

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
³ a hypothesized neural mechanism. Such models are valuable when they give rise to an experimen-
⁴ tally observed phenomenon – whether behavioral or in terms of neural activity – and thus can offer
⁵ insights into neural computation. The operation of these circuits, like all models, critically depends
⁶ on the choices of model parameters. When analytic derivation of the relationship between model pa-
⁷ rameters and computational properties is intractable, approximate inference and simulation-based
⁸ techniques are relied upon for scientific insight. We bring the use of deep generative models for
⁹ probabilistic inference to bear on this problem, learning distributions of parameters that produce
¹⁰ the specified properties of computation. By learning parameter distributions that produce compu-
¹¹ tations – an emergent property, we introduce a novel methodology that is particularly well-suited
¹² to the stochastic dynamical systems models predominant in our field of theoretical neuroscience.
¹³ We motivate this methodology with a worked example analyzing sensitivity in the stomatogastric
¹⁴ ganglion. We then use it to reveal the key factors of variability in a model of primary visual cortex,
¹⁵ gain a mechanistic understanding of rapid task switching in superior colliculus models, and scale
¹⁶ inference of large low-rank RNN’s exhibiting stable amplification. While much use of deep learning
¹⁷ in theoretical neuroscience focuses on drawing analogies between optimized neural architectures
¹⁸ and the brain, this work illustrates how we can further leverage the power of deep learning towards
¹⁹ solving inverse problems in theoretical neuroscience.

20 **2 Introduction**

21 The fundamental practice of theoretical neuroscience is to use a mathematical model to understand
22 neural computation, whether that computation enables perception, action, or some intermediate
23 processing [1]. A neural computation is systematized with a set of equations – the model – and
24 these equations are motivated by biophysics, neurophysiology, and other conceptual considerations.

25 The function of this system is governed by the choice of model parameters, which when configured
26 in a particular way, give rise to a measurable signature of a computation. The work of analyzing a
27 model then requires solving the inverse problem: given a computation of interest, how can we reason
28 about these particular parameter configurations? The inverse problem is crucial for reasoning about
29 likely parameter values, uniquenesses and degeneracies, and predictions made by the model.

30 Consider the idealized practice: one carefully designs a model and analytically derives how model
31 parameters govern the computation. Seminal examples of this gold standard (which often adopt
32 approaches from statistical physics) include our field’s understanding of memory capacity in asso-
33 ciative neural networks [2], chaos and autocorrelation timescales in random neural networks [3],
34 the paradoxical effect [4], and decision making [5]. Unfortunately, as circuit models include more
35 biological realism, theory via analytical derivation becomes intractable. Alternatively, statistical
36 inference can be run to obtain model parameters likely to produce some model output, and local
37 sensitivity analyses can be performed at inferred parameter values. Since most neural circuit mod-
38 els stipulate a noisy system of differential equations that can only be sampled or realized through
39 forward simulation, they lack the explicit likelihood central to the probabilistic modeling toolkit.
40 Therefore, the most popular approaches to the inverse problem have been likelihood-free methods
41 such as approximate Bayesian computation (ABC) [6], in which a set of reasonable parameters
42 estimates is obtained via simulation and rejection.

43 Of course, the challenge of doing inference in complex models has arisen in many scientific fields.
44 In response, the machine learning community has made remarkable progress in recent years, via
45 the use of deep neural networks as a powerful inference engine: a flexible function family that can
46 map observations back to probability distributions quantifying the likely parameter configurations.
47 One celebrated example of this approach from machine learning, of which we draw key inspiration
48 for this work, is the variational autoencoder (VAE) [7, 8], which uses a deep neural network to
49 induce an (approximate) posterior distribution on hidden variables in a latent variable model, given
50 data. Indeed, these tools have been used to great success in neuroscience as well, in particular for

51 interrogating parameters (sometimes treated as hidden states) in models of both cortical population
52 activity [9, 10, 11, 12] and animal behavior [13, 14, 15]. These works have used deep neural networks
53 to expand the expressivity and accuracy of statistical models of neural data [16].

54 Existing approaches to the inverse problem in theoretical neuroscience fall short in three key ways.
55 First, theoretical models of neural computation aim to reflect a complex biological reality, and
56 as a result, such models lack tractable likelihoods. Thus, standard approaches from statistical
57 inference are unavailable. The parameter sets obtained from likelihood-free ABC lack a formal-
58 ized link to Bayesian inference (except in the unrealistic 0-distance scenario), and lack parameter
59 probabilities. Second is the undesirable trade-off between the flexibility and tractability of the ap-
60 proximated posterior distribution. While sampling-based approaches like ABC and Markov chain
61 Monte Carlo (MCMC) can produce flexible posterior approximations, they must be run continu-
62 ally for increasing samples. While VAE approaches can result in tractable posterior sampling and
63 sensitivity measurements post-optimization, existing approaches have relied on simplified classes
64 of distributions, which restrict the flexibility of the posterior approximation. And third, one can
65 never assume what inferred model parameters may predict. This is well understood when con-
66 sidering Box’s loop and the role of posterior predictive checks in the development and critique of
67 scientific models [17, 18]. Uncertainty about the properties of inferred model predictions introduce
68 a conceptual degree of freedom to the inverse problem that may be unnecessary and undesirable
69 given the scientific motivation.

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 [19]. First, we use stochastic gradient techniques in the
77 spirit of likelihood-free variational inference [20] to enable inference in likelihood-free models of
78 neural computation. Second, we stipulate a bijective deep neural network that induces a flexible
79 family of probability distributions over model parameterizations with a probability density we can
80 calculate [21, 22, 23], which confers fast sampling and sensitivity measurements. Third, we quan-
81 tify the notion of emergent properties as a set of moment constraints on datasets generated by the
82 model. Thus, an emergent property is not a single data realization, but a phenomenon or a feature

83 of the model, which is ultimately the object of interest in theoretical neuroscience. Conditioning
84 on an emergent property requires a variant of deep probabilistic inference methods, which we have
85 previously introduced [24]. Taken together, emergent property inference (EPI) provides a method-
86 ology for inferring parameter configurations consistent with a particular emergent phenomena in
87 theoretical models. We use a classic example of parametric degeneracy in a biological system, the
88 stomatogastric ganglion [25], to motivate and clarify the technical details 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 abstract cognitive function), and network scale (from four to
93 hundreds of neurons). First, we use EPI to understand the characteristics of noise that govern
94 Fano factor in a stochastic four neuron-type model of primary visual cortex. Second, we discover
95 connectivity patterns in superior colliculus resilient to optogenetic perturbation by using EPI to
96 condition on rapid task switching. The novel scientific insights offered by EPI contextualize and
97 clarify the previous studies exploring these models [26, 27]. Third, we emphasize the methodological
98 advancement of EPI by inferring high-dimensional distributions of RNN connectivities exhibiting
99 stable amplification. These results point to the value of deep inference for the interrogation of
100 biologically relevant models.

101 3 Results

102 3.1 Motivating emergent property inference of theoretical models

103 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-
104 erty inference. First, one designs or chooses an existing model that, it is hypothesized, captures
105 the computation of interest. To ground this process in a well-known example, consider the stom-
106 atogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic
107 muscle activation patterns for digestion [28]. Despite full knowledge of STG connectivity and a
108 precise characterization of its rhythmic pattern generation, biophysical models of the STG have
109 complicated relationships between circuit parameters and neural activity [25, 29]. A subcircuit
110 model of the STG [30] is shown schematically in Figure 1A, and note that the behavior of this
111 model will be critically dependent on its parameterization – the choices of conductance parameters
112 $\mathbf{z} = [g_{el}, g_{synA}]$. Specifically, the two fast neurons ($f1$ and $f2$) mutually inhibit one another, and

113 oscillate at a faster frequency than the mutually inhibiting slow neurons (s_1 and s_2). The hub
114 neuron (hub) couples with either the fast or slow population or both.

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

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

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

137 3.2 A deep generative modeling approach to emergent property inference

138 Emergent property inference (EPI) systematizes the three-step procedure of the previous section.
139 First, we consider the model as a coupled set of differential equations [30]. In the running STG
140 example, the model activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{\text{hub}}, x_{s1}, x_{s2}]$ is the membrane potential for each neuron,
141 which evolves according to the biophysical conductance-based equation:

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

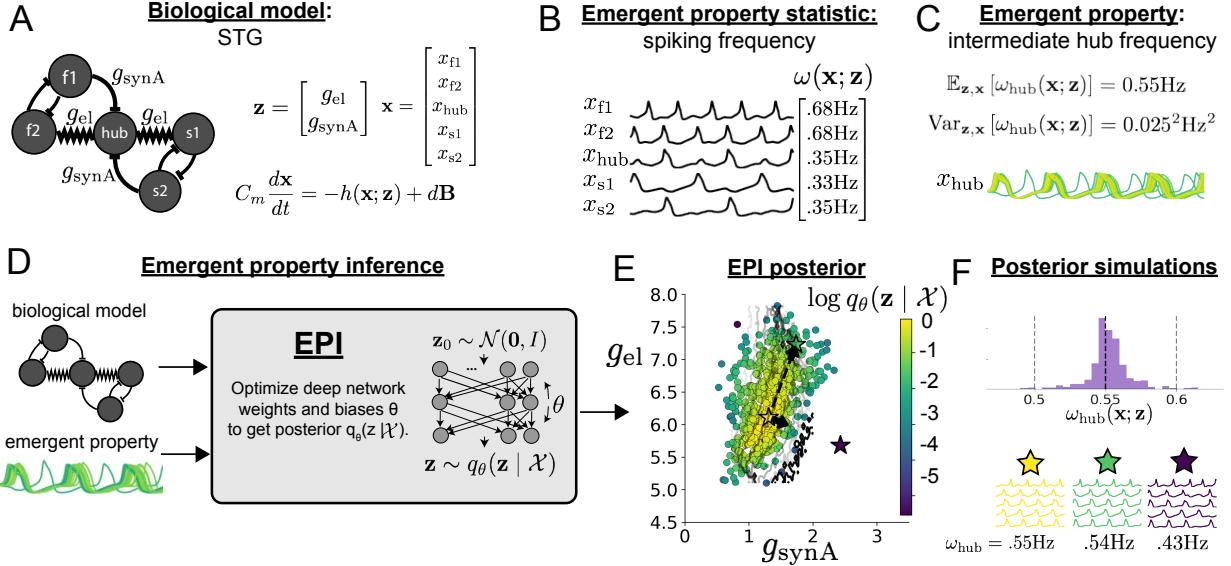


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based biophysical model of the STG subcircuit. In the STG model, jagged connections indicate electrical coupling having electrical conductance g_{el} . Other connections in the diagram are inhibitory synaptic 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 through a deep neural network with weights and biases $\boldsymbol{\theta}$ to parameters $\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0)$ 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.

