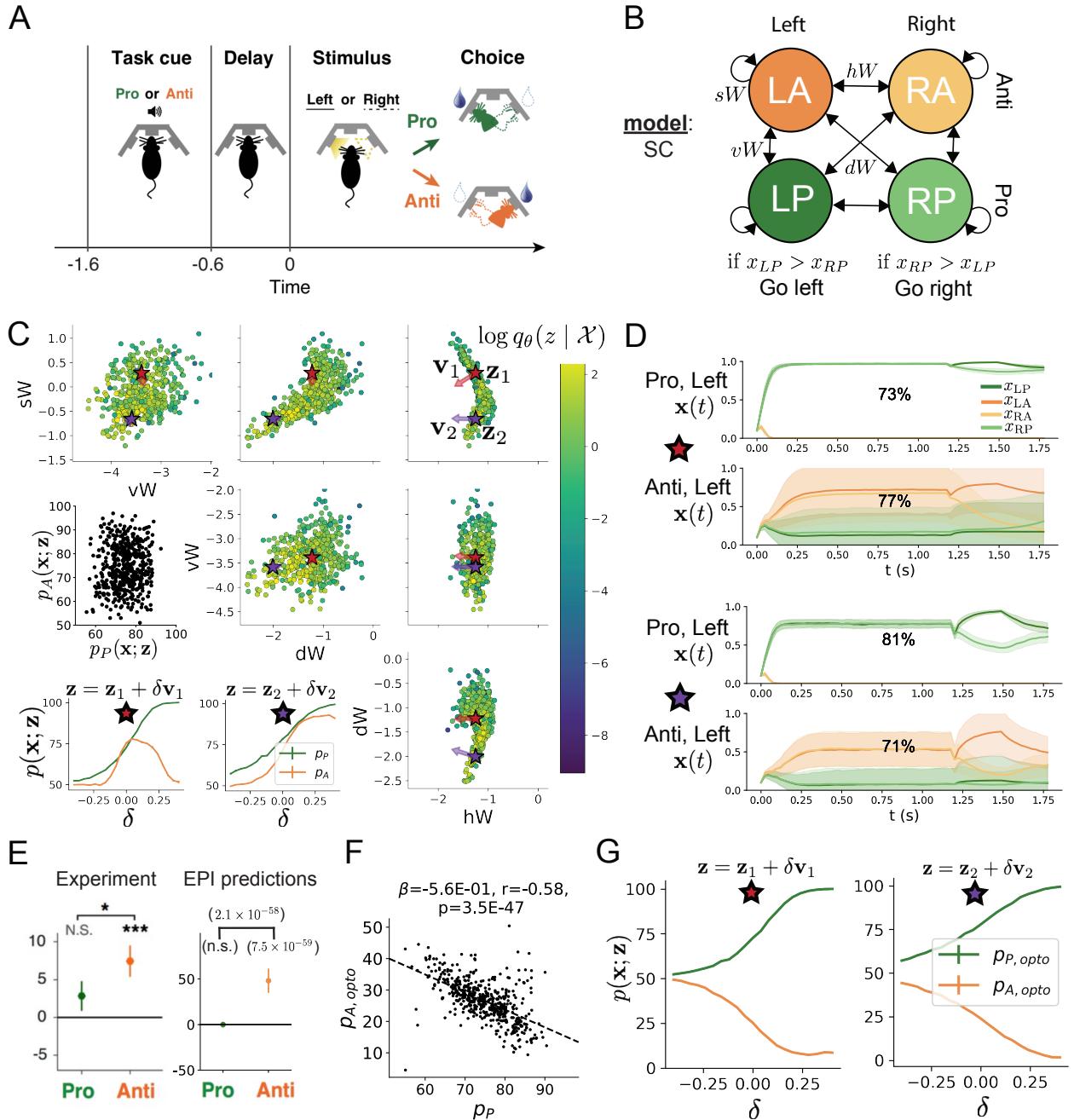


Figure 3: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. Subscripts P and A of connectivity weights indicate Pro or Anti populations. **C.** The EPI parameter distribution of rapid task switching networks. Black star indicates parameter choice of simulations (D). **D.** Simulations of an SC network from the EPI distribution with 75% accuracy in each task. Top row shows no inactivation during Pro and Anti trials, and bottom row shows simulations with delay period inactivation (optogenetic strength $\gamma = 0.7$). Shading indicates standard deviation across trials. **E.** Difference in performance of each task during inactivation. Inactivation level γ scales from no inactivation (0) to full inactivation (1). We compare delay period inactivation $1.2 < t < 1.5$ (blue), choice period inactivation $1.5 < t$ (red), and total inactivation $0 \leq t \leq 1.8$ (purple). **F.** The effect of delay period inactivation on Anti accuracy versus dynamics eigenvalues.

dominance in the dynamics result in better execution of the Pro task. By visualizing these four variables together by p_A (Fig. 13B), we see that low λ_{task} and λ_{diag} producing strong Anti accuracy also have high λ_{side} and λ_{all} . Thus, stronger hemispheric dominance, relaxed task and diag mode dynamics, and slower circuit-wide decay result in greater Anti accuracy.

In agreement with experimental results from Duan et al., we found that inactivation above nominal strength during the delay period consistently decreased performance in the Anti task, but had no consistent effect on the Pro task (Fig. 3E) e.g. (Fig. 3D, bottom). This difference in resiliency across tasks to delay perturbation is a prediction made by the inferred EPI distribution, rather than an emergent property that was conditioned upon. Even though p_P and p_A are anticorrelated in the EPI posterior ($r = -0.15$, $p = 3.68 \times 10^{-12}$), greater p_P and p_A both result in decreased resiliency to delay perturbation in the Anti task (Fig. 14). Ultimately, lower λ_{side} and λ_{all} and greater λ_{task} produce networks more robust to delay perturbation in the Anti task (Fig. 3F)).

In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task switching. This posterior revealed the mechanisms leading to greater accuracy in each task as well as those increasing resiliency to perturbation in the Anti task. Importantly, every connectivity from this inferred distribution predicts fragility and robustness of performance in the Anti and Pro tasks, respectively. EPI allows us to conclude that since *all* parameters of this model producing rapid task switching make such an experimentally verified prediction, we have a well-chosen model.

3.5 EPI scales well to high-dimensional parameter spaces

Here, we are interested in the scalability of EPI in number of parameters $|\mathbf{z}|$. We consider rank-2 RNN with N neurons of connectivity

$$W = UV^\top + g\chi \quad (11)$$

and dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x} \quad (12)$$

where $U = [\mathbf{u}_1 \ \mathbf{u}_2]$, $V = [\mathbf{v}_1 \ \mathbf{v}_2]$, $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2 \in [-1, 1]^N$, and $g = 0.01$.

We want to learn distributions of connectivity that produce stable amplification. Two conditions are both necessary and sufficient for RNNs to exhibit stable amplification [?]. These conditions are

299 inequalities on $\text{real}(\lambda_1)$ and λ_1^s the maximal real eigenvalue of W and the maximum eigenvalue of
300 $W^s = \frac{W+W^\top}{2}$, respectively.

301 In our analysis, we seek to condition rank-2 networks of increasing size on a regime of stable ampli-
302 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.
303 EPI can naturally condition on this emergent property

$$\begin{aligned}\mathcal{X} &: \mathbb{E}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}. \end{aligned}\tag{13}$$

304 In contrast, SNPE cannot condition on the variance of observations across posterior. Thus, we
305 condition on an observation x_0 located at the mean of our desired emergent property.

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}\tag{14}$$

306 ABC methods define tolerance ϵ and distance for observed data x_0 . Here, we chose $\epsilon = 0.5$, an $l - 2$
307 distance, and the same choice for x_0 as in Equation 14.

308 EPI is capable of scaling to higher dimensional parameter spaces than ABC and SNPE. EPI consis-
309 tently produces the same posterior predictive distribution independent of the dimensionality. SMC
310 produces a limited variety of parameters due to the nature of its proposal generation algorithm,
311 yet all parameters obtained produce stable amplification. SNPE's posterior predictive distribution
312 is not necessarily close to the conditioning point, and is very dependent on dimensionality.

