

Interrogating theoretical models of neural computation with emergent property 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 a pattern of neural activity – and thus can
⁵ offer insights into neural computation. The operation of these circuits, like all models, critically
⁶ depends on the choice of model parameters. A key step is then to identify the model parameters
⁷ consistent with observed phenomena: to solve the inverse problem. In this work, we present a
⁸ novel technique, emergent property inference (EPI), that brings the modern probabilistic modeling
⁹ toolkit to theoretical neuroscience. When theorizing circuit models, theoreticians predominantly
¹⁰ focus on reproducing computational properties rather than a particular dataset. Our method uses
¹¹ deep neural networks to learn parameter distributions with these computational properties. This
¹² methodology is introduced through a motivational example inferring conductance parameters in a
¹³ circuit model of the stomatogastric ganglion. Then, with recurrent neural networks of increasing
¹⁴ size, we show that EPI allows precise control over the behavior of inferred parameters, and that
¹⁵ EPI scales better in parameter dimension than alternative techniques. In the remainder of this
¹⁶ work, we present novel theoretical findings gained through the examination of complex parametric
¹⁷ structure captured by EPI. In a model of primary visual cortex, we discovered how connectivity
¹⁸ with multiple inhibitory subtypes shapes variability in the excitatory population. Finally, in a
¹⁹ model of superior colliculus, we identified and characterized two distinct regimes of connectivity

20 that facilitate switching between opposite tasks amidst interleaved trials, characterized each regime
21 via insights afforded by EPI, and found conditions where these circuit models reproduce results
22 from optogenetic silencing experiments. Beyond its scientific contribution, this work illustrates
23 the variety of analyses possible once deep learning is harnessed towards solving theoretical inverse
24 problems.

25 2 Introduction

26 The fundamental practice of theoretical neuroscience is to use a mathematical model to understand
27 neural computation, whether that computation enables perception, action, or some intermediate
28 processing. A neural circuit is systematized with a set of equations – the model – and these
29 equations are motivated by biophysics, neurophysiology, and other conceptual considerations [1–5].

30 The function of this system is governed by the choice of model *parameters*, which when configured
31 in a particular way, give rise to a measurable signature of a computation. The work of analyzing
32 a model then requires solving the inverse problem: given a computation of interest, how can we
33 reason about the distribution of parameters that give rise to it? The inverse problem is crucial for
34 reasoning about likely parameter values, uniquenesses and degeneracies, and predictions made by
35 the model [6–8].

36 Ideally, one carefully designs a model and analytically derives how computational properties deter-
37 mine model parameters. Seminal examples of this gold standard include our field’s understanding
38 of memory capacity in associative neural networks [9], chaos and autocorrelation timescales in ran-
39 dom neural networks [10], central pattern generation [11], the paradoxical effect [12], and decision
40 making [13]. Unfortunately, as circuit models include more biological realism, theory via analytical
41 derivation becomes intractable. Absent this analysis, statistical inference offers a toolkit by which
42 to solve the inverse problem by identifying, at least approximately, the distribution of parameters
43 that produce computations in a biologically realistic model [14–19].

44 Statistical inference, of course, requires quantification of the vague term *computation*. In neuro-
45 science, two perspectives are dominant. First, often we directly use an *exemplar dataset*: a collection
46 of samples that express the computation of interest, this data being gathered either experimen-
47 tally in the lab or from a computer simulation. Though a natural choice given its connection to
48 experiment [20], some drawbacks exist: these data are well known to have features irrelevant to the
49 computation of interest [21–23], confounding inferences made on such data. Related to this point,

50 use of a conventional dataset encourages conventional data likelihoods or loss functions, which focus
51 on some global metric like squared error or marginal evidence, rather than the computation itself.
52 Alternatively, researchers often quantify an *emergent property* (EP): a statistic of data that directly
53 quantifies the computation of interest, wherein the dataset is implicit. While such a choice may
54 seem esoteric, it is not: the above “gold standard” examples [9–13] all quantify and focus on
55 some derived feature of the data, rather than the data drawn from the model. An emergent
56 property is of course a dataset by another name, but it suggests different approach to solving
57 the same inverse problem: here we directly specify the desired emergent property – a statistic
58 of data drawn from the model – and the value we wish that property to have, and we set up
59 an optimization program to find the distribution of parameters that produce this computation.
60 This statistical framework is not new: it is intimately connected to the literature on approximate
61 bayesian computation [24–26], parameter sensitivity analyses [27–30], maximum entropy modeling
62 [31–33], and approximate bayesian inference [34,35]; we detail these connections in Section 5.1.1.
63 The parameter distributions producing a computation may be curved or multimodal along vari-
64 ous parameter axes and combinations. It is by quantifying this complex structure that EPI offers
65 scientific insight. Traditional approximation families (e.g. mean-field or mixture of gaussians) are
66 limited in the distributional structure they may learn. To address such restrictions on expressivity,
67 advances in machine learning have used deep probability distributions as flexible approximating
68 families for such complicated distributions [36,37] (see Section 5.1.2). However, the adaptation of
69 deep probability distributions to the problem of theoretical circuit analysis requires recent devel-
70 opments in deep learning for constrained optimization [38], and architectural choices for efficient
71 and expressive deep generative modeling [39,40]. We detail our method, which we call emergent
72 property inference (EPI) in Section 3.2.
73 Equipped with this method, we demonstrate the capabilities of EPI and present novel theoretical
74 findings from its analysis. First, we show EPI’s ability to handle biologically realistic circuit models
75 using a five-neuron model of the stomatogastric ganglion [41]: a neural circuit whose parametric
76 degeneracy is closely studied [42]. Then, we show EPI’s scalability to high dimensional parameter
77 distributions by inferring connectivities of recurrent neural networks (RNNs) that exhibit stable,
78 yet amplified responses – a hallmark of neural responses throughout the brain [43–45]. In a model
79 of primary visual cortex [46,47], EPI reveals how the recurrent processing across different neuron-
80 type populations shapes excitatory variability: a finding that we show is analytically intractable.
81 Finally, we investigated the possible connectivities of a superior colliculus model that allow execu-

tion of different tasks on interleaved trials [48]. EPI discovered a rich distribution containing two connectivity regimes with different solution classes. We queried the deep probability distribution learned by EPI to produce a mechanistic understanding of neural responses in each regime. Intriguingly, the inferred connectivities of each regime reproduced results from optogenetic inactivation experiments in markedly different ways. These theoretical insights afforded by EPI illustrate the value of deep inference for the interrogation of neural circuit models.

3 Results

3.1 Motivating emergent property inference of theoretical models

Consideration of the typical workflow of theoretical modeling clarifies the need for emergent property inference. First, one designs or chooses an existing circuit model that, it is hypothesized, captures the computation of interest. To ground this process in a well-known example, consider the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic muscle activation patterns for digestion [49]. Despite full knowledge of STG connectivity and a precise characterization of its rhythmic pattern generation, biophysical models of the STG have complicated relationships between circuit parameters and computation [15, 42].

A subcircuit model of the STG [41] is shown schematically in Figure 1A. The fast population (f_1 and f_2) represents the subnetwork generating the pyloric rhythm and the slow population (s_1 and s_2) represents the subnetwork of the gastric mill rhythm. The two fast neurons mutually inhibit one another, and spike at a greater frequency than the mutually inhibiting slow neurons. The hub neuron couples with either the fast or slow population, or both depending on modulatory conditions. The jagged connections indicate electrical coupling having electrical conductance g_{el} , smooth connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto the hub neuron, and $g_{synB} = 5nS$ for mutual inhibitory connections. Note that the behavior of this model will be critically dependent on its parameterization – the choices of conductance parameters $\mathbf{z} = [g_{el}, g_{synA}]$.