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

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

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

149 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
 150 $\boldsymbol{\sigma}^2$ define the emergent property denoted \mathcal{X} .

151 Third, we perform emergent property inference: we find a distribution over parameter configura-
 152 tions \mathbf{z} , and insist that samples from this distribution produce the emergent property; in other
 153 words, they obey the constraints introduced in Equation 2. This distribution will be chosen from a
 154 family of probability distributions $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$, defined by a deep generative distribution
 155 of the normalizing flow class [21, 22, 23] – neural networks which transform a simple distribution
 156 into a suitably complicated distribution (as is needed here). This deep distribution is represented
 157 in Figure 1C (see Section 5.1). Then, mathematically, we must solve the following optimization
 158 program:

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

159 where $f(\mathbf{x}, \mathbf{z})$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ are defined as in Equation 10. According to the emergent property of
 160 interest, $f(\mathbf{x}, \mathbf{z})$ may contain multiple statistics, in which case the mean and variance vectors $\boldsymbol{\mu}$
 161 and $\boldsymbol{\sigma}^2$ match this dimension. Finally, we recognize that many distributions in \mathcal{Q} will respect
 162 the emergent property constraints, so we select that which has maximum entropy. This principle,
 163 captured in Equation 3 by the primal objective H , identifies parameter distributions with minimal
 164 assumptions beyond some chosen structure [31, 32, 24, 33]. Such a normative principle of maximum
 165 entropy, which is also that of Bayesian inference, naturally fits with our scientific objective of
 166 reasoning about parametric sensitivity and robustness. The recovered distribution of EPI is as
 167 variable as possible along each parametric manifold such that it produces the emergent property.

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

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

186 **3.3 EPI reveals how noise across neural population types governs Fano factor
187 in a stochastic inhibition stabilized network**

188 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output
189 function have succeeded in explaining a host of experimentally documented phenomena. In a regime
190 characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to
191 paradoxical responses [4], selective amplification [34, 35], surround suppression [36] and normal-
192 ization [37]. Despite their strong predictive power, E-I circuit models rely on the assumption that
193 inhibition can be studied as an indivisible unit. However, experimental evidence shows that inhibi-
194 tion is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) – composing
195 80% of GABAergic interneurons in V1 [38, 39, 40], and that these inhibitory cell types follow
196 specific connectivity patterns (Fig. 2A) [41]. Recent theoretical advances [26, 42, 43], have only
197 started to address the consequences of this multiplicity in the dynamics of V1, strongly relying on
198 linear theoretical tools. Here, we use EPI to characterize the properties of slow noise in a stochastic

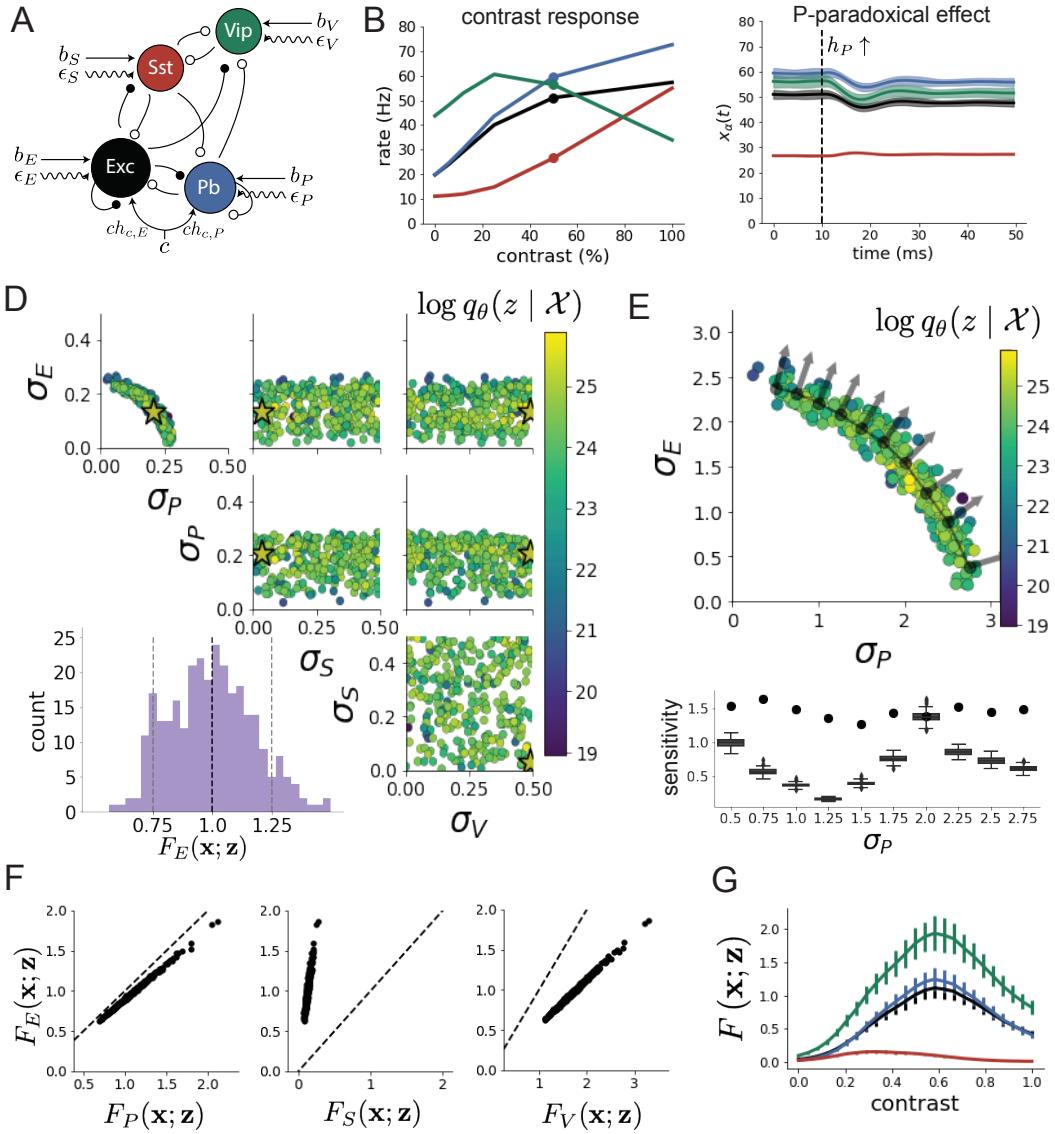


Figure 2: Emergent property inference of a stochastic stabilized supralinear network. **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). 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 smodel ($\epsilon = \mathbf{0}$) 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.

199 version of this model, which result in biologically realistic responses.

200 We considered the contrast response of a nonlinear dynamical V1 circuit model (Fig. 2A) with
 201 a state comprised of each neuron-type population's rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population
 202 receives recurrent input $W\mathbf{x}$ from synaptic projections of effective connectivity W and an external
 203 input \mathbf{h} , which determine the population rate via nonlinearity $\phi = \|\cdot\|_+^2$ (see Section 5.2.2). The
 204 circuit model evolves from an initial condition $\mathbf{x}(0) \sim \mathcal{U}([10, 25])$ with time constant $\tau = 1\text{ms}$
 205 according to a contrast-dependent input \mathbf{h} and slow noise ϵ of time constant $\tau_{\text{noise}} = 5\text{ms}$. This
 206 model is the stochastic stabilized supralinear network (SSSN) [44] generalized to have inhibitory
 207 multiplicity

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

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

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

210 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
 211 mean contrast responses in mice with the deterministic model [45] ($\epsilon = \mathbf{0}$, Fig. 2B, see Section
 212 5.2.2). At all contrasts, the E-population of this SSSN is unstable without recurrent inhibitory
 213 feedback. At 50% contrast, only the P-population exhibits the paradoxical effect (2C, Fig. 9), so
 214 the network is P-stabilized.

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

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

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

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

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

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

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

221 We ran EPI to obtain a posterior $q_{\theta}(\mathbf{z} | \mathcal{X})$, where each parameter \mathbf{z} produces biologically realistic
 222 levels of E-population variability (Fig. 2D). From the marginal distribution of σ_E and σ_P (Fig.

223 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
 224 posterior obtained through EPI offers exactly how this sensitivity changes along this ridge of the
 225 posterior (Fig. 2E). σ_S and σ_V are degenerate with respect to $F_E(\mathbf{x}; \mathbf{z})$ evidenced by the uniform
 226 distribution in those dimensions of the posterior (Fig. 2D, bottom-right). Together, this posterior
 227 indicates a parametric manifold of degeneracy with respect to Fano factor: the ridge visualized in
 228 the σ_E - σ_P marginal (Fig. 10) and the dimensions of σ_S and σ_V .