313 4 Discussion

314 NOTE: This is the old discussion section. I will rewrite this based on our discussion of
315 the rest of the draft.

316 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets
317 [15, 16, 17, 18, 20, 22, 24, 26, 27, 28, 29] (see review, [30]). Such careful inference procedures
318 are developed for these statistical models allowing precise, quantitative reasoning, which clarifies
319 the way data informs beliefs about the model parameters. However, these statistical models lack
320 resemblance to the underlying biology, making it unclear how to go from the structure revealed by
321 these methods, to the neural mechanisms giving rise to it. In contrast, theoretical neuroscience has

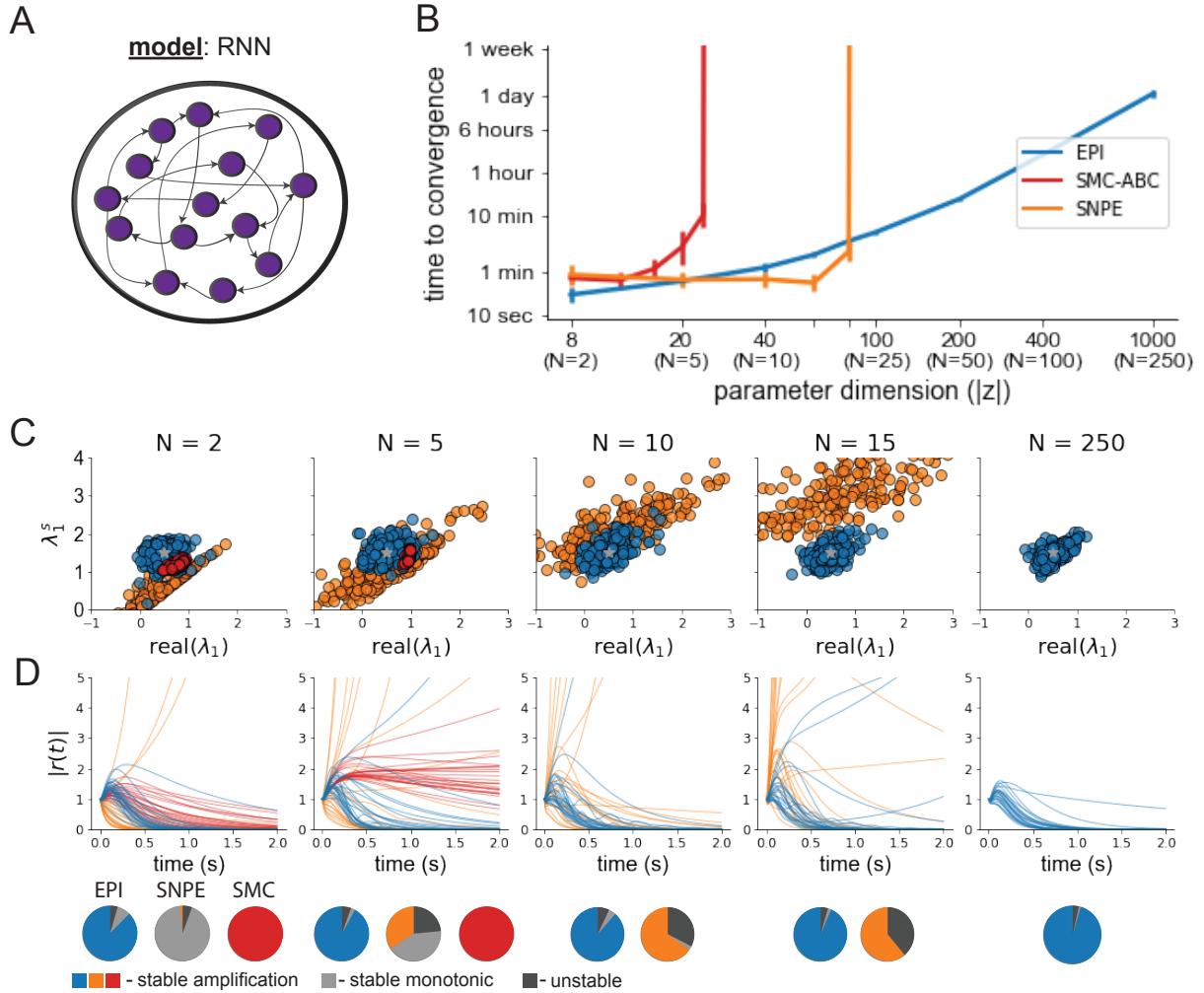


Figure 4: **A.** Recurrent neural network. **B.** EPI scales with z to high dimensions. Convergence definitions: EPI (blue) - satisfies all moment constraints, SNPE (orange)- produces at least $2/n_{\text{train}}$ parameter samples are in the bounds of emergent property (mean ± 0.5), and SMC-ABC (red) - 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 . **E.** Ratio of obtained samples producing stable amplification.

322 focused on careful mechanistic modeling and the production of emergent properties of computation.
323 The careful steps of *i.*) model design and *ii.*) emergent property definition, are followed by *iii.*)
324 practical inference methods resulting in an opaque characterization of the way model parameters
325 govern computation. In this work, we replaced this opaque procedure of parameter identification
326 in theoretical neuroscience with emergent property inference, opening the door to careful inference
327 in careful models of neural computation.

328 Biologically realistic models of neural circuits often prove formidable to analyze. Two main factors
329 contribute to the difficulty of this endeavor. First, in most neural circuit models, the number
330 of parameters scales quadratically with the number of neurons, limiting analysis of its parameter
331 space. Second, even in low dimensional circuits, the structure of the parametric regimes governing
332 emergent properties is intricate. For example, these circuit models can support more than one
333 steady state [74] and non-trivial dynamics on strange attractors [75].

334 In Section 3.3, we advanced the tractability of low-dimensional neural circuit models by showing
335 that EPI offers insights about cell-type specific input-responsivity that cannot be afforded through
336 the available linear analytical methods [54, 69, 70]. By flexibly conditioning this V1 model on
337 different emergent properties, we performed an exploratory analysis of a *model* rather than a
338 dataset, generating a set of testable hypotheses, which were proved out. Furthermore, exploratory
339 analyses can be directed towards formulating hypotheses of a specific form. For example, model
340 parameter dependencies on behavioral performance can be assessed by using EPI to condition on
341 various levels of task accuracy (See Section 3.4). This analysis identified experimentally testable
342 predictions (proved out *in-silico*) of patterns of effective connectivity in SC that should be correlated
343 with increased performance.

344 In our final analysis, we presented a novel procedure for doing statistical inference on interpretable
345 parameterizations of RNNs executing simple tasks. Specifically, we analyzed RNNs solving a pos-
346 terior conditioning problem in the spirit of [76, 77]. This methodology relies on recently extended
347 theory of responses in random neural networks with low-rank structure [78]. While we focused
348 on rank-1 RNNs, which were sufficient for solving this task, this inference procedure generalizes
349 to RNNs of greater rank necessary for more complex tasks. The ability to apply the probabilistic
350 model selection toolkit to RNNs should prove invaluable as their use in neuroscience increases.

351 EPI leverages deep learning technology for neuroscientific inquiry in a categorically different way
352 than approaches focused on training neural networks to execute behavioral tasks [79]. These works
353 focus on examining optimized deep neural networks while considering the objective function, learn-

354 ing rule, and architecture used. This endeavor efficiently obtains sets of parameters that can be
355 reasoned about with respect to such considerations, but lacks the careful probabilistic treatment of
356 parameter inference in EPI. These approaches can be used complementarily to enhance the practice
357 of theoretical neuroscience.

358 **TODO** *merge this point in*

359 While much research in computational neuroscience has focused on optimizing neural architectures
360 to process information and accomplish tasks [79], structure in parameter space of the set of opti-
361 mized solutions is rarely discussed and lacks a probabilistic treatment. Talk about Wykтор’s work
362 here [80].

363 **Acknowledgements:**

364 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
365 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
366 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
367 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
368 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James
369 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

370 **Data availability statement:**

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

373 **Code availability statement:**

374 The software written for the current study is available from the corresponding author upon rea-
375 sonable request.

376 **References**

- 377 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
378 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 379 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
380 21(1):25–45, 1998.
- 381 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.

- 382 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in
383 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 384 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
385 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
386 models. *PLoS Comput Biol*, 3(10):e189, 2007.
- 387 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
388 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
389 sion model. *Neuron*, 82(4):809–821, 2014.
- 390 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-
391 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 392 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
393 networks. *Physical review letters*, 61(3):259, 1988.
- 394 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
395 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
396 17(11):4382–4388, 1997.
- 397 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration
398 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 399 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
400 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 401 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
402 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 403 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
404 model. *PLoS computational biology*, 2(7):e94, 2006.
- 405 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
406 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 407 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
408 13(8):1713–1720, 2001.

- 409 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
410 A statistical paradigm for neural spike train decoding applied to position prediction from
411 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
412 7425, 1998.
- 413 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
414 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 415 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.
416 A point process framework for relating neural spiking activity to spiking history, neural
417 ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 418 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise corre-
419 lations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–
420 1012, 2006.
- 421 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan
422 Segev. A novel multiple objective optimization framework for constraining conductance-based
423 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 424 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature
425 analysis. *Neural computation*, 19(4):1022–1038, 2007.
- 426 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
427 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of
428 neural population activity. In *Advances in neural information processing systems*, pages
429 1881–1888, 2009.
- 430 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and
431 Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural
432 information processing systems*, 24:1350–1358, 2011.
- 433 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In
434 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 435 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-
436 dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,
437 9(3):e1002922, 2013.