Second, once the model is selected, one must specify what the model should produce. In this STG model, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model (Fig. 1B). An emergent property studied by Gutierrez et al. is the hub neuron firing at an intermediate frequency between the intrinsic spiking rates of the fast and slow populations. This emergent property (EP) is shown in Figure 1C at an average frequency of 0.55Hz. To be

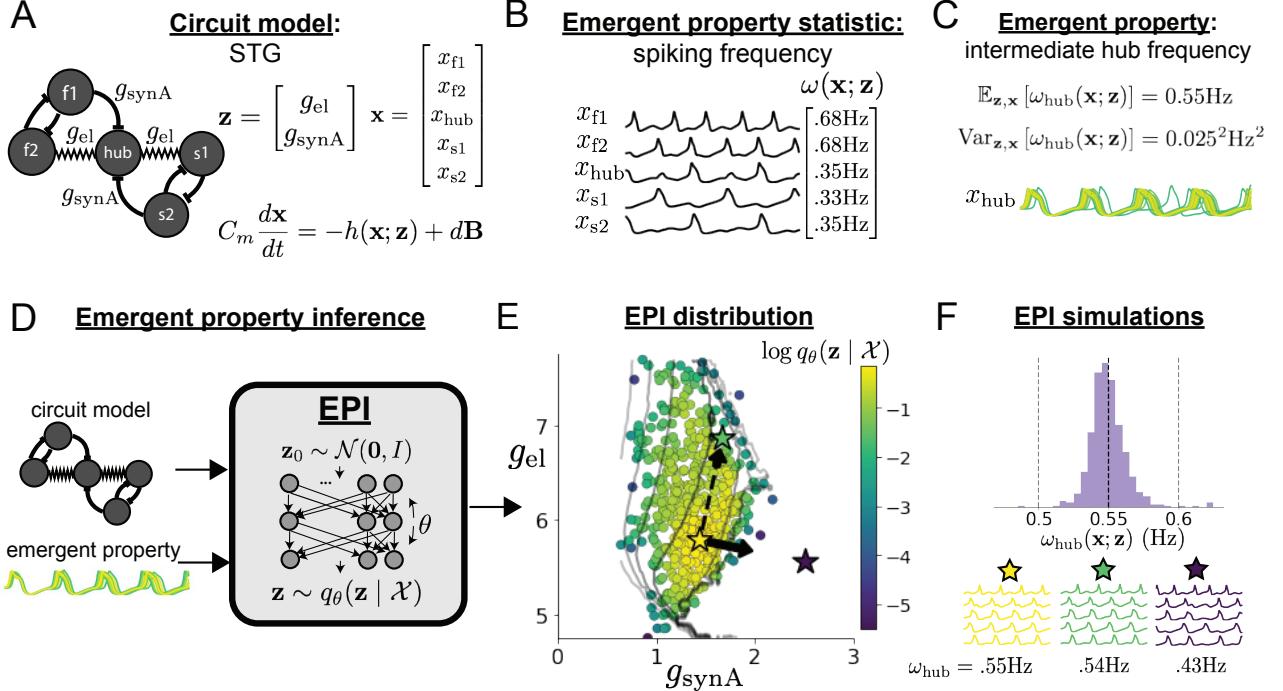


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based subcircuit model of the STG. **B.** Spiking frequency $\omega(\mathbf{x}; \mathbf{z})$ is an emergent property statistic. Simulated at $g_{el} = 4.5 \text{nS}$ and $g_{synA} = 3 \text{nS}$. **C.** The emergent property of intermediate hub frequency. Simulated activity traces are colored by log probability of generating parameters in the EPI distribution (Panel E). **D.** For a choice of circuit model and emergent property, emergent property inference (EPI) learns a deep probability distribution of parameters \mathbf{z} . **E.** The EPI distribution producing intermediate hub frequency. Samples are colored by log probability density. Contours of hub neuron frequency error are shown at levels of $.525, .53, \dots, .575 \text{ Hz}$ (dark to light gray away from mean). Dimension of sensitivity \mathbf{v}_1 (solid arrow) and robustness \mathbf{v}_2 (dashed arrow). **F** (Top) The predictions of the EPI distribution. The black and gray dashed lines show the mean and two standard deviations according the emergent property. (Bottom) Simulations at the starred parameter values.

112 precise, we define intermediate hub frequency not strictly as 0.55Hz, but frequencies of moderate
113 deviation from 0.55Hz between the fast (.35Hz) and slow (.68Hz) frequencies.

114 Third, the model parameters producing the emergent property are inferred. By precisely quantify-
115 ing the emergent property of interest as a statistical feature of the model, we use EPI to condition
116 directly on this emergent property. Before presenting technical details (in the following section), let
117 us understand emergent property inference schematically. EPI (Fig. 1D) takes, as input, the model
118 and the specified emergent property, and as its output, returns the parameter distribution (Fig.
119 1E). This distribution – represented for clarity as samples from the distribution – is a parameter
120 distribution constrained such that the circuit model produces the emergent property. Once EPI
121 is run, the returned distribution can be used to efficiently generate additional parameter samples.
122 Most importantly, the inferred distribution can be efficiently queried to quantify the parametric
123 structure that it captures. By quantifying the parametric structure governing the emergent prop-
124 erty, EPI informs the central question of this inverse problem: what aspects or combinations of
125 model parameters have the desired emergent property?

126 3.2 Emergent property inference via deep generative models

127 Emergent property inference (EPI) formalizes the three-step procedure of the previous section
128 with deep probability distributions [36, 37]. First, as is typical, we consider the model as a
129 coupled set of noisy differential equations. In this STG example, the model activity (or state)
130 $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$ is the membrane potential for each neuron, which evolves according to
131 the biophysical conductance-based equation:

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

132 where $C_m = 1\text{nF}$, and \mathbf{h} is a sum of the leak, calcium, potassium, hyperpolarization, electrical, and
133 synaptic currents, all of which have their own complicated dependence on activity \mathbf{x} and parameters
134 $\mathbf{z} = [g_{el}, g_{synA}]$, and $d\mathbf{B}$ is white gaussian noise [41] (see Section 5.2.1 for more detail).

135 Second, we determine that our model should produce the emergent property of “intermediate hub
136 frequency” (Figure 1C). We stipulate that the hub neuron’s spiking frequency – denoted by statistic
137 $\omega_{hub}(\mathbf{x})$ – is close to a frequency of 0.55Hz, between that of the slow and fast frequencies. Mathe-
138 matically, we define this emergent property with two constraints: that the mean hub frequency is
139 0.55Hz,

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}} [\omega_{hub}(\mathbf{x}; \mathbf{z})] = 0.55 \quad (2)$$

140 and that the variance of the hub frequency is moderate

$$\text{Var}_{\mathbf{z}, \mathbf{x}} [\omega_{\text{hub}}(\mathbf{x}; \mathbf{z})] = 0.025^2 \quad (3)$$

141 In the emergent property of intermediate hub frequency, the statistic of hub neuron frequency is
142 an expectation over the distribution of parameters \mathbf{z} and the distribution of the data \mathbf{x} that those
143 parameters produce. We define the EP as the collection of these two constraints. In general, an
144 emergent property is a collection of constraints on statistical moments that together define the
145 computation of interest.

146 Third, we perform emergent property inference: we find a distribution over parameter configura-
147 tions \mathbf{z} of models that produce the emergent property; in other words, they satisfy the constraints
148 introduced in Equations 2 and 3. This distribution will be chosen from a family of probability
149 distributions $\mathcal{Q} = \{q_{\theta}(\mathbf{z}) : \theta \in \Theta\}$, defined by a deep neural network [36, 37] (Figure 1D, EPI box).