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

236 3.4 EPI identifies neural mechanisms of flexible task switching

237 In a rapid task switching experiment [46], rats were explicitly cued on each trial to either orient
 238 towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti
 239 (A) task (Fig. 3A). Neural recordings in the midbrain superior colliculus (SC) exhibited two
 240 populations of neurons that simultaneously represented both task context (Pro or Anti) and motor
 241 response (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons
 242 [27]. Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is
 243 a four-population dynamical system. We analyzed this model, where the neuron-type populations
 244 are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right
 245 (R)), their connectivity is parameterized geometrically (Fig. 3B). The input-output function of
 246 this model is chosen such that the population responses $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$ are bounded
 247 from 0 to 1 as a function ϕ of a dynamically evolving internal variable \mathbf{u} . The model responds to
 248 the side with greater Pro neuron activation; e.g. the reponse is left if $x_{LP} > x_{RP}$ at the end of the
 249 trial. The dynamics evolve with timescale $\tau = 0.09$ governed by connectivity weights W

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

250 with white noise of variance 0.2². The input \mathbf{h} is comprised of a cue-dependent input to the Pro
 251 or Anti populations, a stimulus orientation input to either the Left or Right populations, and
 252 a choice-period input to the entire network (see Section 5.2.3). Here, we use EPI to determine
 253 the changes in network connectivity $\mathbf{z} = [sW, vW, dW, hW]^\top$ resulting in execution of rapid task
 254 switching behavior.

255 We define rapid task switching behavior as accurate execution of each task. Inferred models should
 256 not exhibit fully random responses (50%), or perfect performance (100%), since perfection is never
 257 attained by even the best trained rats. We formulate rapid task switching as an emergent property
 258 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%
 259 with variance 5%².

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

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

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

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

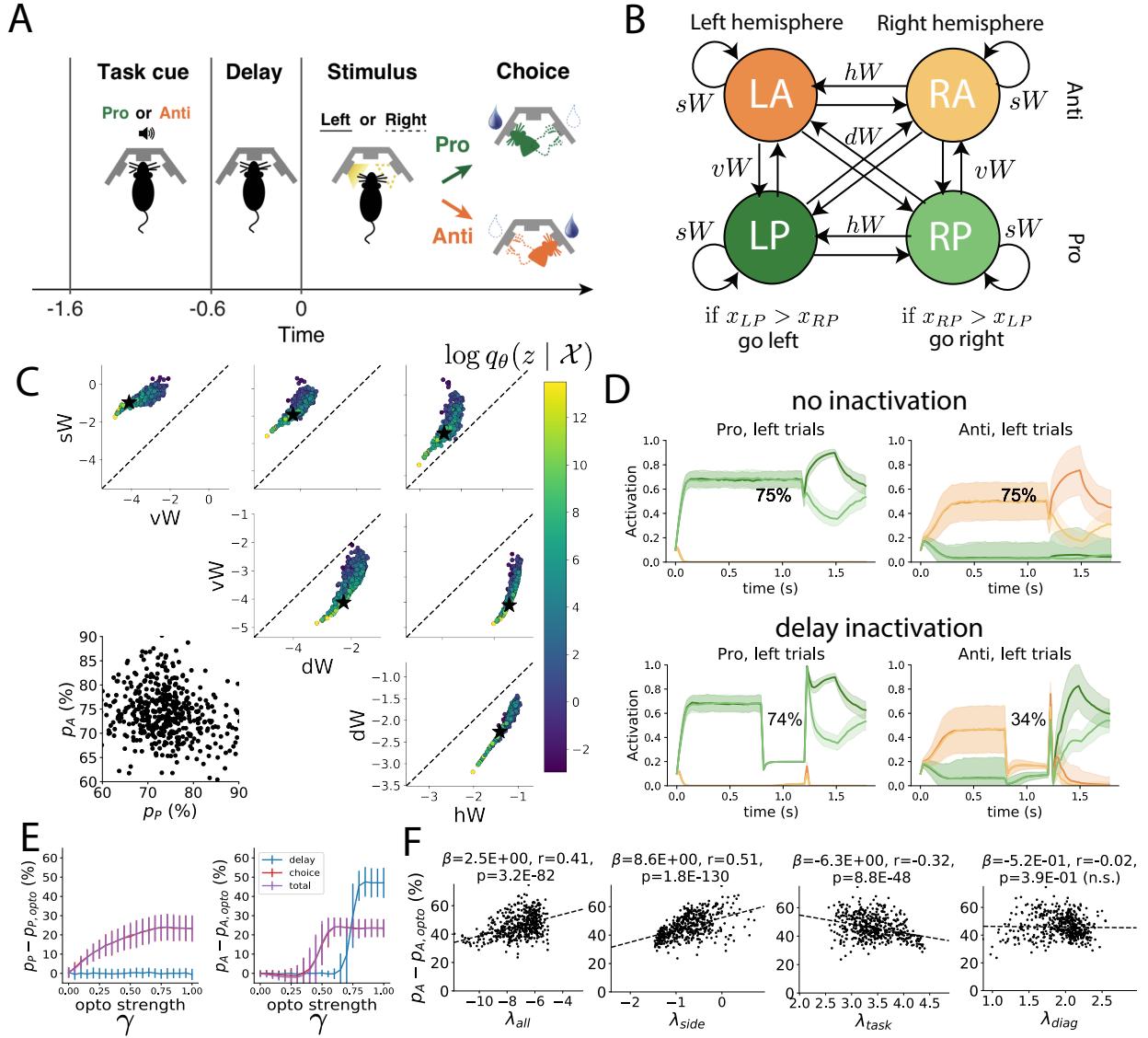


Figure 3: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. Subscripts P and A of connectivity weights 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 “opto strength” 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.

278 In agreement with experimental results from Duan et al., we found that inactivation above nominal
 279 strength during the delay period consistently decreased performance in the Anti task, but had no
 280 consistent effect on the Pro task (Fig. 3E) e.g. (Fig. 3D, bottom). This difference in resiliency
 281 across tasks to delay perturbation is a prediction made by the inferred EPI distribution, rather
 282 than an emergent property that was conditioned upon. Even though p_P and p_A are anticorrelated
 283 in the EPI posterior ($r = -0.15$, $p = 3.68 \times 10^{-12}$), greater p_P and p_A both result in decreased
 284 resiliency to delay perturbation in the Anti task (Fig. 14). Ultimately, lower λ_{side} and λ_{all} and
 285 greater λ_{task} produce networks more robust to delay perturbation in the Anti task (Fig. 3F)).
 286

286 In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task
 287 switching. This posterior revealed the mechanisms leading to greater accuracy in each task as well
 288 as those increasing resiliency to perturbation in the Anti task. Importantly, every connectivity
 289 from this inferred distribution predicts fragility and robustness of performance in the Anti and Pro
 290 tasks, respectively. EPI allows us to conclude that since *all* parameters of this model producing
 291 rapid task switching make such an experimentally verified prediction, we have a well chosen model.
 292

3.5 EPI scales well to high-dimensional parameter spaces

293 Here, we are interested in the scalability of EPI in number of parameters ($|\mathbf{z}|$). We consider rank-2
 294 RNN with N neurons of connectivity
 295

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

and dynamics

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

296 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 $g = 0.01$.
 297

297 We want to learn distributions of connectivity that produce stable amplification. Two conditions
 298 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

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

303 EPI can naturally condition on this emergent property

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

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.

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

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 consistently
309 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 large-scale neural
317 datasets [47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57] (see review, [16]). 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
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