- 438 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
439 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
440 *Science*, 349(6244):184–187, 2015.
- 441 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq
442 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in
443 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 444 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
445 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th In-*
446 *ternational Conference on Machine Learning*, 2019.
- 447 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.
448 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.
449 *Nature Communications*, 10(4933), 2019.
- 450 [30] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
451 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
452 2018.
- 453 [31] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
454 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 455 [32] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
456 A framework for parameter estimation and model selection from experimental data in systems
457 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 458 [33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Con-*
459 *ference on Learning Representations*, 2014.
- 460 [34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropaga-
461 tion and variational inference in deep latent gaussian models. *International Conference on*
462 *Machine Learning*, 2014.
- 463 [35] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
464 neural population models through nonlinear embeddings. In *Advances in neural information*
465 *processing systems*, pages 163–171, 2016.

- 466 [36] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-
467 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 468 [37] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-
469 encoders. *bioRxiv*, page 399246, 2018.
- 470 [38] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
471 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R
472 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-
473 encoders. *Nature methods*, page 1, 2018.
- 474 [39] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
475 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
476 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 477 [40] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
478 Datta. Composing graphical models with neural networks for structured representations and
479 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 480 [41] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
481 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
482 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural
483 Information Processing Systems*, 2019.
- 484 [42] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-
485 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 486 [43] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
487 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
488 2003.
- 489 [44] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
490 1970.
- 491 [45] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
492 Edward Teller. Equation of state calculations by fast computing machines. *The journal of
493 chemical physics*, 21(6):1087–1092, 1953.

- 494 [46] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
495 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 496 [47] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
497 *International Conference on Machine Learning*, 2015.
- 498 [48] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R
499 Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics,
500 biology, and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.
- 501 [49] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
502 free variational inference. In *Advances in Neural Information Processing Systems*, pages
503 5523–5533, 2017.
- 504 [50] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
505 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 506 [51] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for
507 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,
508 2017.
- 509 [52] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
510 networks. *International Conference on Learning Representations*, 2017.
- 511 [53] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-
512 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,
513 2001.
- 514 [54] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and
515 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysi-
516 ology*, 115(3):1399–1409, 2016.
- 517 [55] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
518 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
519 routing. *bioRxiv*, page 245613, 2018.
- 520 [56] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
521 tion. *Neural Networks*, 15(4-6):479–493, 2002.

- 522 [57] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
523 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
524 *Neuron*, 77(5):845–858, 2013.
- 525 [58] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
526 1957.
- 527 [59] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
528 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 529 [60] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
530 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 531 [61] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
532 2009.
- 533 [62] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
534 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 535 [63] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
536 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
537 592, 2009.
- 538 [64] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear-
539 ear network: a unifying circuit motif underlying multi-input integration in sensory cortex.
540 *Neuron*, 85(2):402–417, 2015.
- 541 [65] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
542 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
543 5(10):793, 2004.
- 544 [66] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
545 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-*
546 *robiology*, 71(1):45–61, 2011.
- 547 [67] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-
548 tex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.

- 549 [68] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
550 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
551 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 552 [69] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing
553 Wang. Paradoxical response reversal of top- down modulation in cortical circuits with three
554 interneuron types. *Elife*, 6:1–15, 2017.
- 555 [70] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying
556 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 557 [71] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
558 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
559 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 560 [72] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
561 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
562 operating regime of cortex. *bioRxiv*, 2020.
- 563 [73] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and
564 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,
565 2015.
- 566 [74] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give
567 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of
568 Sciences*, 115(13):3464–3469, 2018.
- 569 [75] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emerg-
570 ent dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint
571 arXiv:1605.04463*, 2016.
- 572 [76] Xaq Pitkow and Dora E Angelaki. Inference in the brain: statistics flowing in redundant
573 population codes. *Neuron*, 94(5):943–953, 2017.
- 574 [77] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-
575 like dynamics in recurrent circuits optimized for sampling-based probabilistic inference.
576 *bioRxiv*, page 696088, 2019.

- 577 [78] Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and compu-
578 tations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- 579 [79] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,
580 2019.
- 581 [80] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical
582 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.
- 583 [81] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
584 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
585 73(2):123–214, 2011.
- 586 [82] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-
587 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
588 820, 2011.
- 589 [83] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based infer-
590 ence. *Proceedings of the National Academy of Sciences*, 2020.
- 591 [84] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
592 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience
593 Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 594 [85] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
595 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience
596 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 597 [86] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
598 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
599 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 600 [87] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-
601 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,
602 Germany*, 2019.
- 603 [88] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
604 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P

- 605 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
606 dynamics. *bioRxiv*, page 838383, 2019.
- 607 [89] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
608 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
609 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 610 [90] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
611 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on*
612 *Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 613 [91] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
614 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
615 4248. PMLR, 2020.
- 616 [92] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
617 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 618 [93] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
619 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 620 [94] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
621 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
622 2012.
- 623 [95] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
624 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
625 2018.
- 626 [96] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
627 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 628 [97] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
629 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
630 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
631 1929, 2009.
- 632 [98] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
633 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.

- 634 [99] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angò. Parameter identifiability of
635 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 636 [100] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-
637 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv*
638 *preprint arXiv:1912.02762*, 2019.
- 639 [101] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
640 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 641 [102] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
642 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-*
643 *tion processing systems*, 29:4743–4751, 2016.
- 644 [103] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Internation-
645 al Conference on Learning Representations*, 2015.
- 646 [104] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for
647 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

648 **5 Methods**

649 **5.1 Emergent property inference (EPI)**

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

659 Consider model parameterization \mathbf{z} , which is a collection of scientifically interesting variables that
 660 govern the complex simulation of data \mathbf{x} . For example (see Section 3.1), \mathbf{z} may be the electrical
 661 conductance parameters of an STG subcircuit, and \mathbf{x} the evolving membrane potentials of the five
 662 neurons. In terms of statistical modeling, this circuit model has an intractable likelihood $p(\mathbf{x} | \mathbf{z})$,
 663 which is predicated by the stochastic differential equations that define the model. Even so, we do
 664 not scientifically reason about how \mathbf{z} governs all of \mathbf{x} , but rather specific phenomena that are a
 665 function of the data $f(\mathbf{x}; \mathbf{z})$. In the STG example, $f(\mathbf{x}; \mathbf{z})$ measures hub neuron frequency from the
 666 evolution of \mathbf{x} governed by \mathbf{z} . With EPI, we learn distributions of \mathbf{z} that results in an average and
 667 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
 668 emergent property. Such emergent properties \mathcal{X} are defined through choice of $f(\mathbf{x}; \mathbf{z})$ (which may
 669 be one or multiple statistics), $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}^2$

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (15)$$

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

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

$$\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0) = g_l(\dots g_1(\mathbf{z}_0)) \sim q_{\boldsymbol{\theta}}(\mathbf{z}). \quad (16)$$

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

679 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
680 the neural network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep
681 variational family \mathcal{Q} producing the emergent property \mathcal{X} :

$$q_{\theta}(\mathbf{z} \mid \mathcal{X}) = q_{\theta}^*(\mathbf{z}) = \operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}(\mathbf{z})) \quad (17)$$

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

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

688 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.
689 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-
690 ond, we describe the special class of deep probability distributions used in EPI called normalizing
691 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-
692 tion 17 (Section 5.1.3). Then, we demonstrate the details of this optimization in a toy example
693 (Section 5.1.4). Finally, we establish the known relationship between maximum entropy distribu-
694 tions and exponential families (Section 5.1.5), which is used to explain the relation between EPI
695 and variational inference (Section 5.1.6).

696 5.1.1 Related approaches

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

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

707 Likelihood-free (or “simulation-based”) inference (LFI) [83] is model parameter inference in the
708 absence of a tractable likelihood function. The most prevalent approach to LFI is approximate
709 Bayesian computation [42], in which satisfactory parameter samples are kept from random prior
710 sampling according to a rejection heuristic. The obtained set of parameters do not have a prob-
711 abilities, and further insight about the model must be gained from examination of the parameter
712 set and their generated activity. Methodological advances to ABC methods have come through
713 the use of Markov chain Monte Carlo (MCMC-ABC) [43] and sequential Monte Carlo (SMC-ABC)
714 [31] sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still
715 struggles to scale in dimensionality (cf. Fig. 4). Furthermore, once a parameter set has been
716 obtained by SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again
717 with a new population of initialized particles to obtain additional samples.

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

729 We note that, during our preparation and early presentation of this work [84, 85], another work
730 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
731 circuits ([86, 87, 88]). We are encouraged by this general problem being recognized by others in the
732 community, and we emphasize that these works offer complementary neuroscientific contributions
733 (different theoretical models of focus) and use different technical methodologies (ours is built on
734 our prior work [52], theirs similarly [89]).

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

736 recently developed LFI methods in which two neural networks are used for posterior inference.
737 This first neural network is a normalizing flow used to estimate the posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE)
738 or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural likelihood (SNL [90])). A recent advance uses an
739 unconstrained neural network to estimate the likelihood ratio (sequential neural ratio estimation
740 (SNRE [91])). In SNL and SNRE, MCMC sampling techniques are used to obtain samples from
741 the approximated posterior. This contrasts with EPI and SNPE, which afford a normalizing flow
742 approximation to the posterior, which facilitates immediate measurements of sample probability,
743 gradient, or Hessian for system analysis. The second neural network in this sequential class of
744 methods is the amortizer. This network maps data \mathbf{x} (or statistics $f(\mathbf{x}; \mathbf{z})$ or model parameters \mathbf{z}
745 to the weights and biases of the first neural network. These methods are optimized on a conditional
746 density (or ratio) estimation objective on a sequentially adapting finite sample-based approximation
747 to the posterior.