150 Deep probability distributions map a simple random variable \mathbf{z}_0 (e.g. an isotropic gaussian) through
151 a deep neural network with weights and biases θ to parameters $\mathbf{z} = g_{\theta}(\mathbf{z}_0)$ of a suitably compli-
152 cated distribution (see Section 5.1.2 for more details). Many distributions in \mathcal{Q} will respect the
153 emergent property constraints, so we select the most random (highest entropy) distribution, which
154 also means this approach is equivalent to bayesian variational inference (see Section 5.1.6). In EPI
155 optimization, stochastic gradient steps in θ are taken such that entropy is maximized, and the
156 emergent property \mathcal{X} is produced (see Section 5.1). We then denote the inferred EPI distribution
157 as $q_{\theta}(\mathbf{z} | \mathcal{X})$, since the structure of the learned parameter distribution is determined by weights
158 and biases θ , and this distribution is conditioned upon emergent property \mathcal{X} .

159 The structure of the inferred parameter distributions of EPI can be analyzed to reveal key infor-
160 mation about how the circuit model produces the emergent property. As probability in the EPI
161 distribution decreases away from the mode of $q_{\theta}(\mathbf{z} | \mathcal{X})$ (Fig. 1E yellow star), the emergent prop-
162 erty deteriorates. Perturbing \mathbf{z} along a dimension in which $q_{\theta}(\mathbf{z} | \mathcal{X})$ changes little will not disturb
163 the emergent property, making this parameter combination *robust* with respect to the emergent
164 property. In contrast, if \mathbf{z} is perturbed along a dimension with strongly decreasing $q_{\theta}(\mathbf{z} | \mathcal{X})$,
165 that parameter combination is deemed *sensitive* [27, 30]. By querying the second order derivative
166 (Hessian) of $\log q_{\theta}(\mathbf{z} | \mathcal{X})$ at a mode, we can quantitatively identify how sensitive (or robust) each
167 eigenvector is by its eigenvalue; the more negative, the more sensitive and the closer to zero, the
168 more robust (see Section 5.2.4). Indeed, samples equidistant from the mode along these dimensions
169 of sensitivity (\mathbf{v}_1 , smaller eigenvalue) and robustness (\mathbf{v}_2 , greater eigenvalue) (Fig. 1E, arrows)
170 agree with error contours (Fig. 1E contours) and have diminished or preserved hub frequency, re-

171 spectsively (Fig. 1F activity traces). The directionality of \mathbf{v}_2 suggests that changes in conductance
 172 along this parameter combination will most preserve hub neuron firing between the intrinsic rates
 173 of the pyloric and gastric mill rhythms. Importantly and unlike alternative techniques, once an
 174 EPI distribution has been learned, the modes and Hessians of the distribution can be measured
 175 with trivial computation (see Section 5.1.2).

176 In the following sections, we demonstrate EPI on three neural circuit models across ranges of
 177 biological realism, neural system function, and network scale. First, we demonstrate the superior
 178 scalability of EPI compared to alternative techniques by inferring high-dimensional distributions
 179 of recurrent neural network connectivities that exhibit amplified, yet stable responses. Next, in a
 180 model of primary visual cortex [46,47], we show how EPI discovers parametric degeneracy, revealing
 181 how input variability across neuron types affects the excitatory population. Finally, in a model of
 182 superior colliculus [48], we used EPI to capture multiple parametric regimes of task switching, and
 183 queried the dimensions of parameter sensitivity to characterize each regime.

184 **3.3 Scaling inference of recurrent neural network connectivity with EPI**

185 To understand how EPI scales in comparison to existing techniques, we consider recurrent neu-
 186 ral networks (RNNs). Transient amplification is a hallmark of neural activity throughout cortex,
 187 and is often thought to be intrinsically generated by recurrent connectivity in the responding cor-
 188 tical area [43–45]. It has been shown that to generate such amplified, yet stabilized responses,
 189 the connectivity of RNNs must be non-normal [43, 50], and satisfy additional constraints [51]. In
 190 theoretical neuroscience, RNNs are optimized and then examined to show how dynamical systems
 191 could execute a given computation [52, 53], but such biologically realistic constraints on connec-
 192 tivity [43, 50, 51] are ignored for simplicity or because constrained optimization is difficult. In
 193 general, access to distributions of connectivity that produce theoretical criteria like stable amplifi-
 194 cation, chaotic fluctuations [10], or low tangling [54] would add scientific value to existing research
 195 with RNNs. Here, we use EPI to learn RNN connectivities producing stable amplification, and
 196 demonstrate the superior scalability and efficiency of EPI to alternative approaches.

197 We consider a rank-2 RNN with N neurons having connectivity $W = UV^\top$ and dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}, \quad (4)$$

198 where $U = [\mathbf{U}_1 \ \mathbf{U}_2] + g\chi^{(U)}$, $V = [\mathbf{V}_1 \ \mathbf{V}_2] + g\chi^{(V)}$, $\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}_1, \mathbf{V}_2 \in [-1, 1]^N$, and $\chi_{i,j}^{(U)}, \chi_{i,j}^{(V)} \sim$
 199 $\mathcal{N}(0, 1)$. We infer connectivity parameters $\mathbf{z} = [\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}_1, \mathbf{V}_2]$ that produce stable amplification.

200 Two conditions are necessary and sufficient for RNNs to exhibit stable amplification [51]: $\text{real}(\lambda_1) <$
 201 1 and $\lambda_1^s > 1$, where λ_1 is the eigenvalue of W with greatest real part and λ^s is the maximum
 202 eigenvalue of $W^s = \frac{W+W^\top}{2}$. RNNs with $\text{real}(\lambda_1) = 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$ will be stable with
 203 modest decay rate ($\text{real}(\lambda_1)$ close to its upper bound of 1) and exhibit modest amplification (λ_1^s
 204 close to its lower bound of 1). EPI can naturally condition on this emergent property

$$\begin{aligned}\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}.\end{aligned}\quad (5)$$

205 Variance constraints predicate that the majority of the distribution (within two standard devia-
 206 tions) are within the specified ranges.

207 For comparison, we infer the parameters \mathbf{z} likely to produce stable amplification using two al-
 208 ternative simulation-based inference approaches. Sequential Monte Carlo approximate bayesian
 209 computation (SMC-ABC) [26] is a rejection sampling approach that uses SMC techniques to im-
 210 prove efficiency, and sequential neural posterior estimation (SNPE) [35] approximates posteriors
 211 with deep probability distributions (see Section 5.1.1). Unlike EPI, these statistical inference tech-
 212 niques do not constrain the predictions of the inferred distribution, so they were run by conditioning
 213 on an exemplar dataset $\mathbf{x}_0 = \boldsymbol{\mu}$, following standard practice with these methods [26, 35]. To com-
 214 pare the efficiency of these different techniques, we measured the time and number of simulations
 215 necessary for the distance of the predictive mean to be less than 0.5 from $\boldsymbol{\mu} = \mathbf{x}_0$ (see Section 5.3).

216 As the number of neurons N in the RNN, and thus the dimension of the parameter space $\mathbf{z} \in$
 217 $[-1, 1]^{4N}$, is scaled, we see that EPI converges at greater speed and at greater dimension than
 218 SMC-ABC and SNPE (Fig. 2A). It also becomes most efficient to use EPI in terms of simulation
 219 count at $N = 50$ (Fig. 2B). It is well known that ABC techniques struggle in parameter spaces
 220 of modest dimension [55], yet we were careful to assess the scalability of SNPE, which is a more
 221 closely related methodology to EPI. Between EPI and SNPE, we closely controlled the number of
 222 parameters in deep probability distributions by dimensionality (Fig. S5), and tested more aggressive
 223 SNPE hyperparameter choices when SNPE failed to converge (Fig. S6). In this analysis, we see that
 224 deep inference techniques EPI and SNPE are far more amenable to inference of high dimensional
 225 RNN connectivities than rejection sampling techniques like SMC-ABC, and that EPI outperforms
 226 SNPE in both wall time (elapsed real time) and simulation count.