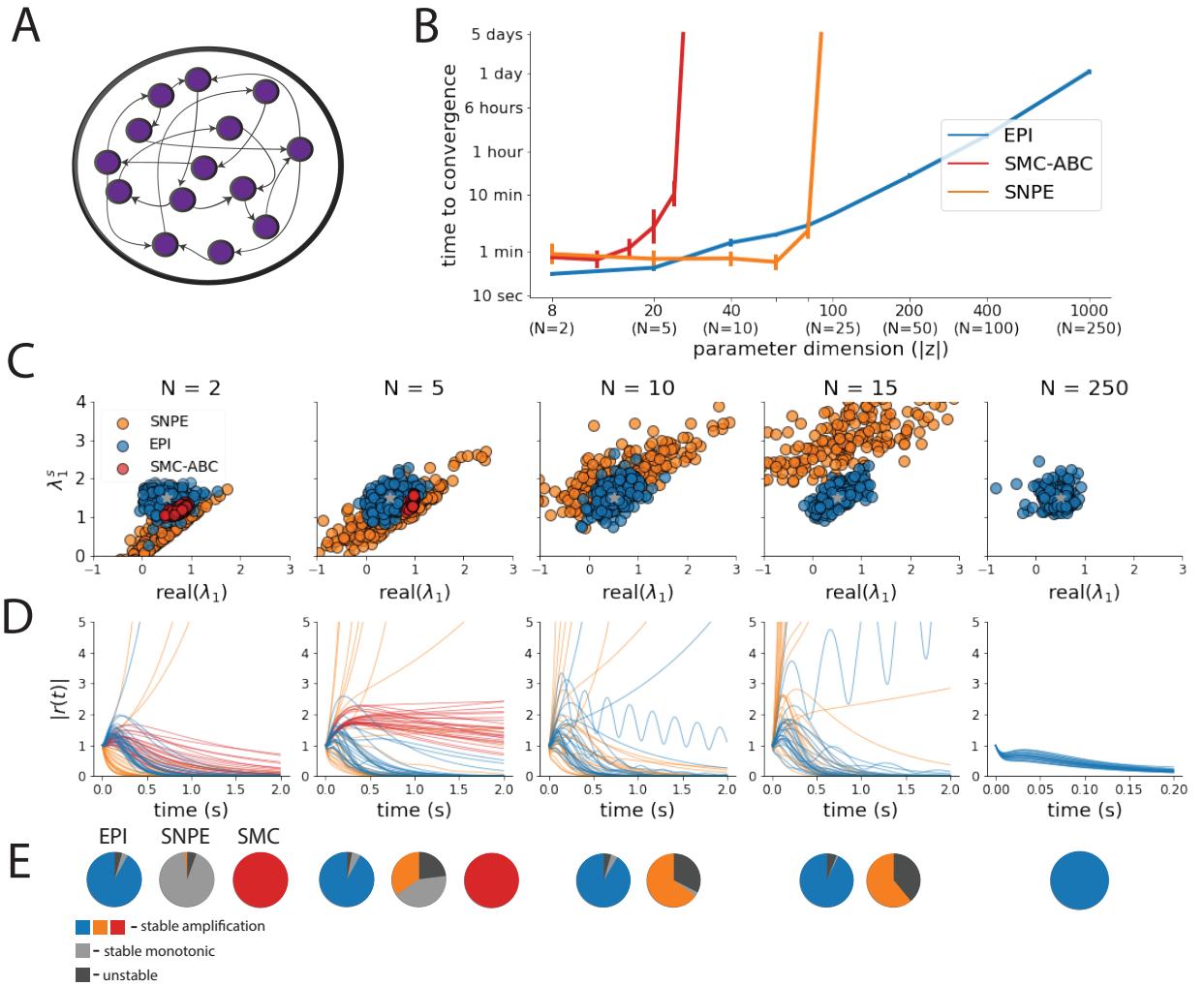


Figure 4: **A.** Recurrent neural network. **B.** EPI scales with z to high dimensions. Convergence definitions: EPI (blue) - satisfies all moment constraints, SNPE (orange)- produces at least $2/n_{\text{train}}$ parameter samples are in the bounds of emergent property (mean ± 0.5), and SMC-ABC (red) - 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 ($\tau = 100\text{ms}$). Each trace corresponds to simulation of one z . **E.** Ratio of obtained samples producing stable amplification.

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 [58] and non-trivial dynamics on strange attractors [59].

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 [26, 42, 43]. 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 [60, 61]. This methodology relies on recently extended
347 theory of responses in random neural networks with low-rank structure [62]. 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 [63]. 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 **Acknowledgements:**

359 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
360 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
361 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
362 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
363 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James
364 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

365 **Data availability statement:**

366 The datasets generated during and/or analysed during the current study are available from the
367 corresponding author upon reasonable request.

368 **Code availability statement:**

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

371 **References**

- 372 [1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 373 [2] John J Hopfield. Neural networks and physical systems with emergent collective computational
374 abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 375 [3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
376 networks. *Physical review letters*, 61(3):259, 1988.
- 377 [4] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
378 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
379 17(11):4382–4388, 1997.
- 380 [5] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in
381 perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 382 [6] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
383 A framework for parameter estimation and model selection from experimental data in systems
384 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.

- 385 [7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.
- 386
- 387 [8] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation
388 and variational inference in deep latent gaussian models. *International Conference on Machine
389 Learning*, 2014.
- 390 [9] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
391 neural population models through nonlinear embeddings. In *Advances in neural information
392 processing systems*, pages 163–171, 2016.
- 393 [10] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear
394 dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 395 [11] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders.
396 *bioRxiv*, page 399246, 2018.
- 397 [12] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
398 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg,
399 et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature
400 methods*, page 1, 2018.
- 401 [13] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
402 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
403 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 404 [14] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
405 Datta. Composing graphical models with neural networks for structured representations and
406 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 407 [15] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
408 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
409 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural
410 Information Processing Systems*, 2019.
- 411 [16] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
412 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
413 2018.

- 414 [17] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics.
415 *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.
- 416 [18] David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models.
417 2014.
- 418 [19] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R
419 Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology,
420 and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.
- 421 [20] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
422 free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–
423 5533, 2017.
- 424 [21] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
425 *International Conference on Machine Learning*, 2015.
- 426 [22] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
427 *arXiv preprint arXiv:1605.08803*, 2016.
- 428 [23] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density
429 estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.
- 430 [24] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
431 networks. *International Conference on Learning Representations*, 2017.
- 432 [25] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness,
433 and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.
- 434 [26] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and vi-
435 sual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,
436 115(3):1399–1409, 2016.
- 437 [27] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
438 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
439 routing. *bioRxiv*, page 245613, 2018.
- 440 [28] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
441 tion. *Neural Networks*, 15(4-6):479–493, 2002.

- 442 [29] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
443 parameters. *Nature neuroscience*, 7(12):1345, 2004.
- 444 [30] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
445 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
446 *Neuron*, 77(5):845–858, 2013.
- 447 [31] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
448 1957.
- 449 [32] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
450 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 451 [33] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
452 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 453 [34] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
454 2009.
- 455 [35] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
456 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 457 [36] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
458 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
459 592, 2009.
- 460 [37] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
461 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,
462 85(2):402–417, 2015.
- 463 [38] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
464 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
465 5(10):793, 2004.
- 466 [39] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
467 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-*
468 *biology*, 71(1):45–61, 2011.
- 469 [40] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex:
470 From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.

- 471 [41] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
472 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
473 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 474 [42] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing Wang.
475 Paradoxical response reversal of top- down modulation in cortical circuits with three interneu-
476 ron types. *Elife*, 6:1–15, 2017.
- 477 [43] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying
478 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 479 [44] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
480 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
481 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 482 [45] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
483 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
484 operating regime of cortex. *bioRxiv*, 2020.
- 485 [46] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain
486 regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.
- 487 [47] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
488 13(8):1713–1720, 2001.
- 489 [48] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
490 A statistical paradigm for neural spike train decoding applied to position prediction from
491 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
492 7425, 1998.
- 493 [49] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
494 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 495 [50] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A
496 point process framework for relating neural spiking activity to spiking history, neural ensemble,
497 and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

- 498 [51] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan
499 Segev. A novel multiple objective optimization framework for constraining conductance-based
500 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 501 [52] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
502 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis
503 of neural population activity. In *Advances in neural information processing systems*, pages
504 1881–1888, 2009.
- 505 [53] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In
506 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 507 [54] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
508 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
509 *Science*, 349(6244):184–187, 2015.
- 510 [55] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq
511 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in
512 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 513 [56] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
514 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th Inter-*
515 *national Conference on Machine Learning*, 2019.
- 516 [57] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.
517 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.
518 *Nature Communications*, 10(4933), 2019.
- 519 [58] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give
520 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of*
521 *Sciences*, 115(13):3464–3469, 2018.
- 522 [59] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emergent
523 dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint*
524 *arXiv:1605.04463*, 2016.
- 525 [60] Xaq Pitkow and Dora E Angelaki. Inference in the brain: statistics flowing in redundant
526 population codes. *Neuron*, 94(5):943–953, 2017.

- 527 [61] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like
528 dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*,
529 page 696088, 2019.
- 530 [62] Francesca Mastrogiovisepppe and Srdjan Ostojic. Linking connectivity, dynamics, and computa-
531 tions in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.
- 532 [63] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,
533 2019.
- 534 [64] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
535 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
536 2012.
- 537 [65] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
538 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
539 2018.
- 540 [66] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
541 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 542 [67] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
543 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
544 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
545 1929, 2009.
- 546 [68] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
547 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 548 [69] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
549 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 550 [70] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International
551 Conference on Learning Representations*, 2015.
- 552 [71] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
553 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 554 [72] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for
555 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

556 **5 Methods**

557 **5.1 Emergent property inference (EPI)**

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

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

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

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

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

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

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

587 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
588 the neural network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep
589 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. $\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2$.

590 Entropy is chosen as the normative selection principle, since we want the posterior to only contain
591 structure predicated by the emergent property [31, 32]. This choice of selection principle is also
592 that of standard Bayesian inference, and we derive an exact relation between EPI and variational
593 inference (see Section 5.1.5). However, a key difference is that variational inference and other
594 Bayesian methods do not constrain the predictions of their inferred posteriors. This optimization
595 is executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [24].

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

604 5.1.1 Related approaches

605 Inverse problems across scientific fields have been approached in a variety of ways. The goals of
606 such approaches are generally to a.) identify parametric sensitivities and manifolds of degeneracy
607 with respect to some model output (structural identifiability analysis), and to reason about the
608 parameters that are most probable to have produced a model output (statistical inference). While
609 much research in computational neuroscience has focused on optimizing neural architectures to
610 process information and accomplish tasks [63], structure in parameter space of the set of optimized
611 solutions is rarely discussed and lacks a probabilistic treatment.

612 To understand sensitivity around a point, one can measure the Jacobian. One approach that scales
613 well is EAR [64]. A popular efficient approach for systems of ODEs has been neural ODE adjoint
614 [65] and its stochastic adaptation [66]. Casting identifiability as a statistical estimation problem,
615 the profile likelihood can assess via iterated optimization while holding parameters fixed [67]. An
616 exciting recent method is capable of recovering the functional form of such unidentifiabilities away
617 from a point [68]. Global structural non-identifiabilities can be found for models with polynomial
618 or rational dynamics equations using DAISY [69].

619 MCMC approaches

620 ABC approaches

621 NF LFI approaches

622 5.1.2 Normalizing flows

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

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

626 However, this computation has cubic complexity in dimensionality for fully connected layers. By
627 restricting our layers to normalizing flows [21] – bijective functions with fast log determinant Ja-
628 cobian computations, we can tractably optimize deep generative models with objectives that are a
629 function of sample density, like entropy. TODO: (clean up) We use Real NVP because it's a cou-
630 pling architecture, which is fast to run either forwards (probability with samples) and backwards
631 (prroability or hessian). Normalizing flow architectures for deep probability distributions used in
632 EPI are specified by the number of masks, neural network layers per mask, units per layer, and
633 batch normalization momentum parameter.