748 The approximating fidelity of the first neural network in sequential approaches is optimized to
749 generalize across the entire distribution it is conditioned upon. This optimization towards gen-
750 eralization of sequential methods can reduce the accuracy at the singular posterior of interest.
751 Whereas in EPI, the entire expressivity of the normalizing flow is dedicated to learning a single
752 distribution as well as possible. While amortization is not possible in EPI parameterized by the
753 mean parameter μ (due to the inverse mapping problem [92]), we have shown this two-network
754 amortization approach to be effective in exponential family distributions defined by their natural
755 parameterization [93].

756 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in
757 natural models. Around a point, one can measure the Jacobian. One approach that scales well is
758 EAR [94]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [95] and
759 its stochastic adaptation [96]. Casting identifiability as a statistical estimation problem, the profile
760 likelihood can assess via iterated optimization while holding parameters fixed [97]. An exciting
761 recent method is capable of recovering the functional form of such unidentifiabilities away from a
762 point by following degenerate dimensions of the fisher information matrix [98]. Global structural
763 non-identifiabilities can be found for models with polynomial or rational dynamics equations using
764 DAISY [99]. With EPI, we have all the benefits given by a statistical inference method plus the
765 ability to query the gradient or Hessian of the inferred distribution at any chosen parameter value.

766 **5.1.2 Normalizing flows**

767 Deep probability distributions are comprised of multiple layers of fully connected neural networks
 768 (Equation). When each neural network layer is restricted to be a bijective function, the sample
 769 density can be calculated using the change of variables formula at each layer of the network. For
 770 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

$$p(\mathbf{z}_i) = p(g_i^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial g_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} \right| = p(\mathbf{z}_{i-1}) \left| \det \frac{\partial g_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right|^{-1}. \quad (18)$$

771 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 772 restricting our layers to normalizing flows [47, 100] – bijective functions with fast log determinant
 773 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log
 774 probability calculation confers efficient optimization of the maximum entropy objective (see Section
 775 5.1.3). We use the Real NVP [50] normalizing flow class, because its coupling architecture confers
 776 both fast sampling (forward) and fast log probability evaluation (backward). Fast probability
 777 evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout
 778 parameter space. Glow permutations were used in between coupling stages [101]. This is in contrast
 779 to autoregressive architectures [51, 102], in which only forward or backward passes are efficient. In
 780 this work, normalizing flows are used as flexible posterior approximations $q_{\boldsymbol{\theta}}(\mathbf{z})$ having weights and
 781 biases $\boldsymbol{\theta}$. We specify the architecture used in each application by the number of Real-NVP affine
 782 coupling stages, and the number of neural network layers and units per layer of the conditioning
 783 functions.

784 **5.1.3 Augmented Lagrangian optimization**

785 To optimize $q_{\boldsymbol{\theta}}(\mathbf{z})$ in Equation 17, the constrained maximum entropy optimization is executed using
 786 the augmented Lagrangian method. The following objective is minimized:

$$L(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{opt}}, c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top R(\boldsymbol{\theta}) + \frac{c}{2} \|R(\boldsymbol{\theta})\|^2 \quad (19)$$

787 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
 788 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.
 789 The sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter $\boldsymbol{\mu}_{\text{opt}}$ are determined by the means $\boldsymbol{\mu}$ and
 790 variances $\boldsymbol{\sigma}^2$ of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 17. Specifically, $T(\mathbf{x}; \mathbf{z})$ is
 791 a concatenation of the first and second moments, $\boldsymbol{\mu}_{\text{opt}}$ is a concatenation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ (see section
 792 5.1.5), and the Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of exponential

793 families (see Section 5.1.6). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution are optimized
794 according to Equation 19 using the Adam optimizer with learning rate 10^{-3} [103].

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

$$H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \int -q_{\boldsymbol{\theta}}(\mathbf{z}) \log(q_{\boldsymbol{\theta}}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [-\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (20)$$

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

$$\nabla_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\nabla_{\boldsymbol{\theta}} \log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (21)$$

800 The lagrangian parameters $\boldsymbol{\eta}_{\text{opt}}$ are initialized to zero and adapted following each augmented
801 Lagrangian epoch, which is a period of optimization with fixed $(\boldsymbol{\eta}_{\text{opt}}, c)$ for a given number of
802 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
803 after each epoch based on constraint error reduction. The penalty coefficient is updated based
804 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
805 $\mathbb{E}[|R(\boldsymbol{\theta}_{k+1})|] > \gamma \mathbb{E}[|R(\boldsymbol{\theta}_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
806 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
807 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
808 according to application.

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

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

822 To assess whether the EPI distribution $q_{\theta}(\mathbf{z})$ produces the emergent property, we assess whether
 823 each individual constraint on the means and variances of $f(\mathbf{x}; \mathbf{z})$ is satisfied. We consider the EPI
 824 to have converged when a null hypothesis test of constraint violations $R(\boldsymbol{\theta})_i$ being zero is accepted
 825 for all constraints $i \in \{1, \dots, m\}$ at a significance threshold $\alpha = 0.05$. This significance threshold is
 826 adjusted through Bonferroni correction according to the number of constraints m . The p-values for
 827 each constraint are calculated according to a two-tailed nonparametric test, where 200 estimations
 828 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
 829 Lagrangian epoch.

830 When assessing the suitability of EPI for a particular modeling question, there are some important
 831 technical considerations. First and foremost, as in any optimization problem, the defined emergent
 832 property should always be appropriately conditioned (constraints should not have wildly different
 833 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution
 834 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-
 835 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,
 836 one should consider the computational cost of the gradient calculations. In the best circumstance,
 837 there is a simple, closed form expression (e.g. Section 5.2.4) for the emergent property statistic
 838 given the model parameters. On the other end of the spectrum, many forward simulation iterations
 839 may be required before a high quality measurement of the emergent property statistic is available
 840 (e.g. Section 5.2.1). In such cases, backpropagating gradients through the SDE evolution will be
 841 expensive.

842 5.1.4 Example: 2D LDS

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

$$845 \quad \tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (22)$$

with

$$846 \quad A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (23)$$

847 To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fix-
 848 ing $\tau = 1$), the emergent property statistics $T(\mathbf{x})$ were chosen to contain the first and second
 849 moments of the oscillatory frequency, $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscil-
 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 $(0.1\text{Hz})^2$:

$$\mathbb{E}[T(\mathbf{x})] \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}. \quad (24)$$

854

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:

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (25)$$

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 [92]. 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 probability 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 .