227 No matter the number of neurons, EPI always produces connectivity distributions with mean and

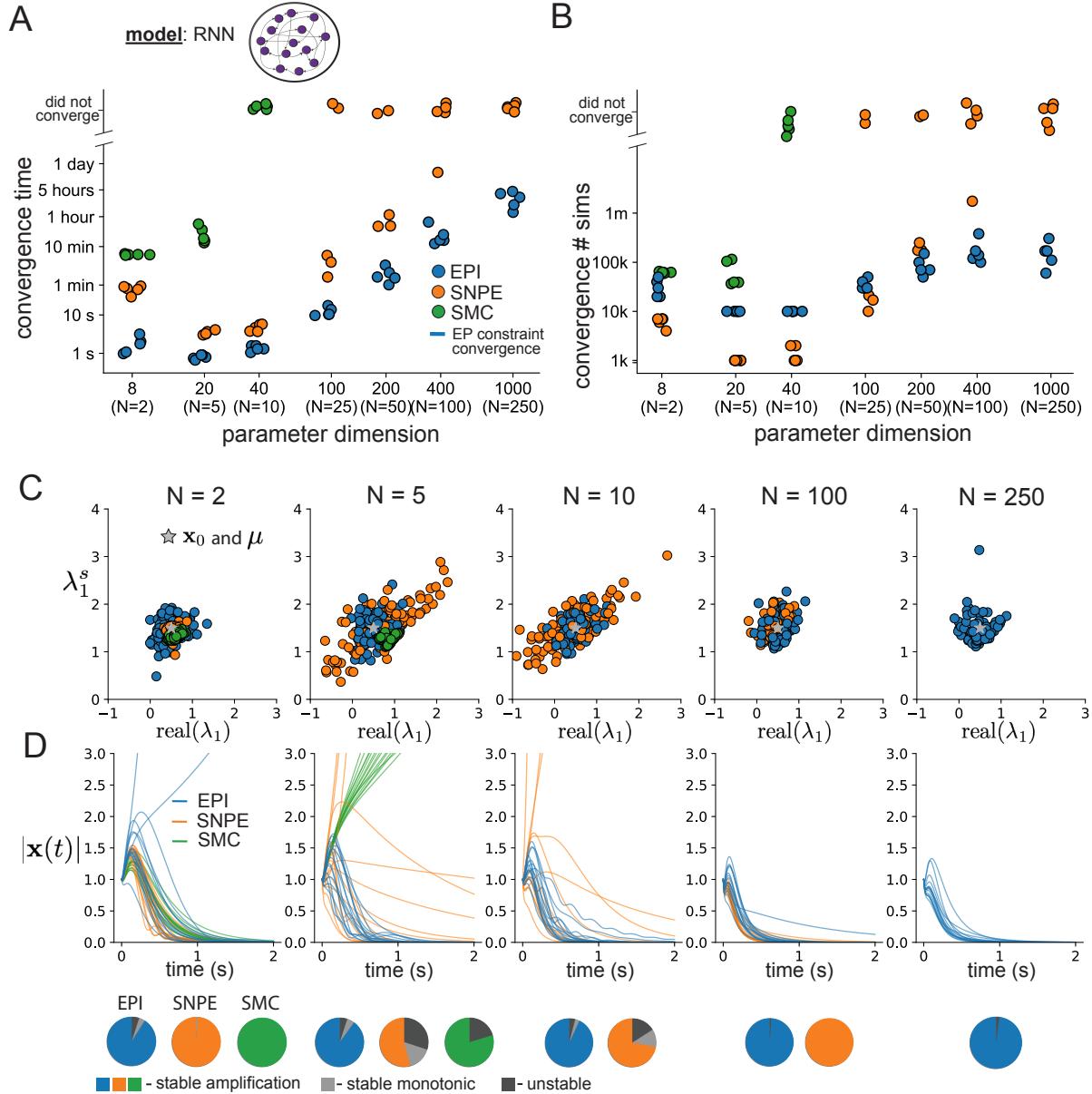


Figure 2: **A.** Wall time of EPI (blue), SNPE (orange), and SMC-ABC (green) to converge on RNN connectivities producing stable amplification. Each dot shows convergence time for an individual random seed. For reference, the mean wall time for EPI to achieve its full constraint convergence (means and variances) is shown (blue line). **B.** Simulation count of each algorithm to achieve convergence. Same conventions as A. **C.** The predictive distributions of connectivities inferred by EPI (blue), SNPE (orange), and SMC-ABC (green), with reference to $x_0 = \mu$ (gray star). **D.** Simulations of networks inferred by each method ($\tau = 100ms$). Each trace (15 per algorithm) corresponds to simulation of one z . (Below) Ratio of obtained samples producing stable amplification, monotonic decay, and instability.

variance of $\text{real}(\lambda_1)$ and λ_1^s according to \mathcal{X} (Fig. 2C, blue). For the dimensionalities in which SMC-ABC is tractable, the inferred parameters are concentrated and offset from the exemplary dataset \mathbf{x}_0 (Fig. 2C, green). When using SNPE, the predictions of the inferred parameters are highly concentrated at some RNN sizes and widely varied in others (Fig. 2C, orange). We see these properties reflected in simulations from the inferred distributions: EPI produces a consistent variety of stable, amplified activity norms $|\mathbf{x}(t)|$, SMC-ABC produces a limited variety of responses, and the changing variety of responses from SNPE emphasizes the control of EPI on parameter predictions (Fig. 2D). Even for moderate neuron counts, the predictions of the inferred distribution of SNPE are highly dependent on N and g , while EPI maintains the emergent property across choices of RNN (see Section 5.3.5).

To understand these differences, note that EPI outperforms SNPE in high dimensions by using gradient information (from $\nabla_{\mathbf{z}}f(\mathbf{x}; \mathbf{z}) = \nabla_{\mathbf{z}}[\text{real}(\lambda_1), \lambda_1^s]^\top$). This choice agrees with recent speculation that such gradient information could improve the efficiency of simulation-based inference techniques [56], as well as reflecting the classic tradeoff between gradient-based and sampling-based estimators (scaling and speed versus generality). Since gradients of the emergent property are necessary in EPI optimization, gradient tractability is a key criteria when determining the suitability of a simulation-based inference technique. If the emergent property gradient is efficiently calculated, EPI is a clear choice for inferring high dimensional parameter distributions. In the next two sections, we use EPI for novel scientific insight by examining the structure of inferred distributions.

3.4 EPI reveals how recurrence with multiple inhibitory subtypes governs excitatory variability in a V1 model

Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output function have succeeded in explaining a host of experimentally documented phenomena in primary visual cortex (V1). In a regime characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to paradoxical responses [12], selective amplification [43, 50], surround suppression [57] and normalization [58]. Recent theoretical work [59] shows that stabilized E-I models reproduce the effect of variability suppression [60]. Furthermore, experimental evidence shows that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) – composing 80% of GABAergic interneurons in V1 [61–63], and that these inhibitory cell types follow specific connectivity patterns (Fig. 3A) [64]. Here, we use EPI on a model of V1 with biologically realistic connectivity to show how the structure of input across neuron types affects

259 the variability of the excitatory population – the population largely responsible for projecting to
 260 other brain areas [65].