634 optimization with respect to entropy, and to confer immediate probability evaluations after opti-
635 mization.

636 5.1.3 Augmented Lagrangian optimization

637 Since we are optimizing parameters θ of our deep probability distribution with respect to the
638 entropy $H(q_\theta(\mathbf{z}))$, we must take gradients with respect to the log probability density of samples from

639 the deep probability distribution. Entropy of $q_{\theta}(\mathbf{z})$ can be expressed using the reparameterization
 640 trick as an expectation of the negative log density of parameter samples \mathbf{z} over the randomness in
 641 the parameterless initial distribution $q_0(\mathbf{z}_0)$:

$$H(q_{\theta}(\mathbf{z})) = \int -q_{\theta}(\mathbf{z}) \log(q_{\theta}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [-\log(q_{\theta}(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\log(q_{\theta}(g_{\theta}(\mathbf{z}_0)))]. \quad (19)$$

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

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

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

$$L(\theta; \eta_{\text{opt}}, c) = -H(q_{\theta}) + \eta_{\text{opt}}^{\top} R(\theta) + \frac{c}{2} \|R(\theta)\|^2 \quad (21)$$

646 where $R(\theta) = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [T(\mathbf{x}; \mathbf{z}) - \mu_{\text{opt}}]]$, $\eta_{\text{opt}} \in \mathbb{R}^m$ are the Lagrange multipliers where
 647 $m = |\mu_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})|$, and c is the penalty coefficient. These Lagrange multipliers are closely
 648 related to the natural parameters η of exponential families (see Section 5.1.6). Deep neural network
 649 weights and biases θ of the deep probability distribution are optimized according to Equation 21
 650 using the Adam optimizer with its standard parameterization [70]. η_{opt} is initialized to the zero
 651 vector and adapted following each augmented Lagrangian epoch, which is a period of optimization
 652 with fixed (η_{opt}, c) for a given number of stochastic optimization iterations. A low value of c is
 653 used initially, and conditionally increased after each epoch based on constraint error reduction. For
 654 example, the initial value of c was $c_0 = 10^{-3}$ during EPI with the oscillating 2D LDS (Fig. S1C).
 655 The penalty coefficient is updated based on the result of a hypothesis test regarding the reduction in
 656 constraint violation. The p-value of $\mathbb{E}[|R(\theta_{k+1})|] > \gamma \mathbb{E}[|R(\theta_k)|]$ is computed, and c_{k+1} is updated
 657 to βc_k with probability $1-p$. The other update rule is $\eta_{\text{opt},k+1} = \eta_{\text{opt},k} + c_k \frac{1}{n} \sum_{i=1}^n (T(\mathbf{x}^{(i)}) - \mu)$ given
 658 a batch size n . Throughout the study, $\beta = 4.0$, $\gamma = 0.25$, and the batch size was a hyperparameter,
 659 which varied according to the application of EPI.

660 The intention is that c and η_{opt} start at values encouraging entropic growth early in optimization.
 661 With each training epoch in which the update rule for c is invoked by unsatisfactory constraint
 662 error reduction, the constraint satisfaction terms are increasingly weighted, resulting in a decreased
 663 entropy. This encourages the discovery of suitable regions of parameter space, and the subsequent
 664 refinement of the distribution to produce the emergent property. In the oscillating 2D LDS example,
 665 each augmented Lagrangian epoch ran for 2,000 iterations (Fig. S1C-D). Notice the initial entropic

666 growth, and subsequent reduction upon each update of η_{opt} and c . The momentum parameters of
667 the Adam optimizer were reset at the end of each augmented Lagrangian epoch.

668 Rather than starting optimization from some θ drawn from a randomized distribution, we found
669 that initializing $q_{\theta}(\mathbf{z})$ to approximate an isotropic Gaussian distribution conferred more stable, con-
670 sistent optimization. The parameters of the Gaussain initialization were chosen on an application-
671 specific basis. Throughout the study, we chose isotropic Gaussian initializations with mean μ_{init}
672 at the center of the distribution support and some standard deviation σ_{init} , except for one case,
673 where an initialization informed by random search was used (see Section 5.2.2).

674 To assess whether EPI distribution $q_{\theta}(\mathbf{z})$ produces the emergent property, we defined a hypothesis
675 testing convergence criteria. The algorithm has converged when a null hypothesis test of constraint
676 violations $R(\theta)_i$ being zero is accepted for all constraints $i \in \{1, \dots, m\}$ at a significance threshold
677 $\alpha = 0.05$. This significance threshold is adjusted through Bonferroni correction according to the
678 number of constraints m . The p-values for each constraint are calculated according to a two-tailed
679 nonparametric test, where 200 estimations of the sample mean $R(\theta)^i$ are made from k resamplings
680 of \mathbf{z} from a finite sample of size n taken at the end of the augmented Lagrangian epoch. k is
681 determined by a fraction of the batch size ν , which varies according to the application. In the
682 linear two-dimensional system example, we used a batch size of $n = 1000$ and set $\nu = 0.1$ resulting
683 in convergence after the ninth epoch of optimization. (Fig. S1C-D black dotted line).

684 When assessing the suitability of EPI for a particular modeling question, there are some important
685 technical considerations. First and foremost, as in any optimization problem, the defined emergent
686 property should always be appropriately conditioned (constraints should not have wildly different
687 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution
688 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-
689 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,
690 one should consider the computational cost of the gradient calculations. In the best circumstance,
691 there is a simple, closed form expression (e.g. Section 5.1.4) for the emergent property statistic
692 given the model parameters. On the other end of the spectrum, many forward simulation iterations
693 may be required before a high quality measurement of the emergent property statistic is available
694 (e.g. Section 5.2.1). In such cases, optimization will be expensive.

695 **5.1.4 Example: 2D LDS**

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

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

698 with

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

699 To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fixing $\tau = 1$), the emergent property statistics $T(\mathbf{x})$ were chosen to contain the first and second
 700 moments of the oscillatory frequency, $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscil-
 701 lating system. λ_1 is the eigenvalue of greatest real part when the imaginary component is zero, and
 702 alternatively of positive imaginary component when the eigenvalues are complex conjugate pairs.
 703 To learn the distribution of real entries of A that produce a band of oscillating systems around
 704 1Hz, we formalized this emergent property as $\text{real}(\lambda_1)$ having mean zero with variance 0.25^2 , and
 705 the oscillation frequency $2\pi\text{imag}(\lambda_1)$ having mean $\omega = 1$ Hz with variance $(0.1\text{Hz})^2$:

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

707

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

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

711 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
 712 does not require simulation, we cannot derive the distribution $q_{\boldsymbol{\theta}}^*$ directly. This fact is due to the
 713 formally hard problem of the backward mapping: finding the natural parameters η from the mean
 714 parameters $\boldsymbol{\mu}$ of an exponential family distribution [71]. Instead, we used EPI to approximate this
 715 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two
 716 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped
 717 onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

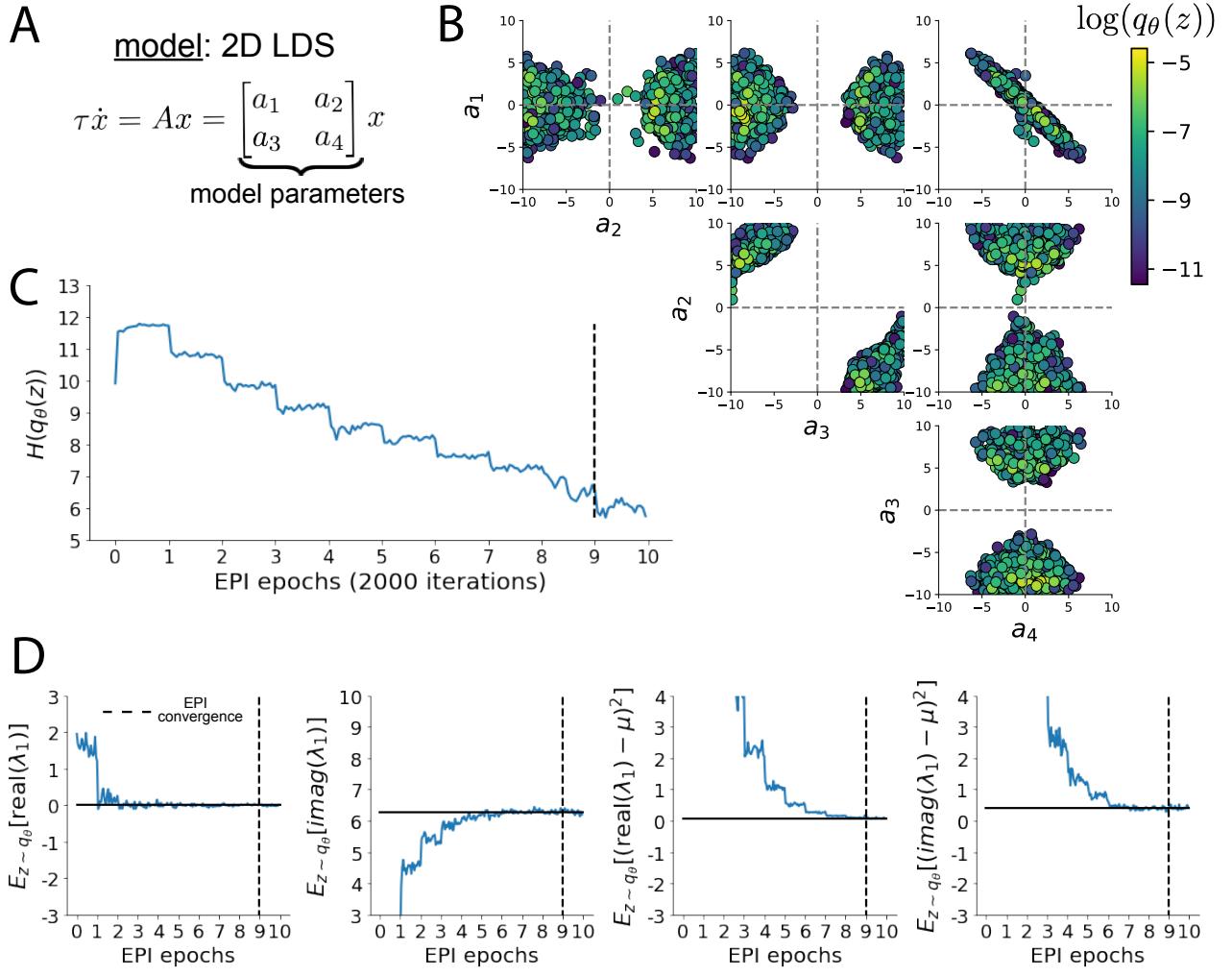


Figure 5: (LDS1): A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed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