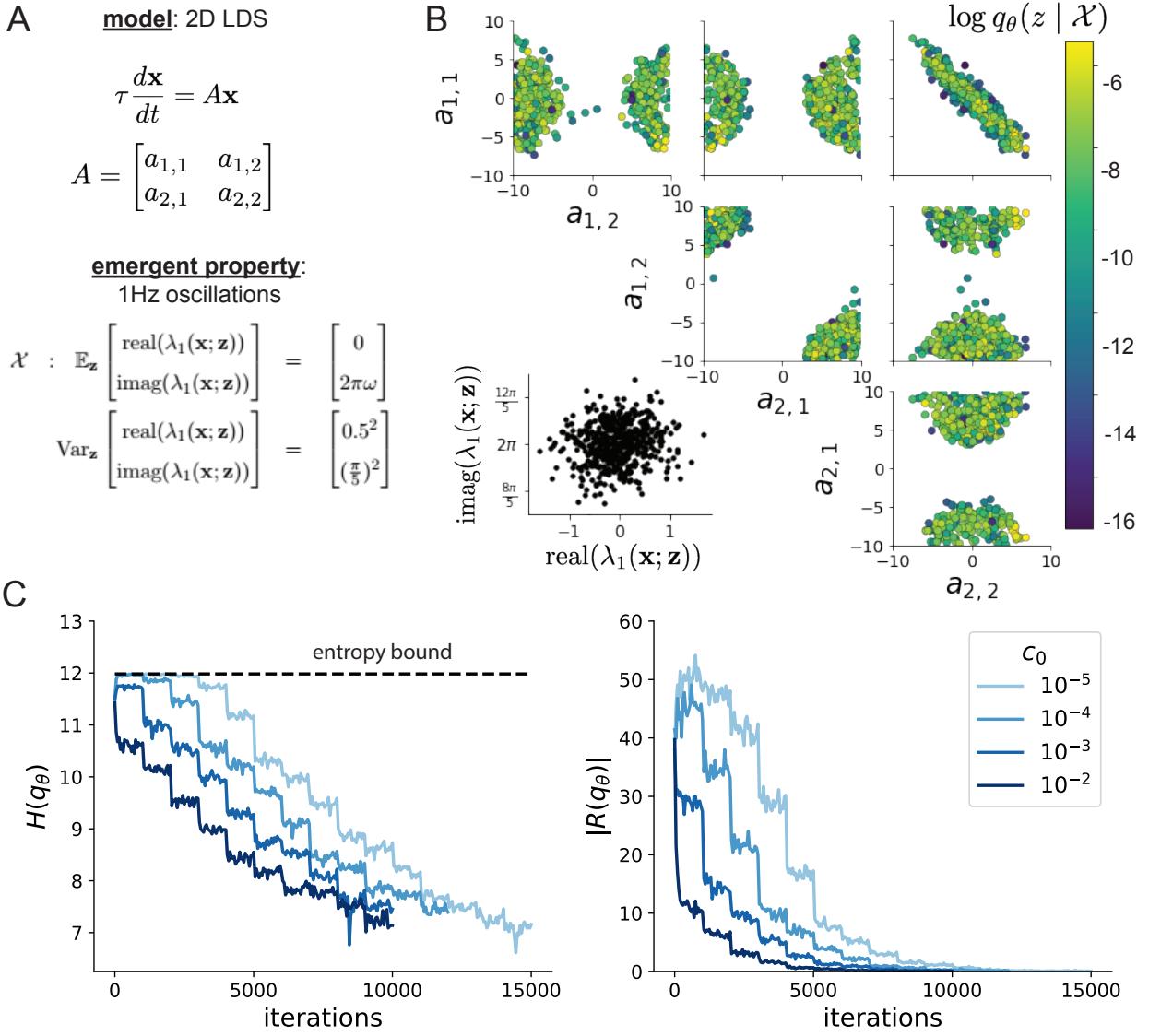


Figure 5: (LDS1): A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed 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.

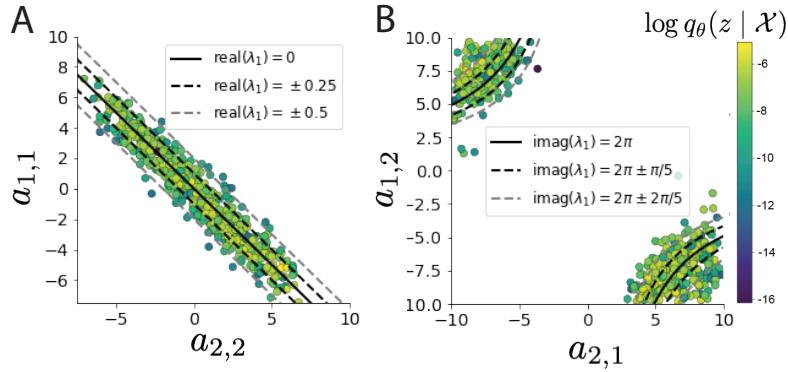


Figure 6: (LDS2): A. Probability contours in the a_1 - a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. B. Probability contours in the a_2 - a_3 plane were derived from the emergent property statistic of oscillation frequency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

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4 - a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (26)$$

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

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \quad (27)$$

877 Since we constrained $\mathbb{E}_{z \sim q_\theta} [\text{imag}(\lambda)] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation
878 $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{z \sim q_\theta} (|a_1a_4 - E_{q_\theta}[a_1a_4]|)$,
879 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
880 dotted). This validates the curved structure of the inferred distribution learned through EPI. We
881 took steps in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive
882 values a_1a_4 in the learned distribution. Subtler combinations of model and emergent property will
883 have more complexity, further motivating the use of EPI for understanding these systems. As we
884 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz
885 (Fig. S3).

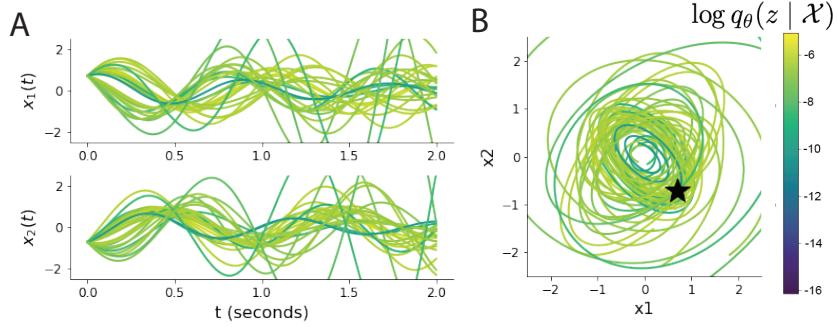


Figure 7: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\theta}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. A. Each dimension of the simulated trajectories throughout time. B The simulated trajectories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

886 5.1.5 Maximum entropy distributions and exponential families

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

$$p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \quad (28)$$

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

889 will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \quad (29)$$

890 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
891 formally hard to identify [92].

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

893 2

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (30)$$

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

$$\operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \mathbb{E}_{\mathbf{z}, \mathbf{x}}[(f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2] \quad (31)$$

895 As a general maximum entropy distribution (Equation 28), the sufficient statistics vector contains
896 both first and second order moments of $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (32)$$

897 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (33)$$

898 5.1.6 EPI as variational inference

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

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \quad (34)$$

903 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
904 However, conjugacy is rare in practice, and alternative methods, such as variational inference [104],
905 are utilized.

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

$$q_{\boldsymbol{\theta}}^*(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})). \quad (35)$$

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

$$KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})) = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] - \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{z} | \mathbf{x}))] \quad (36)$$

$$= -H(q_{\boldsymbol{\theta}}) - \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z})) - \log(p(\mathbf{x}))] \quad (37)$$

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

$$\underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}} || p(\mathbf{z} | \mathbf{x})) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmax}} H(q_{\boldsymbol{\theta}}) + \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z}))]. \quad (38)$$

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

$$p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) | \boldsymbol{\mu}_f, \Sigma_f), \quad (39)$$

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

$$\log p(\mathbf{x} | \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \quad (40)$$

917 where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\mu_f}{\sigma_f^2} \\ \frac{-1}{2\sigma_f^2} \end{bmatrix}, \text{ and} \quad (41)$$

918

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \mu_f)^2 \end{bmatrix}. \quad (42)$$

919 The variational objective is then

$$\operatorname{argmax}_{q_{\theta} \in Q} H(q_{\theta}) + \boldsymbol{\eta}_f^\top \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [T(\mathbf{x}; \mathbf{z})] \quad (43)$$

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

$$\begin{aligned} q_{\theta}^*(\mathbf{z}) &= \operatorname{argmin}_{q_{\theta} \in Q} -H(q_{\theta}) + \boldsymbol{\eta}_{\text{opt}}^\top (\mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}_{\text{opt}}) \\ &= \operatorname{argmin}_{q_{\theta} \in Q} -H(q_{\theta}) + \boldsymbol{\eta}_{\text{opt}}^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})]. \end{aligned} \quad (44)$$

922 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
923 (Equation 32). Thus, EPI is implicitly executing variational inference with a uniform prior and a
924 mean-field gaussian likelihood on the emergent property statistics. The data \mathbf{x} used by this implicit
925 variational inference program would be that generated by the adapting variational approximation
926 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{z})q_{\theta}(\mathbf{z})$, and the likelihood parameters $\boldsymbol{\eta}_f$ of EPI optimization epoch k are predicated
927 by $\boldsymbol{\eta}_{\text{opt},k}$. However, in EPI we have not specified a prior distribution, or collected data, which can
928 inform us about model parameters. Instead we have a mathematical specification of an emergent
929 property, which the model must produce, and a maximum entropy selection principle. Accordingly,
930 we replace the notation of $p(\mathbf{z} | \mathbf{x})$ with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys
931 emergent property \mathcal{X} (see Section 5.1).

932 5.2 Theoretical models

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

935 5.2.1 Stomatogastric ganglion

936 We analyze how the parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$ govern the emergent phenomena of intermediate
937 hub frequency in a model of the stomatogastric ganglion (STG) [57] shown in Figure 1A with

938 activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{\text{hub}}, x_{s1}, x_{s2}]$, using the same hyperparameter choices as Gutierrez et al.
 939 Each neuron's membrane potential $x_\alpha(t)$ for $\alpha \in \{f1, f2, \text{hub}, s1, s2\}$ is the solution of the following
 940 stochastic differential equation:

$$C_m \frac{dx_\alpha}{dt} = -[h_{\text{leak}}(\mathbf{x}; \mathbf{z}) + h_{Ca}(\mathbf{x}; \mathbf{z}) + h_K(\mathbf{x}; \mathbf{z}) + h_{hyp}(\mathbf{x}; \mathbf{z}) + h_{elec}(\mathbf{x}; \mathbf{z}) + h_{syn}(\mathbf{x}; \mathbf{z})] + dB. \quad (45)$$

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

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

$$h_{\text{leak}}(\mathbf{x}; \mathbf{z}) = g_{\text{leak}}(x_\alpha - V_{\text{leak}}) \quad (46)$$

$$h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{\text{post}} - x_\alpha^{\text{pre}}) \quad (47)$$

$$h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{\text{pre}}(x_\alpha^{\text{post}} - V_{\text{syn}}) \quad (48)$$

$$h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (49)$$

$$h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (50)$$

$$h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (51)$$

952 The reversal potentials were set to $V_{\text{leak}} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 953 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{\text{leak}} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 954 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 955 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
 956 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
 957 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

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

$$M_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_1}{v_2} \right) \right) \quad (52)$$

$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (53)$$

$$N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_3}{v_4} \right) \right) \quad (54)$$

962

$$\lambda_N = \phi_N \cosh \left(\frac{x_\alpha - v_3}{2v_4} \right) \quad (55)$$

963

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (56)$$

964

$$H_\infty = \frac{1}{1 + \exp \left(\frac{x_\alpha + v_5}{v_6} \right)} \quad (57)$$

965

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp \left(\frac{-x_\alpha + v_7}{v_8} \right)} \right). \quad (58)$$

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 $v_{th} = -25mV$.