261 We considered response variability of a nonlinear dynamical V1 circuit model (Fig. 3A) with a state
 262 comprised of each neuron-type population’s rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population receives
 263 recurrent input $W\mathbf{x}$, where W is the effective connectivity matrix (see Section 5.4) and an external
 264 input with mean \mathbf{h} , which determines population rate via supralinear nonlinearity $\phi(\cdot) = [\cdot]_+^2$. The
 265 external input has an additive noisy component ϵ with variance $\sigma^2 = [\sigma_E^2, \sigma_P^2, \sigma_S^2, \sigma_V^2]$. This noise
 266 has a slower dynamical timescale $\tau_{\text{noise}} > \tau$ than the population rate, allowing fluctuations around
 267 a stimulus-dependent steady-state (Fig. 3B). This model is the stochastic stabilized supralinear
 268 network (SSSN) [59]

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \epsilon), \quad (6)$$

269 generalized to have multiple inhibitory neuron types. It introduces stochasticity to four neuron-
 270 type models of V1 [46]. Stochasticity and inhibitory multiplicity introduce substantial complexity
 271 to the mathematical treatment of this problem (see Section 5.4.5) motivating the analysis of this
 272 model with EPI. Here, we consider fixed weights W and input \mathbf{h} [47], and study the effect of input
 273 variability $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on excitatory variability.

274 We quantify levels of E-population variability by studying two emergent properties

$$\begin{aligned} \mathcal{X}(5\text{Hz}) : \mathbb{E}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= 5\text{Hz} & \mathcal{X}(10\text{Hz}) : \mathbb{E}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= 10\text{Hz} \\ \text{Var}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= 1\text{Hz}^2 & \text{Var}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= 1\text{Hz}^2, \end{aligned} \quad (7)$$

275 where $s_E(\mathbf{x}; \mathbf{z})$ is the standard deviation of the stochastic E -population response about its steady
 276 state (Fig. 3C). In the following analyses, we select 1Hz^2 variance such that the two emergent
 277 properties do not overlap in $s_E(\mathbf{z}; \mathbf{x})$.

278 First, we ran EPI to obtain parameter distribution $q_{\theta}(\mathbf{z} \mid \mathcal{X}(5\text{Hz}))$ producing E-population vari-
 279 ability around 5Hz (Fig. 3D). From the marginal distribution of σ_E and σ_P (Fig. 3D, top-left),
 280 we can see that $s_E(\mathbf{x}; \mathbf{z})$ is sensitive to various combinations of σ_E and σ_P . Alternatively, both σ_S
 281 and σ_V are degenerate with respect to $s_E(\mathbf{x}; \mathbf{z})$ evidenced by the unexpectedly high variability in
 282 those dimensions (Fig. 3D, bottom-right). Together, these observations imply a curved path with
 283 respect to $s_E(\mathbf{x}; \mathbf{z})$ of 5Hz, which is indicated by the modes along σ_P (Fig. 3E).

284 Figure 3E suggests a quadratic relationship in E-population fluctuations and the standard deviation
 285 of E- and P-population input; as the square of either σ_E or σ_P increases, the other compensates by
 286 decreasing to preserve the level of $s_E(\mathbf{x}; \mathbf{z})$. This quadratic relationship is preserved at greater level

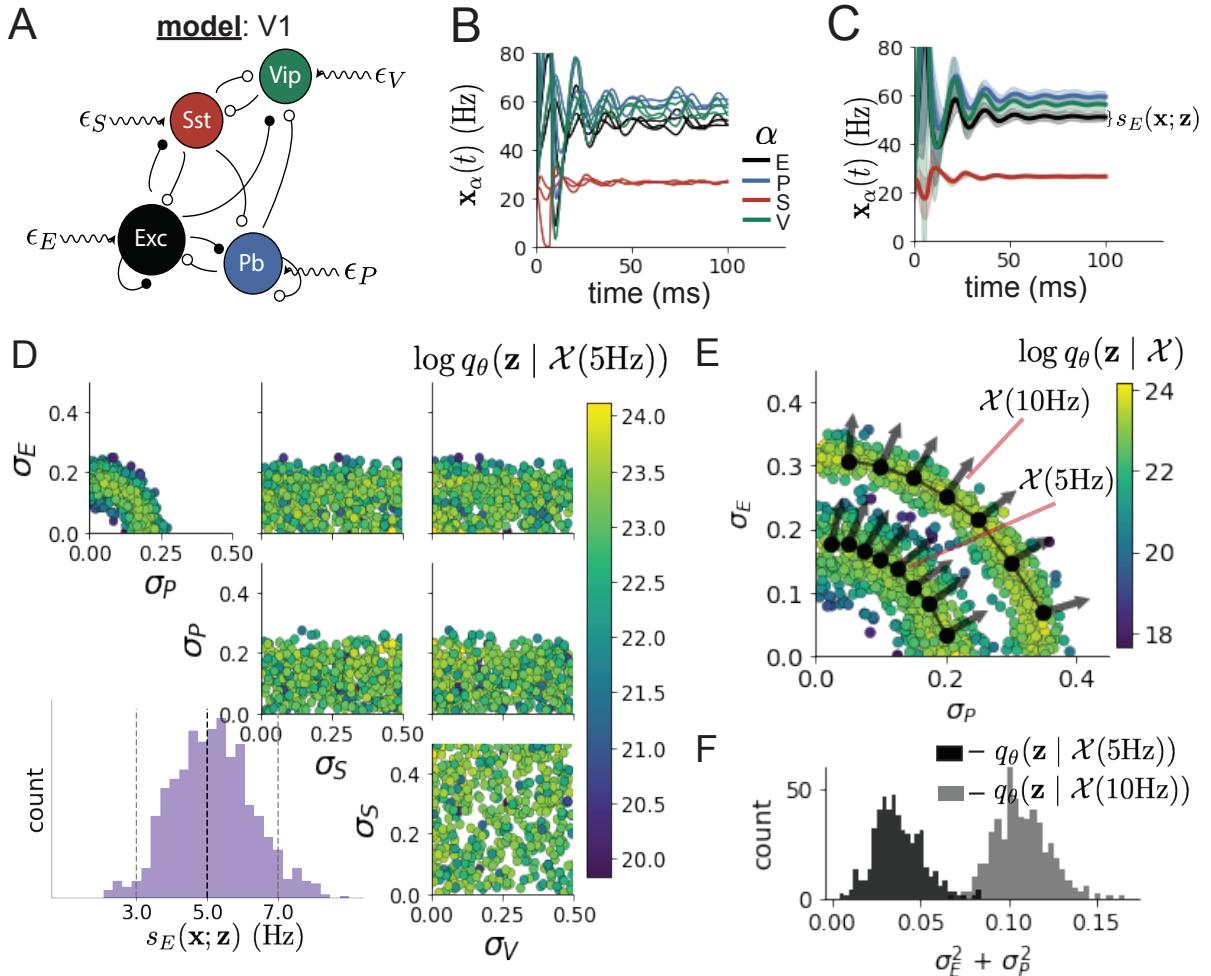


Figure 3: Emergent property inference in the stochastic stabilized supralinear network (SSSN)

A. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). Some neuron-types largely do not form synaptic projections to others ($|W_{\alpha_1, \alpha_2}| < 0.025$). Each neural population receives a baseline input \mathbf{h}_b , and the E- and P- populations also receive a contrast-dependent input \mathbf{h}_c . Additionally, each neural population receives a slow noisy input ϵ . **B.** Transient network responses of the SSSN model. Traces are independent trials with varying initialization $\mathbf{x}(0)$ and noise ϵ . **C.** Mean (solid line) and standard deviation $s_E(\mathbf{x}; \mathbf{z})$ (shading) across 100 trials. **D.** EPI distribution of noise parameters \mathbf{z} conditioned on E-population variability. The EPI predictive distribution of $s_E(\mathbf{x}; \mathbf{z})$ is show on the bottom-left. **E.** (Top) Enlarged visualization of the σ_E - σ_P marginal distribution of EPI $q_\theta(\mathbf{z} | \mathcal{X}(5\text{Hz}))$ and $q_\theta(\mathbf{z} | \mathcal{X}(10\text{Hz}))$. Each black dot shows the mode at each σ_P . The arrows show the most sensitive dimensions of the Hessian evaluated at these modes. **F.** The predictive distributions of $\sigma_E^2 + \sigma_P^2$ of each inferred distribution $q_\theta(\mathbf{z} | \mathcal{X}(5\text{Hz}))$ and $q_\theta(\mathbf{z} | \mathcal{X}(10\text{Hz}))$.