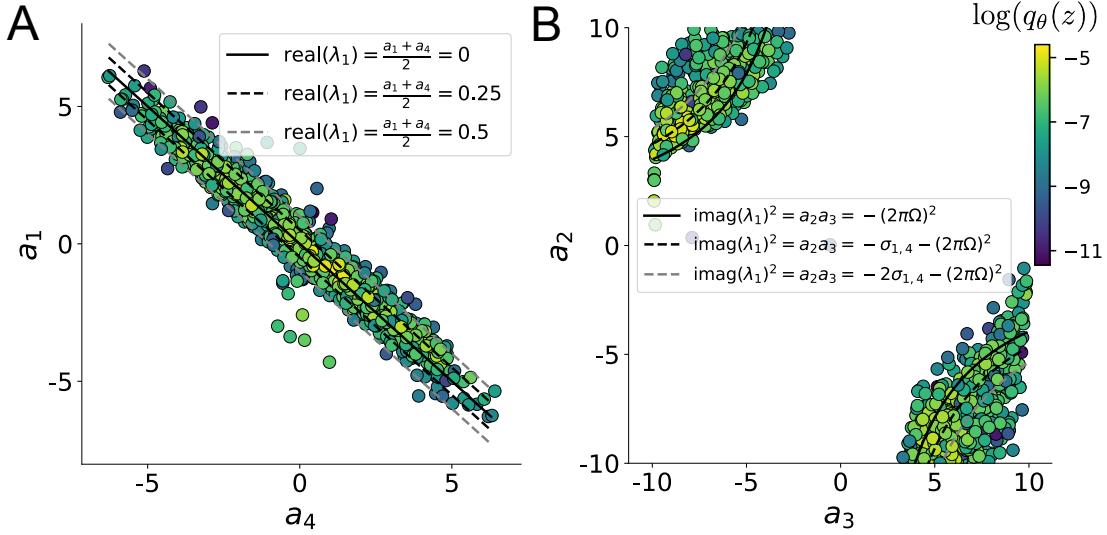


Figure 6: (LDS2): A. Probability contours in the a_1 - a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. B. Probability contours in the a_2 - a_3 plane were derived from the emergent property statistic of oscillation frequency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

718 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the
 719 parameter distribution. To validate our method, we analytically derived the contours of the prob-
 720 ability density from the emergent property statistics and values. In the a_1 - a_4 plane, the black
 721 line at $\text{real}(\lambda_1) = \frac{a_1 + a_4}{2} = 0$, dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1 + a_4}{2} \pm 0.25$,
 722 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
 723 of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired
 724 statistical constraints and model degeneracy in the sum of a_1 and a_4 . Intuitively, the parameters
 725 equivalent with respect to emergent property statistic $\text{real}(\lambda_1)$ have similar log densities.

726 To explain the bimodality of the EPI distribution, we examined the imaginary component of λ_1 .
 727 When $\text{real}(\lambda_1) = \frac{a_1 + a_4}{2} = 0$, we have

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1 a_4 - a_2 a_3}{\tau}}, & \text{if } a_1 a_4 < a_2 a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (26)$$

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

$$\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 \quad (27)$$

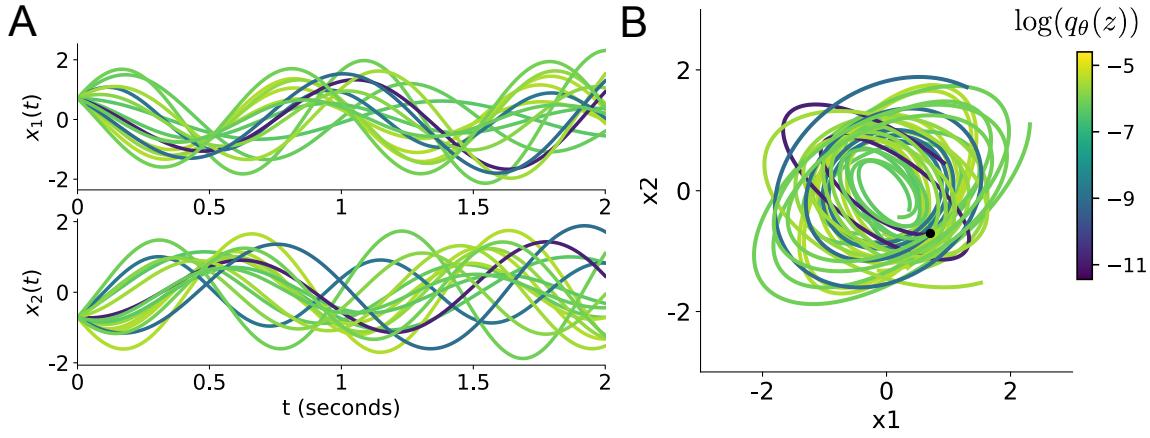


Figure 7: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\theta}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. A. Each dimension of the simulated trajectories throughout time. B The simulated trajectories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

730 Since we constrained $\mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\text{imag}(\lambda)] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation
 731 $\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 = (2\pi)^2$ for various $a_1 a_4$ (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} (|a_1 a_4 - E_{q_{\theta}}[a_1 a_4]|)$,
 732 we show the contours as $a_1 a_4 = 0$ (black), $a_1 a_4 = -\sigma_{1,4}$ (black dotted), and $a_1 a_4 = -2\sigma_{1,4}$ (grey
 733 dotted). This validates the curved structure of the inferred distribution learned through EPI. We
 734 took steps in negative standard deviation of $a_1 a_4$ (dotted and gray lines), since there are few positive
 735 values $a_1 a_4$ in the learned distribution. Subtler combinations of model and emergent property will
 736 have more complexity, further motivating the use of EPI for understanding these systems. As we
 737 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz
 738 (Fig. S3).

739 5.1.5 Maximum entropy distributions and exponential families

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

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

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

742 will have probability density in the exponential family:

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

743 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
 744 formally hard to identify [71].

745 In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equation
 746 2

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

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

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

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

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

750 which are constrained to the chosen means and variances

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

751 5.1.6 EPI as variational inference

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

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

756 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
 757 However, conjugacy is rare in practice, and alternative methods, such as variational inference [72],
 758 are utilized.

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

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

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

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

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

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

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

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

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

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

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

770 where

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

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

772 The variational objective is then

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

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

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

775 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
 776 (Equation 32). Thus, EPI is implicitly executing variational inference with a uniform prior and a
 777 mean-field gaussian likelihood on the emergent property statistics. The data \mathbf{x} used by this implicit
 778 variational inference program would be that generated by the adapting variational approximation

779 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{z})q_{\boldsymbol{\theta}}(\mathbf{z})$, and the likelihood parameters $\boldsymbol{\eta}_f$ of EPI optimization epoch k are predicated
 780 by $\boldsymbol{\eta}_{\text{opt},k}$. However, in EPI we have not specified a prior distribution, or collected data, which can
 781 inform us about model parameters. Instead we have a mathematical specification of an emergent
 782 property, which the model must produce, and a maximum entropy selection principle. Accordingly,
 783 we replace the notation of $p(\mathbf{z} | \mathbf{x})$ with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys
 784 emergent property \mathcal{X} (see Section 5.1).

785 5.2 Theoretical models

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

788 5.2.1 Stomatogastric ganglion

789 We analyze how the parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$ govern the emergent phenomena of intermediate
 790 hub frequency in a model of the stomatogastric ganglion (STG) [30] shown in Figure 1A with
 791 activity $\mathbf{x} = [x_{\text{f1}}, x_{\text{f2}}, x_{\text{hub}}, x_{\text{s1}}, x_{\text{s2}}]$, using the same hyperparameter choices as Gutierrez et al.
 792 Each neuron's membrane potential $x_{\alpha}(t)$ for $\alpha \in \{\text{f1}, \text{f2}, \text{hub}, \text{s1}, \text{s2}\}$ is the solution of the following
 793 stochastic differential equation:

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

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

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

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

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

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

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

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

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

805 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 806 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 807 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 808 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
 809 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
 810 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

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

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

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

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

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

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

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

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

819 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 820 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

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

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

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

825 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for
 826 $T = 300$ time steps of $dt = 25ms$. The chosen dt and T were the most computationally convenient

827 choices yielding accurate frequency measurement. We used a basis of complex exponentials with
 828 frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

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

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

833 To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in$
 834 $\mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in$
 835 $\{f1, f2, \text{hub}, s1, s2\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

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

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

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

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

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

843

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

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

846 For EPI in Fig 1E, we used a real NVP architecture with three coupling layers of affine transforma-
 847 tions parameterized by two-layer neural networks of 25 units per layer. The initial distribution was
 848 a standard isotropic gaussian $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to a support of $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}] \in [4, 8] \times [0.01, 4]$.