Finally, there is a synaptic gating variable as well:

$$S_\infty = \frac{1}{1 + \exp \left(\frac{v_{th} - x_\alpha}{v_9} \right)}. \quad (59)$$

When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear dynamical system. Gaussian noise of variance $(1 \times 10^{-12})^2$ amps 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 $dt = 25ms$. 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

$$\Phi = [0.0, 0.01, \dots, 1.0]^\top .. \quad (60)$$

To measure spiking frequency, we processed simulated membrane potentials with a relu (spike extraction) and low-pass filter with averaging window of size 20, then took the frequency with the maximum absolute value of the complex exponential basis coefficients of the processed time-series. 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 $\alpha \in \{f1, f2, \text{hub}, s1, s2\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta |X_\alpha| \odot i), \quad (61)$$

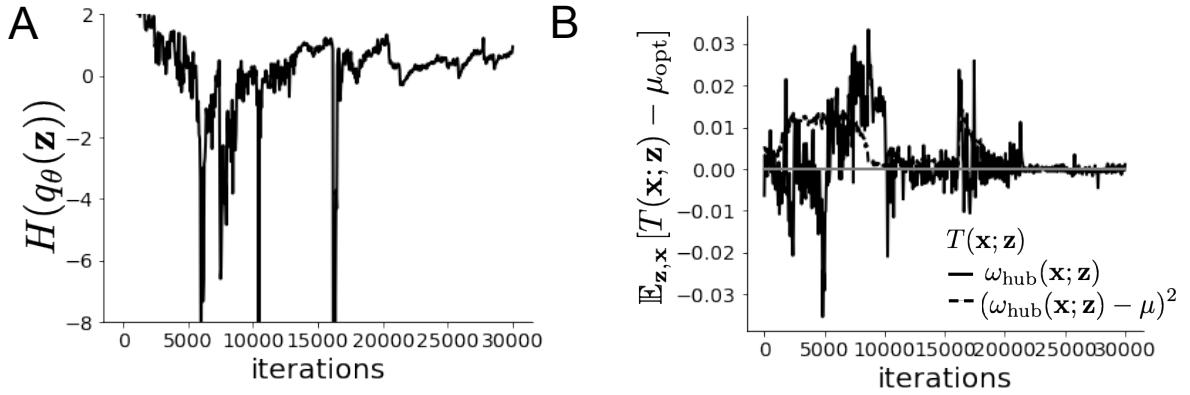


Figure 8: (STG1): EPI optimization of the STG model producing network syncing. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

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

$$\omega_\alpha = 0.01\psi_\alpha \text{Hz}. \quad (62)$$

984 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
 985 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
 986 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be $(0.025\text{Hz})^2$ to
 987 capture variation in frequency between 0.5Hz and 0.6Hz (Equation 2). As a maximum entropy dis-
 988 tribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments of the hub neuron frequency
 989 (as in Equations 32 and 33)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \quad (63)$$

$$\mu_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \quad (64)$$

990
 991 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 992 epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

993 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and
 994 two-layer neural networks of 25 units per layer. The initial distribution was a standard isotropic
 995 gaussian $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to a support of $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}] \in [4, 8] \times [0.01, 4]$. We did not include

996 $g_{\text{synA}} < 0.01$, since conductances that low make the circuit simulations numerically unstable. We
 997 used an augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, $\beta = 2$, $N_{\text{test}} = 100$,
 998 and initialized $q_{\theta}(\mathbf{z})$ to produce a gaussian approximation to samples returned from an initial ABC
 999 search. This initialization had much greater entropy and a different emergent property than the
 1000 returned EPI posterior.

1001 TODO write about specifics of the Hessian analysis.

1002 **5.2.2 Primary visual cortex**

1003 Connectivity (W_{fit}) and input ($\mathbf{h}_{b,\text{fit}}$ and $\mathbf{h}_{c,\text{fit}}$) parameters were fit using the deterministic V1 circuit
 1004 model [72]

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

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

1005 and

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

1006 To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equiv-
 1007 alence class

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

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

1008 and

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

1009 Since the E-population of this network increases exponentially in the absense of recurrent inhibitory
 1010 feedback, we may also observe a paradoxical effect in the inhibitory populations (which is present
 1011 in E-I networks). At 50% contrast (Fig. 2B, dots), this network exhibits a paradoxical effect in
 1012 the P-population (Fig. 2C), but no others (Fig. 9). That is, for a small increase in h_P , $\mathbb{E}_t [x_P]$
 1013 decreases.

1014 Fano factor is calculated as the temporal variance divided by the temporal mean following sometime
 1015 t_{ss} following dynamical evolution from the initial state at $\mathbf{x}(t = 0)$.

1016 5.2.3 Superior colliculus

1017 In the model of Duan et al [55], there are four total units: two in each hemisphere corresponding to
 1018 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right
 1019 Pro (RP) and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related
 1020 by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh \left(\frac{u_\alpha - a}{b} \right) + \frac{1}{2} \right) \quad (71)$$

1021 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-
 1022 earity, respectively. During periods of optogenetic inactivation, activity was decreased proportional

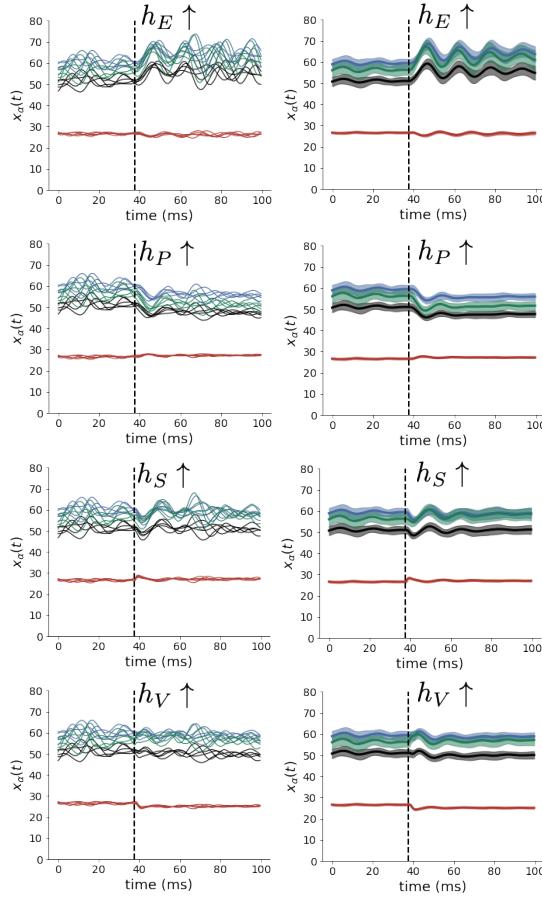


Figure 9: Supplemental Figure: (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average and standard deviation of stochastic fluctuations of responses.

1023 to the optogenetic strength γ

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (72)$$

1024 We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (73)$$

1025 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (74)$$

1026 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2. The weight

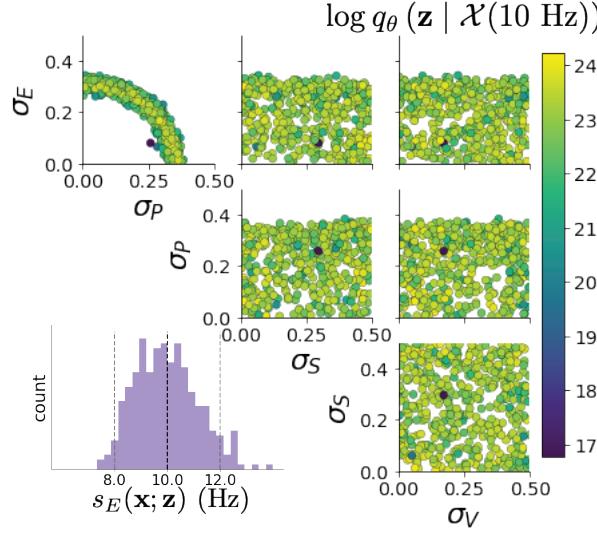


Figure 10: Supplemental Figure: Fano factors along the ridge of the posterior in Fig. 2E.