287 of E-population variability $\mathcal{X}(10\text{Hz})$ (Fig. 3E and S8). Indeed, the sum of squares of σ_E and σ_P is
288 larger in $q_{\theta}(\mathbf{z} \mid \mathcal{X}(10\text{Hz}))$ than $q_{\theta}(\mathbf{z} \mid \mathcal{X}(5\text{Hz}))$ (Fig 3F, $p < 1 \times 10^{-10}$), while the sum of squares of
289 σ_S and σ_V are not significantly different in the two EPI distributions (Fig. S10, $p = .40$), in which
290 parameters were bounded from 0 to 0.5. The strong interaction between E- and P-population input
291 variability on excitatory variability is intriguing, since this circuit exhibits a paradoxical effect in
292 the P-population (and no other inhibitory types) (Fig. S11), meaning that the E-population is
293 P-stabilized. Future research may uncover a link between the population of network stabilization
294 and compensatory interactions governing excitatory variability.

295 EPI revealed the quadratic dependence of excitatory variability on input variability to the E- and
296 P-populations, as well as its independence to input from the other two inhibitory populations.
297 In a simplified model ($\tau = \tau_{\text{noise}}$), it can be shown that surfaces of equal variance are ellipsoids
298 as a function of σ (see Section 5.4.5). Nevertheless, the sensitive and degenerate parameters are
299 intractable to predict mathematically, since the covariance matrix depends on the steady-state
300 solution of the network [59, 66], and terms in the covariance expression increase quadratically with
301 each additional neuron-type population (see also Section 5.4.5). By pointing out this mathematical
302 complexity, we emphasize the value of EPI for gaining understanding about theoretical models
303 when mathematical analysis becomes onerous or impractical.

304 3.5 EPI identifies two regimes of rapid task switching

305 It has been shown that rats can learn to switch from one behavioral task to the next on randomly
306 interleaved trials [67], and an important question is what neural mechanisms produce this compu-
307 tation. In this experimental setup, rats were given an explicit task cue on each trial, either Pro
308 or Anti. After a delay period, rats were shown a stimulus, and made a context (task) dependent
309 response (Fig. 4A). In the Pro task, rats were required to orient towards the stimulus, while in
310 the Anti task, rats were required to orient away from the stimulus. Pharmacological inactivation
311 of the SC impaired rat performance, and time-specific optogenetic inactivation revealed a crucial
312 role for the SC on the cognitively demanding Anti trials [48]. These results motivated a nonlinear
313 dynamical model of the SC containing four functionally-defined neuron-type populations. In Duan
314 et al. 2019, a computationally intensive procedure was used to obtain a set of 373 connectivity
315 parameters that qualitatively reproduced these optogenetic inactivation results. To build upon
316 the insights of this previous work, we use the probabilistic tools afforded by EPI to identify and
317 characterize two linked, yet distinct regimes of rapid task switching connectivity.

318 In this SC model, there are Pro- and Anti-populations in each hemisphere (left (L) and right (R))
 319 with activity variables $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$ [48]. The connectivity of these populations is
 320 parameterized by self sW , vertical vW , diagonal dW and horizontal hW connections (Fig. 4B). The
 321 input \mathbf{h} is comprised of a positive cue-dependent signal to the Pro or Anti populations, a positive
 322 stimulus-dependent input to either the Left or Right populations, and a choice-period input to the
 323 entire network (see Section 5.5.1). Model responses are bounded from 0 to 1 as a function ϕ of an
 324 internal variable \mathbf{u}

$$\begin{aligned}\tau \frac{d\mathbf{u}}{dt} &= -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B} \\ \mathbf{x} &= \phi(\mathbf{u}).\end{aligned}\tag{8}$$

325 The model responds to the side with greater Pro neuron activation; e.g. the response is left if
 326 $x_{LP} > x_{RP}$ at the end of the trial. Here, we use EPI to determine the network connectivity
 327 $\mathbf{z} = [sW, vW, dW, hW]^\top$ that produces rapid task switching.
 328 Rapid task switching is formalized mathematically as an emergent property with two statistics:
 329 accuracy in the Pro task $p_P(\mathbf{x}; \mathbf{z})$ and Anti task $p_A(\mathbf{x}; \mathbf{z})$. We stipulate that accuracy be on average
 330 .75 in each task with variance .075²

$$\begin{aligned}\mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .75 \\ .75 \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .075^2 \\ .075^2 \end{bmatrix}.\end{aligned}\tag{9}$$

331 75% accuracy is a realistic level of performance in each task, and with the chosen variance, inferred
 332 models will not exhibit fully random responses (50%), nor perfect performance (100%).
 333 The EPI inferred distribution (Fig. 4C) produces Pro and Anti task accuracies (Fig. 4C, bottom-left)
 334 consistent with rapid task switching (Equation 9). This parameter distribution has rich structure
 335 that is not captured well by simple linear correlations (Fig. S12). Specifically, the shape
 336 of the EPI distribution is sharply bent, matching ground truth structure indicated by brute-force
 337 sampling (Fig. S18). This is most saliently observed in the marginal distribution of $sW-hW$ (Fig.
 338 4C top-right), where anticorrelation between sW and hW switches to correlation with decreasing
 339 sW . By identifying the modes of the EPI distribution $\mathbf{z}^*(sW)$ at different values of sW (Fig. 4C
 340 red/purple dots), we can quantify this change in distributional structure with the sensitivity dimension
 341 $\mathbf{v}_1(\mathbf{z})$ (Fig. 4C red/purple arrows). Note that the directionality of these sensitivity dimensions
 342 at $\mathbf{z}^*(sW)$ changes distinctly with sW , and are perpendicular to the robust dimensions of the EPI