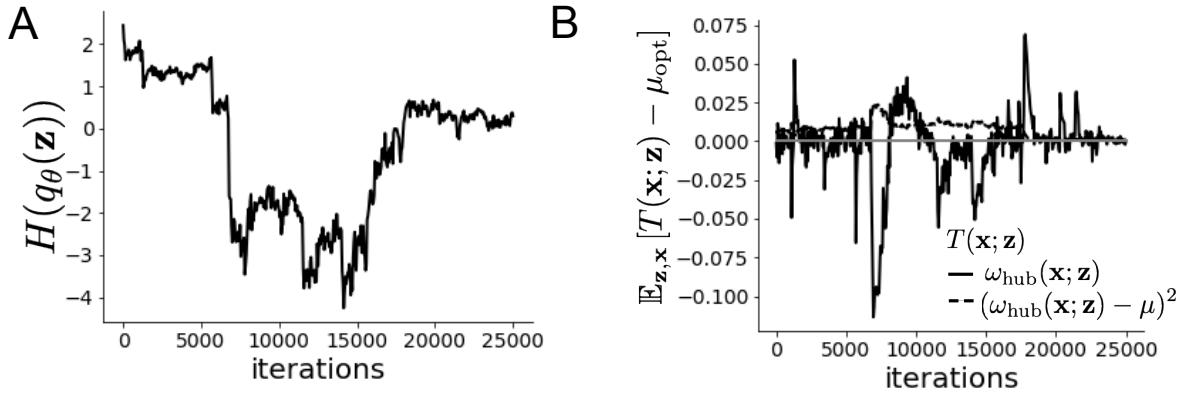


Figure 8: (STG1): EPI optimization of the STG model producing network syncing. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

849 We did not include $g_{\text{synA}} < 0.01$, since conductances that low make the circuit simulations numerically unstable. We used an augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, set $\nu = 0.25$, and initialized $q_\theta(\mathbf{z})$ to produce a gaussian approximation to samples returned from an initial ABC search. This initialization had much greater entropy and a different emergent property than the the returned EPI posterior.

854 TODO write about specifics of the Hessian analysis.

855 5.2.2 Primary visual cortex

856 Connectivity (W_{fit}) and input ($\mathbf{h}_{b,\text{fit}}$ and $\mathbf{h}_{c,\text{fit}}$) parameters were fit using the deterministic V1 circuit model [45]

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

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

858 and

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

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

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

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

861 and

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

862 Since the E-population of this network increases exponentially in the absense of recurrent in-
863 hibitory feedback, we may also observe a paradoxical effect in the inhibitory populations (which
864 is present in E-I networks). At 50% contrast (Fig. 2B, dots), this network exhibits a paradoxical
865 effect in the P-population (Fig. 2C), but no others (Fig. 9). That is, for a small increase in h_P ,
866 $\mathbb{E}_t[x_P]$ decreases.

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

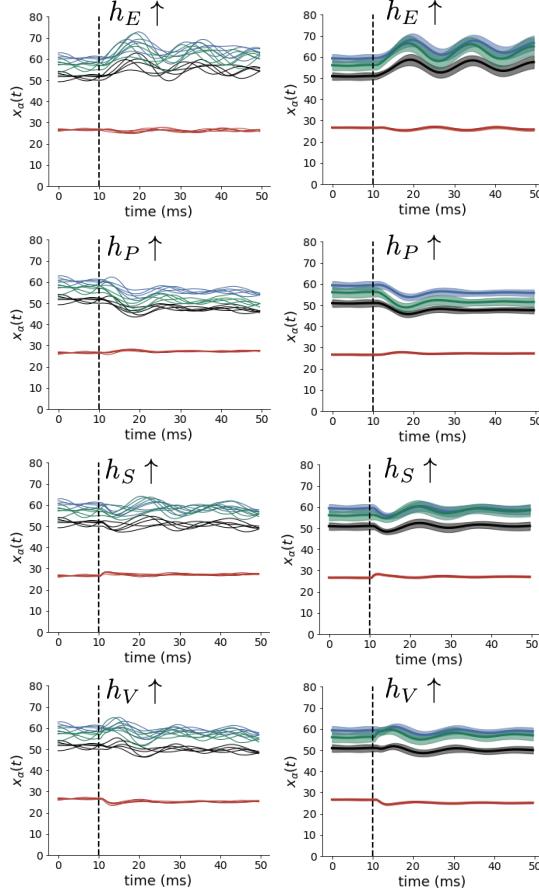


Figure 9: Supplemental Figure: (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average and standard deviation of stochastic fluctuations of responses.

869 5.2.3 Superior colliculus

870 In the model of Duan et al [27], there are four total units: two in each hemisphere corresponding to
 871 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right
 872 Pro (RP) and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related
 873 by

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

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

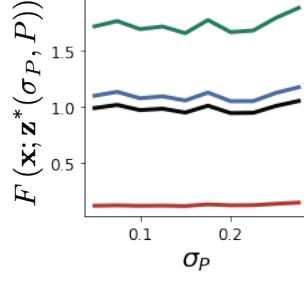


Figure 10: Supplemental Figure: Fano factors along the ridge of the posterior in Fig. 2E.

876 to the optogenetic strength γ

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

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

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

878 which evolve according to

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

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

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

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

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

882 There is a constant input to every population,

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

883 a bias to the Pro populations

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

884 rule-based input depending on the condition

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

885

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

886 a choice-period input

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

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

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

888 The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{P,\text{bias}} = 0.5$, $I_{P,\text{rule}} = 0.6$, $I_{A,\text{rule}} = 0.6$,

889 $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

890 The accuracies of p_P and p_A are calculated as

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

891 and

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

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

893 The Heaviside step function is approximated as

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

894 where $\beta = 100$.

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

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

897

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

898 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 899 epoch of 2,000 iterations(see Section 5.1.3). The optimization converged after six epochs (Fig. 15).

900 For EPI in Fig. 3C, we used a real NVP architecture with three coupling layers of affine transforma-
 901 tions parameterized by two-layer neural networks of 50 units per layer. The initial distribution was
 902 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-
 903 mented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, set $\nu = 0.5$, and initialized $q_{\theta}(\mathbf{z})$
 904 to produce an isotropic gaussian with mean 0 and variance 2.5^2 . Accuracies were estimated over
 905 200 trials of random gaussian noise, which was sampled independently for each drawn parameter \mathbf{z}
 906 and each iteration of the EPI optimization.

907 5.2.4 Rank-2 RNN

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

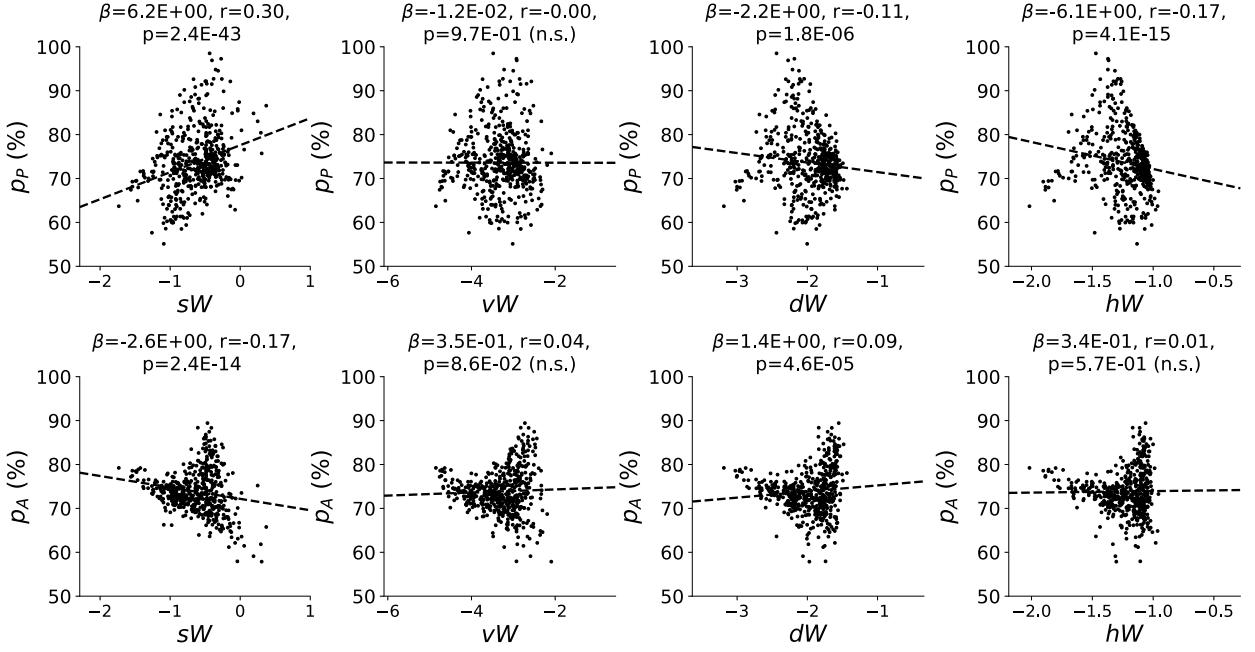


Figure 11: (SC1): Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p value.