¹⁰²⁷ matrix has 4 parameters sW , vW , hW , and dW :

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (75)$$

¹⁰²⁸ The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (76)$$

¹⁰²⁹ There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (77)$$

¹⁰³⁰ a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (78)$$

¹⁰³¹ rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (79)$$

¹⁰³²

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (80)$$

1033 a choice-period input

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

1034 and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases}. \quad (82)$$

1035 The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{P,\text{bias}} = 0.5$, $I_{P,\text{rule}} = 0.6$, $I_{A,\text{rule}} = 0.6$,
1036 $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

1037 The accuracies of p_P and p_A are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (83)$$

1038 and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (84)$$

1039 given that the stimulus is on the left side, where Θ is the Heaviside step function.

1040 The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (85)$$

1041 where $\beta = 100$.

1042 As a maximum entropy distribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments
1043 of the accuracy in each task (as in Equations 32 and 33)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p(\mathbf{x}; \mathbf{z})_P \\ p(\mathbf{x}; \mathbf{z})_A \\ (p(\mathbf{x}; \mathbf{z})_P - 75\%)^2 \\ (p(\mathbf{x}; \mathbf{z})_A - 75\%)^2 \end{bmatrix}, \quad (86)$$

1044

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 5\%^2 \\ 5\%^2 \end{bmatrix}. \quad (87)$$

1045 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
1046 epoch of 2,000 iterations(see Section 5.1.3). The optimization converged after six epochs (Fig. 15).
1047 For EPI in Fig. 3C, we used a real NVP architecture with three coupling layers of affine transforma-
1048 tions parameterized by two-layer neural networks of 50 units per layer. The initial distribution was
1049 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 aug-
1050 mented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, set $\nu = 0.5$, and initialized $q_{\theta}(\mathbf{z})$
1051 to produce an isotropic gaussian with mean 0 and variance 2.5^2 . Accuracies were estimated over
1052 200 trials of random gaussian noise, which was sampled independently for each drawn parameter \mathbf{z}
1053 and each iteration of the EPI optimization.

1054 **5.2.4 Rank-2 RNN**

1055 Traditional approaches to likelihood-free inference – approximate Bayesian computation (ABC)
1056 methods – randomly sample parameters \mathbf{z} until a suitable set is obtained. State-of-the-art ABC
1057 methods leverage sequential Monte Carlo (SMC) sampling techniques to obtain parameter sets more
1058 efficiently. To obtain more parameter samples, SMC-ABC must be run from scratch again. ABC
1059 methods do not confer log probabilities of samples. Like EPI, sequential neural posterior estimation
1060 (SNPE) uses deep learning to produce flexible posterior approximations. Like traditional Bayesian
1061 inference methods, SNPE conditions directly on the statistics of data. This differs from EPI, where
1062 posteriors are conditioned on emergent properties (moment constraints on the posterior predictive
1063 distribution). Peculiarities of SNPE (density estimation approach, two deep networks) make scaling
1064 in \mathbf{z} prohibitive.

1065 SMC-ABC has many hyperparameters, of which pyABC selects automatically by running some ini-
1066 tial diagnostics upon initialization. In concurrence with the literature, SMC-ABC fails to converge
1067 around 25-30 dimensions, since it's proposal samples never get close enough to the target statis-
1068 tics. We searched over many SNPE hyperparameter choices: $n_{\text{train}} \in [2,000, 10,000, 100,000]$ is the
1069 number of simulations run per training epoch, and $n_{\text{mades}} \in [2, 3]$ is the number of masked autore-
1070 gressive density estimators in the deep parameter distribution architecture. The greater n_{train} , the
1071 longer each epoch will take, but the more likely SNPE may converge during that epoch. Greater
1072 n_{mades} increases the flexibility of the deep parameter distribution of SNPE, but slows optimization.
1073 For the timing plot, we show the fastest among all of these choices, and for the convergence plot,
1074 we show the best convergence among all of these choices. During optimization, we used $n_{\text{atom}}=100$
1075 atomic proposals as is recommended.

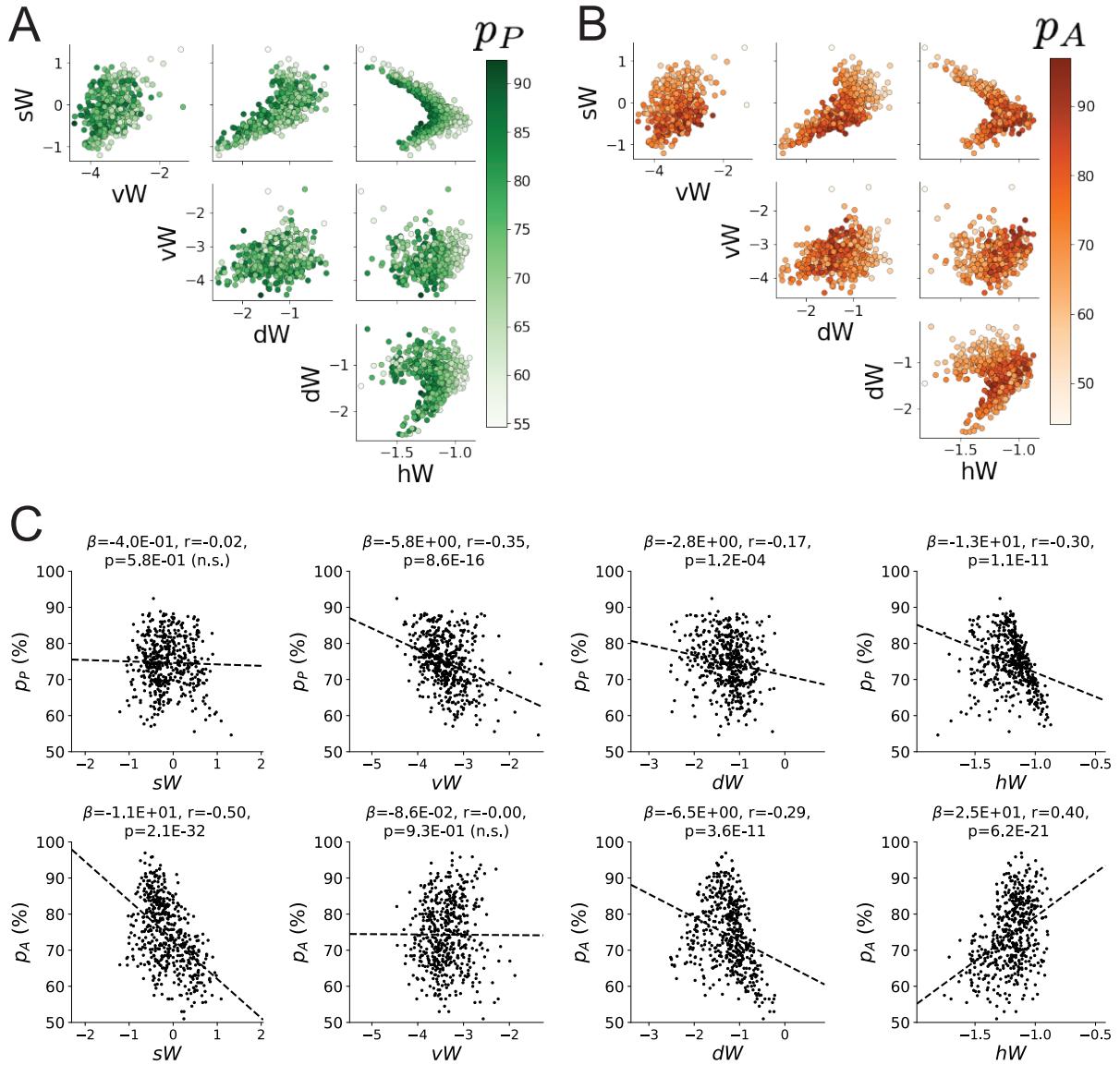


Figure 11: (SC1): Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p value.

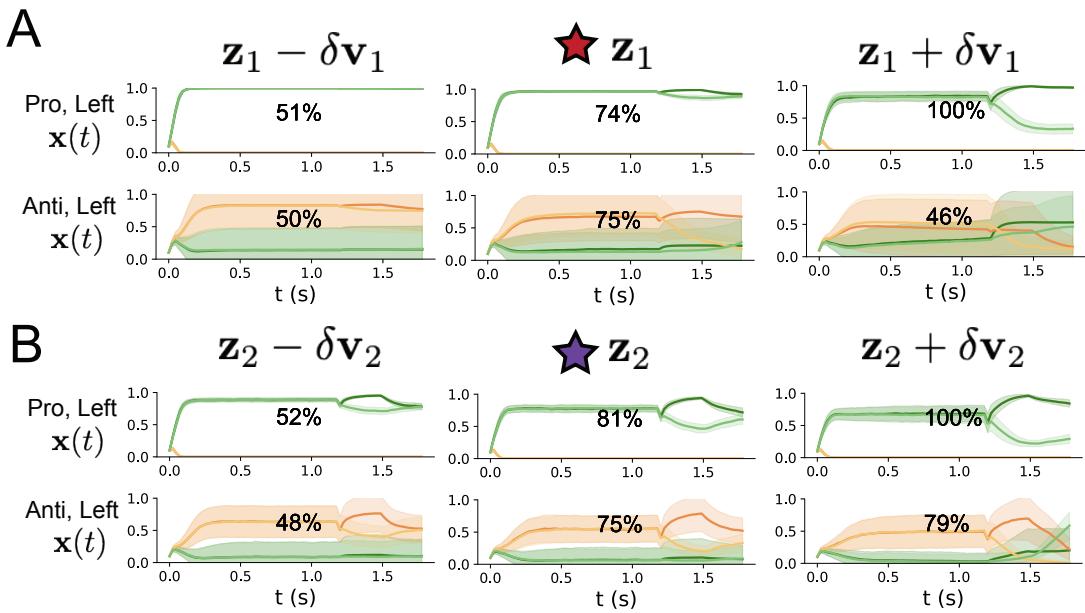


Figure 12: (SC2): A. Invariant eigenvectors of connectivity matrix W . B. Eigenvalues of connectivities of EPI distribution versus task accuracies.