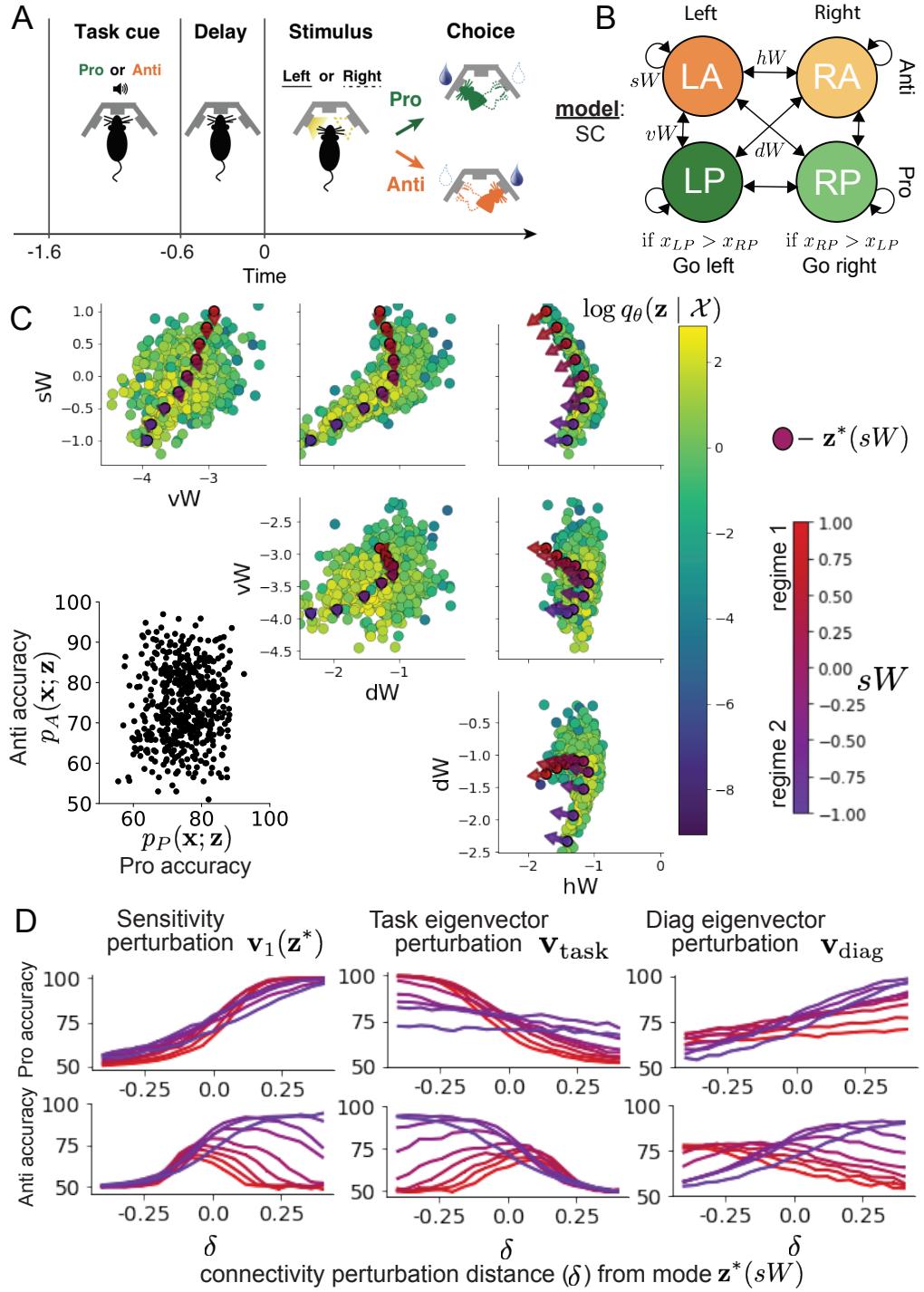


Figure 4: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - Left Pro, RP - Right Pro, LA - Left Anti, RA - Right Anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. **C.** The EPI inferred distribution of rapid task switching networks. Red/purple parameters indicate modes $\mathbf{z}^*(sW)$ colored by sW . Sensitivity vectors $\mathbf{v}_1(\mathbf{z}^*)$ are shown by arrows. (Bottom-left) EPI predictive distribution of task accuracies. **D.** Mean and standard error ($N_{\text{test}} = 25$, bars not visible) of accuracy in Pro (top) and Anti (bottom) tasks after perturbing connectivity away from mode along $\mathbf{v}_1(\mathbf{z}^*)$ (left), \mathbf{v}_{task} (middle), and \mathbf{v}_{diag} (right).

343 distribution that preserve rapid task switching. These two directionalities of sensitivity motivate
344 the distinction of connectivity into two regimes, which produce different types of responses in the
345 Pro and Anti tasks (Fig. S13).

346 When perturbing connectivity along the sensitivity dimension away from the modes

$$\mathbf{z} = \mathbf{z}^*(sW) + \delta\mathbf{v}_1(\mathbf{z}^*(sW)), \quad (10)$$

347 Pro accuracy monotonically increases in both regimes (Fig. 4D, top-left). However, there is a stark
348 difference between regimes in Anti accuracy. Anti accuracy falls in either direction of \mathbf{v}_1 in regime 1,
349 yet monotonically increases along with Pro accuracy in regime 2 (Fig. 4D, bottom-left). The sharp
350 change in local structure of the EPI distribution is therefore explained by distinct sensitivities:
351 Anti accuracy diminishes in only one or both directions of the sensitivity perturbation.

352 To understand the mechanisms differentiating the two regimes, we can make connectivity pertur-
353 bations along dimensions that only modify a single eigenvalue of the connectivity matrix. These
354 eigenvalues λ_{all} , λ_{side} , λ_{task} , and λ_{diag} correspond to connectivity eigenmodes with intuitive roles
355 in processing in this task (Fig. S14A). For example, greater λ_{task} will strengthen internal repre-
356 sentations of task, while greater λ_{diag} will amplify dominance of Pro and Anti pairs in opposite
357 hemispheres (Section 5.5.7). Unlike the sensitivity dimension, the dimensions \mathbf{v}_a that perturb
358 isolated connectivity eigenvalues λ_a for $a \in \{\text{all}, \text{side}, \text{task}, \text{diag}\}$ are independent of $\mathbf{z}^*(sW)$ (see
359 Section 5.5.7), e.g.

$$\mathbf{z} = \mathbf{z}^*(sW) + \delta\mathbf{v}_{\text{task}}. \quad (11)$$

360 Connectivity perturbation analyses reveal that decreasing λ_{task} has a very similar effect on Anti
361 accuracy as perturbations along the sensitivity dimension (Fig. 4D, middle). The similar effects
362 of perturbations along the sensitivity dimension $\mathbf{v}_1(\mathbf{z}^*)$ and reduction of task eigenvalue (via per-
363 turbations along $-\mathbf{v}_{\text{task}}$) suggest that there is a carefully tuned strength of task representation in
364 connectivity regime 1, which if disturbed results in random Anti trial responses. Finally, we rec-
365 ognize that increasing λ_{diag} has opposite effects on Anti accuracy in each regime (Fig. 4D, right).
366 In the next section, we build on these mechanistic characterizations of each regime by examining
367 their resilience to optogenetic inactivation.

368 **3.6 EPI inferred SC connectivities reproduce results from optogenetic inacti-**
369 **vation experiments**

370 During the delay period of this task, the circuit must prepare to execute the correct task according
371 to the presented cue. The circuit must then maintain a representation of task throughout the delay
372 period, which is important for correct execution of the Anti task. Duan et al. found that bilateral
373 optogenetic inactivation of SC during the delay period consistently decreased performance in the
374 Anti task, but had no effect on the Pro task (Fig. 5A) [48]. The distribution of connectivities
375 inferred by EPI exhibited this same effect in simulation at high optogenetic strengths γ , which
376 reduce the network activities $\mathbf{x}(t)$ by a factor $1 - \gamma$ (Fig. 5B) (see Section 5.5.8).

377 To examine how connectivity affects response to delay period inactivation, we grouped connectivi-
378 ties of the EPI distribution along the continuum linking regimes 1 and 2 of Section 3.5:

$$Z(sW) = \{\mathbf{z} \text{ if } \underset{y \in Y}{\operatorname{argmin}} |\mathbf{z} - \mathbf{z}^*(y)|_2 = sW, \text{ for } z \sim q_{\theta}(\mathbf{z} | \mathcal{X})\}, \quad (12)$$

379 where $Y = \{-1., -0.75, ..., 1.\}$ are the values of sW for which we calculated the mode. $Z(sW)$ is
380 then the set of EPI samples for which the closest mode was $\mathbf{z}^*(sW)$ (see Section 5.5.4). In the
381 following analyses, we examine how error, and the influence of connectivity eigenvalue on Anti error
382 change along this continuum of connectivities. Obtaining the parameter samples for these analysis
383 with the learned EPI distribution was more than 20,000 times faster than a brute force approach
384 (see Section 5.5.5).

385 The mean increase in Anti error of the EPI distribution is closest to the experimentally measured
386 value of 7% at $\gamma = 0.675$ (Fig. 5B, black dot). At this level of optogenetic strength, regime
387 1 exhibits an increase in Anti error with delay period silencing (Fig. 5C, left), while regime 2
388 does not. In regime 1, greater λ_{task} and λ_{diag} decrease Anti error (Fig. 5C, right). In other words,
389 stronger task representations and diagonal amplification make the SC model more resilient to delay
390 period silencing in the Anti task. This complements the finding from Duan et al. 2019 [48] that
391 λ_{task} and λ_{diag} improve Anti accuracy.