918 SMC-ABC has many hyperparameters, of which pyABC selects automatically by running some
 919 initial diagnostics upon initialization. In concurrence with the literature, SMC-ABC fails to con-
 920 verge around 25-30 dimensions, since it's proposal samples never get close enough to the target
 921 statistics. We searched over many SNPE hyperparameter choices: $n_{\text{train}} \in [2,000, 10,000, 100,000]$
 922 is the number of simulations run per training epoch, and $n_{\text{mades}} \in [2, 3]$ is the number of masked au-
 923 toregressive density estimators in the deep parameter distribution architecture. The greater n_{train} ,
 924 the longer each epoch will take, but the more likely SNPE may converge during that epoch. Greater
 925 n_{mades} increases the flexibility of the deep parameter distribution of SNPE, but slows optimization.
 926 For the timing plot, we show the fastest among all of these choices, and for the convergence plot,
 927 we show the best convergence among all of these choices. During optimization, we used $n_{\text{atom}}=100$
 928 atomic proposals as is recommended.

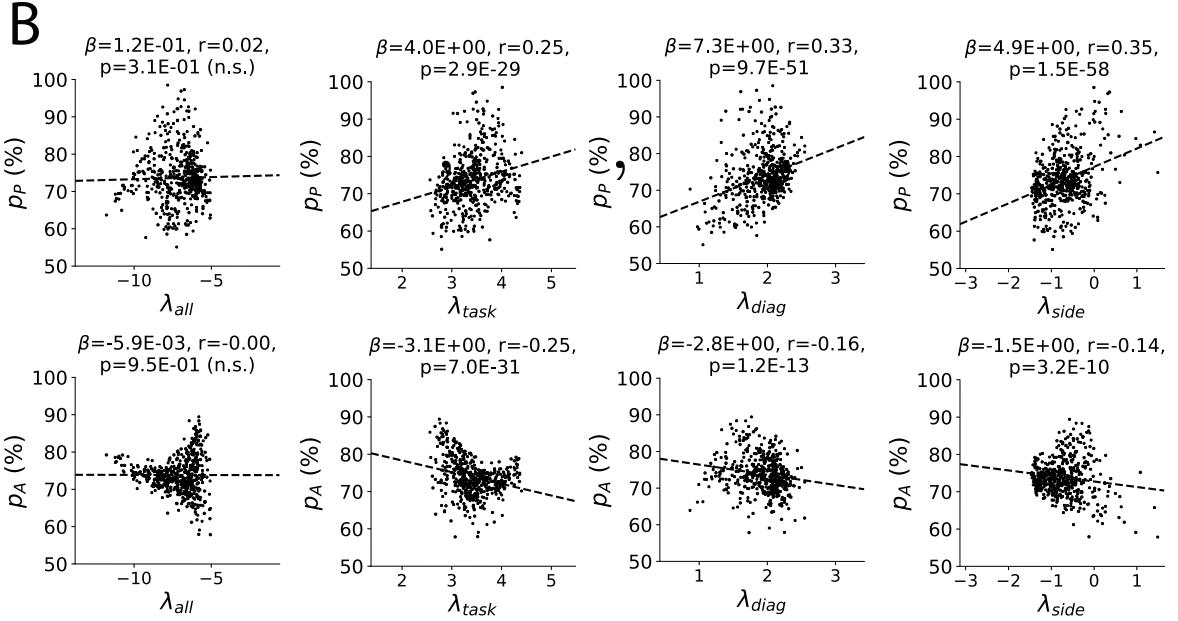
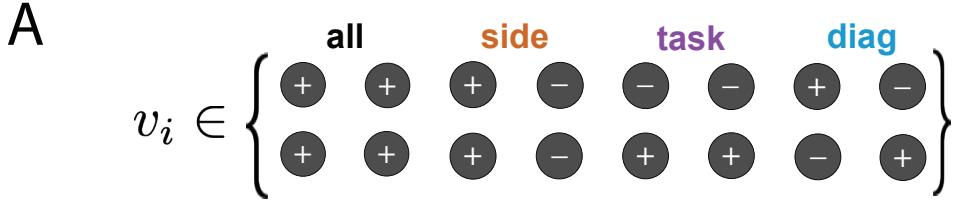


Figure 12: (SC2): A. Invariant eigenvectors of connectivity matrix W . B. Eigenvalues of connectivities of EPI distribution versus task accuracies.

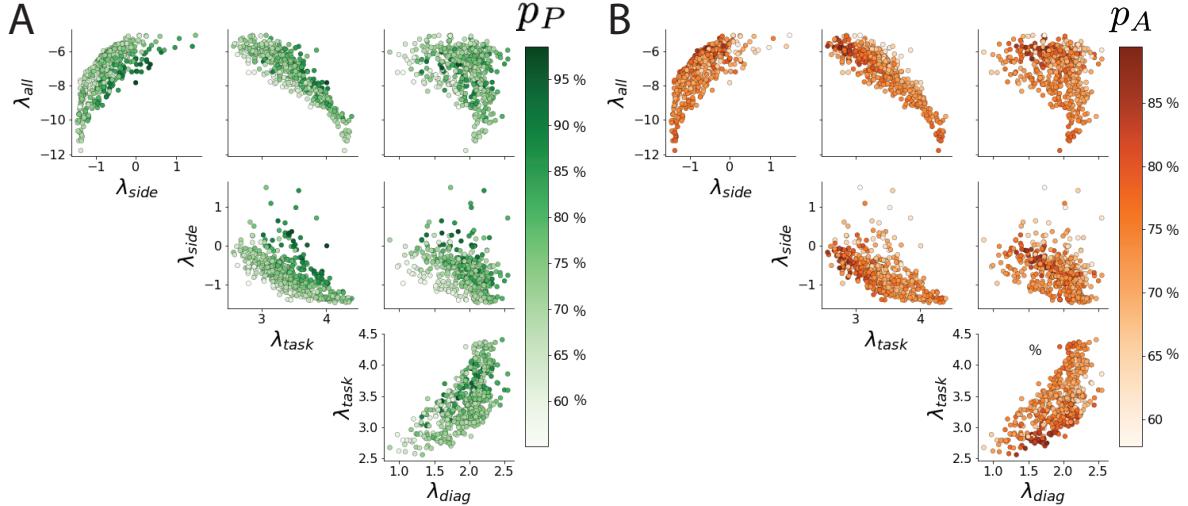


Figure 13: (SC3): A. Connectitivty eigenvalues of EPI parameter distribution colored by Pro task accuracy. B. Same for Anti task.

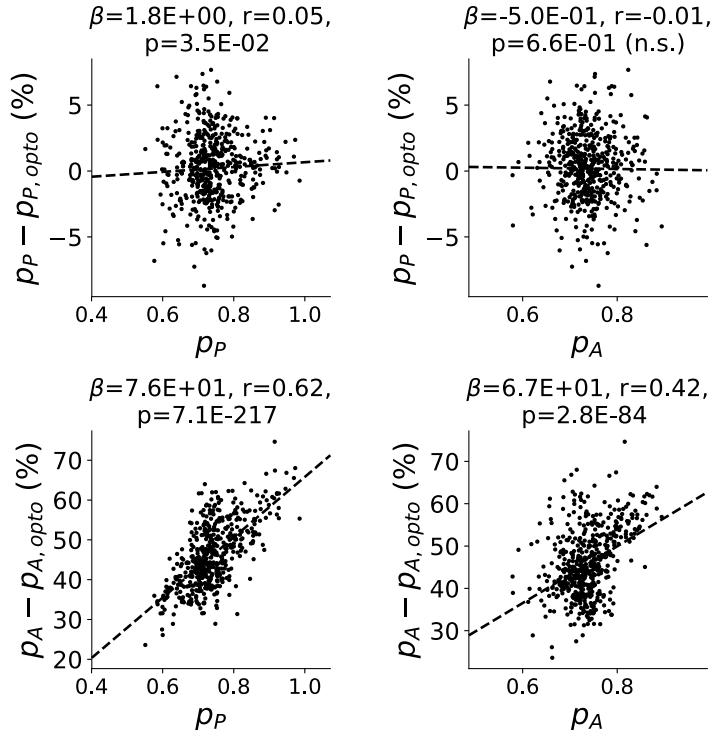


Figure 14: (SC4): Scatters of the effect of delay period inactivation in each task with task accuracy. Plots are shown at an opto strength of 0.8.

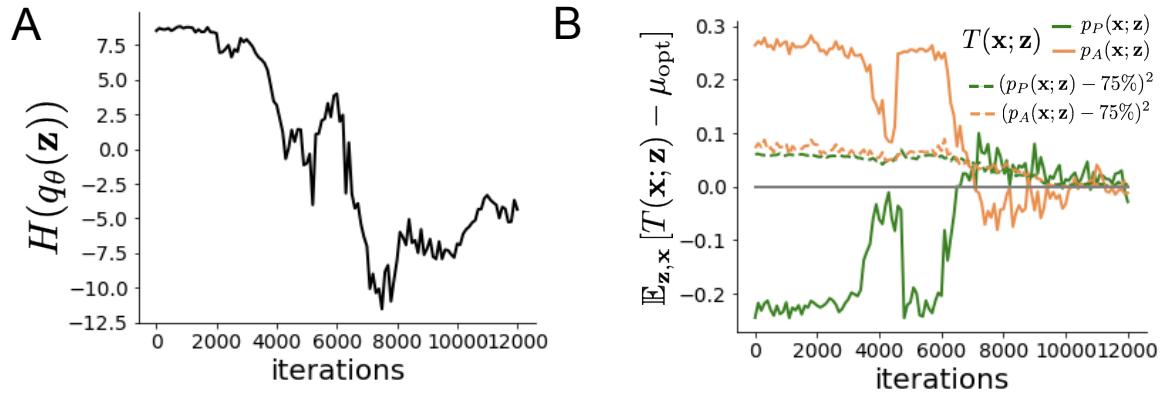


Figure 15: (SC5): EPI optimization of the SC model producing rapid task switching. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 12,000 iterations following the sixth augmented Lagrangian epoch.