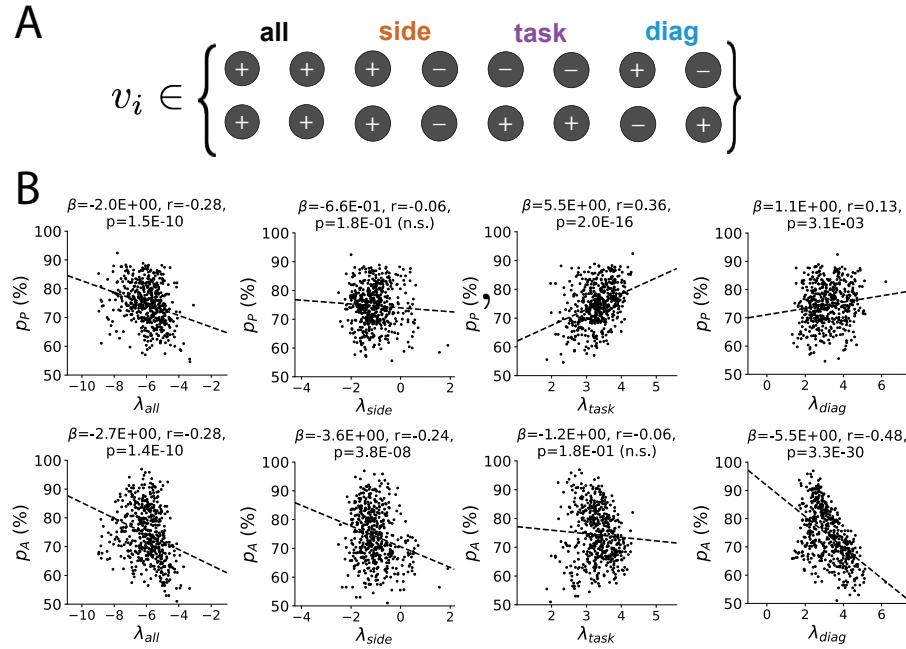


Figure 13: (SC3): A. Connectivity eigenvalues of EPI parameter distribution colored by Pro task accuracy.
B. Same for Anti task.

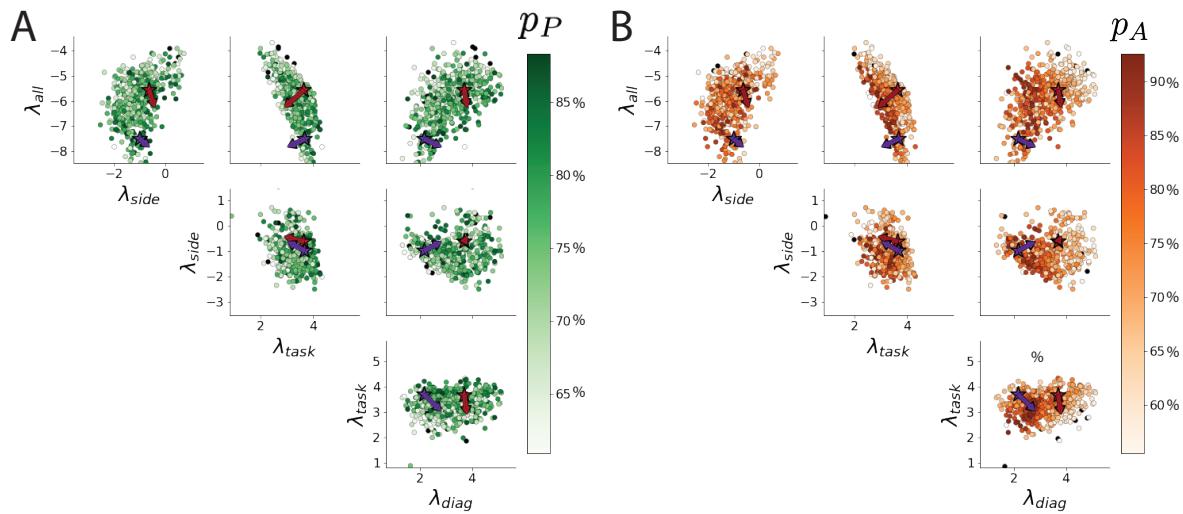


Figure 14: (SC4): Scatters of the effect of delay period inactivation in each task with task accuracy. Plots are shown at an opto strength of 0.8.

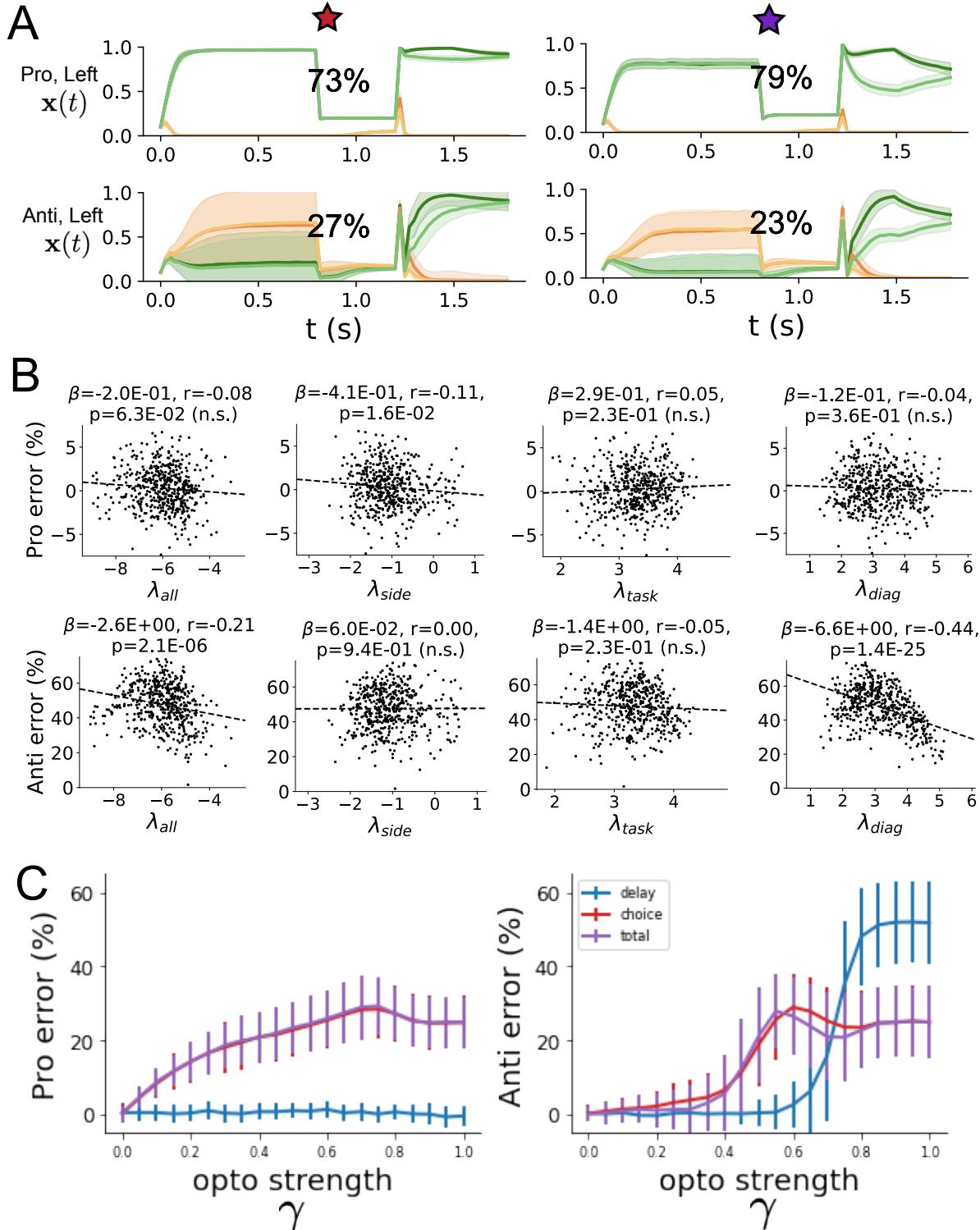


Figure 15: (SC5): EPI optimization of the SC model producing rapid task switching. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 12,000 iterations following the sixth augmented Lagrangian epoch.