392 At roughly $\gamma = 0.85$ (Fig. 5B, gray dot), the Anti error saturates, while Pro error remains at
393 zero. Following delay period inactivation at this optogenetic strength, there are strong similarities
394 in the responses of Pro and Anti trials during the choice period (Fig. 5D, left). We interpreted
395 these similarities to suggest that delay period inactivation at this saturated level flips the internal
396 representation of task (from Anti to Pro) in the circuit model. A flipped task representation
397 would explain why the Anti error saturates at 50%: the average Anti accuracy in EPI inferred

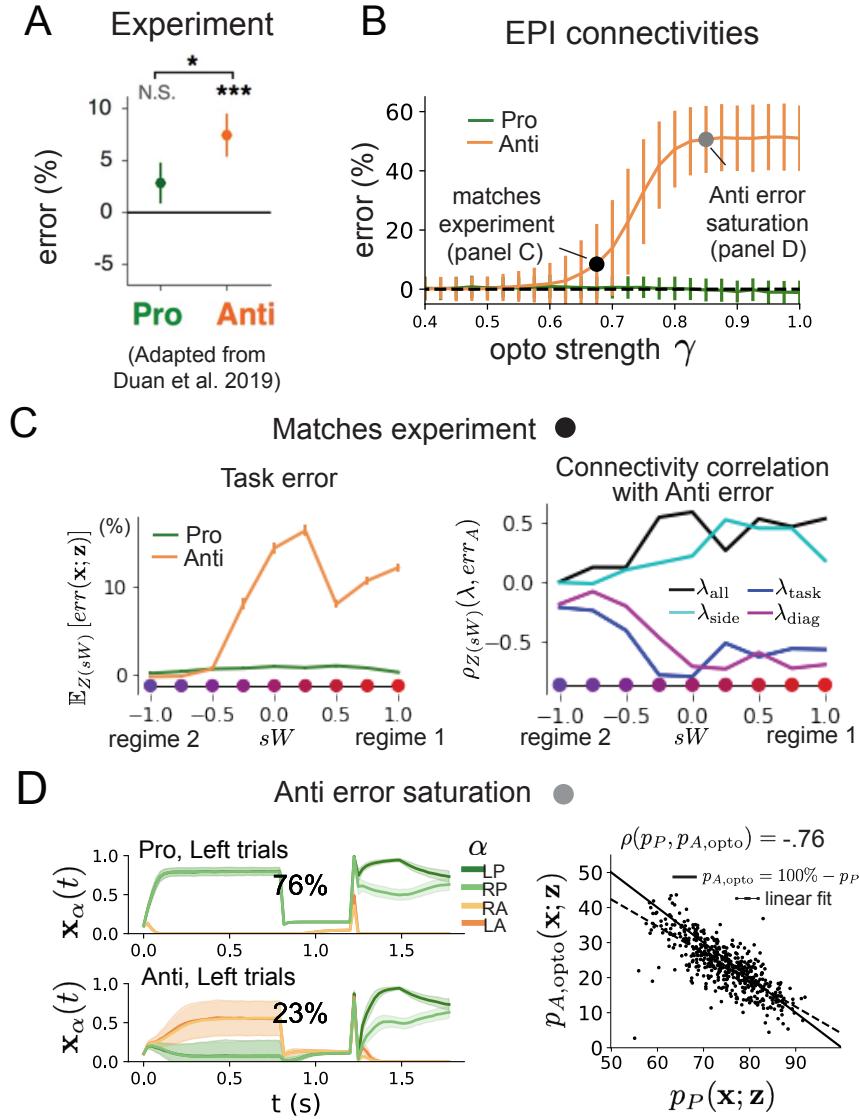


Figure 5: **A.** Mean and standard error (bars) across recording sessions of task error following delay period optogenetic inactivation in rats. **B.** Mean and standard deviation (bars) of task error induced by delay period inactivation of varying optogenetic strength γ across the EPI distribution. **C.** (Left) Mean and standard error of Pro and Anti error from regime 1 to regime 2 at $\gamma = 0.675$. (Right) Correlations of connectivity eigenvalues with Anti error from regime 1 to regime 2 at $\gamma = 0.675$. **D.** (Left) Mean and standard deviation (shading) of responses of the SC model at the mode of the EPI distribution to delay period inactivation at $\gamma = 0.85$. Accuracy in Pro (top) and Anti (bottom) task is shown as a percentage. (Right) Anti accuracy following delay period inactivation at $\gamma = 0.85$ versus accuracy in the Pro task across connectivities in the EPI distribution.

398 connectivities is 75%, but is 25% when the internal representation is flipped during delay period
399 silencing. This hypothesis prescribes a model of Anti accuracy during delay period silencing of
400 $p_{A,\text{opto}} = 100\% - p_P$, which is fit closely across both regimes of the EPI inferred connectivities (Fig.
401 5D, right). Similarities between Pro and Anti trial responses were not present at the experiment-
402 matching level of $\gamma = 0.675$ (Fig. S16 left) and neither was anticorrelation in p_P and $p_{A,\text{opto}}$ (Fig.
403 S16 right).

404 In summary, the connectivity inferred by EPI to perform rapid task switching replicated results
405 from optogenetic silencing experiments. We found that at levels of optogenetic strength matching
406 experimental levels of Anti error, only one regime actually exhibited the effect. This connectivity
407 regime is less resilient to optogenetic perturbation, and perhaps more biologically realistic. Finally,
408 we characterized the pathology in Anti error that occurs in both regimes when optogenetic strength
409 is increased to high levels, leading to a mechanistic hypothesis that is experimentally testable.
410 The probabilistic tools afforded by EPI yielded this insight: we identified two regimes and the
411 continuum of connectivities between them by taking gradients of parameter probabilities in the EPI
412 distribution, we identified sensitivity dimensions by measuring the Hessian of the EPI distribution,
413 and we obtained many parameter samples at each step along the continuum at an efficient rate.

414 4 Discussion

415 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets [20].
416 Careful inference procedures are developed for these statistical models allowing precise, quantitative
417 reasoning, which clarifies the way data informs beliefs about the model parameters. However, these
418 statistical models often lack resemblance to the underlying biology, making it unclear how to go
419 from the structure revealed by these methods, to the neural mechanisms giving rise to it. In
420 contrast, theoretical neuroscience has primarily focused on careful models of neural circuits and
421 the production of emergent properties of computation, rather than measuring structure in neural
422 datasets. In this work, we improve upon parameter inference techniques in theoretical neuroscience
423 with emergent property inference, harnessing deep learning towards parameter inference in neural
424 circuit models (see Section 5.1.1).

425 Methodology for statistical inference in circuit models has evolved considerably in recent years.
426 Early work used rejection sampling techniques [24–26], but EPI and another recently developed
427 methodology [35] employ deep learning to improve efficiency and provide flexible approximations.

428 SNPE has been used for posterior inference of parameters in circuit models conditioned upon
429 exemplar data used to represent computation, but it does not infer parameter distributions that
430 only produce the computation of interest like EPI (see Section 3.3). When strict control over the
431 predictions of the inferred parameters is necessary, EPI uses a constrained optimization technique
432 [38] (see Section 5.1.4) to make inference conditioned on the emergent property possible.

433 A key difference between EPI and SNPE, is that EPI uses gradients of the emergent property
434 throughout optimization. In Section 3.3, we showed that such gradients confer beneficial scaling
435 properties, but a concern remains that emergent property gradients may be too computationally
436 intensive. Even in a case of close biophysical realism with an expensive emergent property gradient,
437 EPI was run successfully on intermediate hub frequency in a 5-neuron subcircuit model of the
438 STG (Section 3.1). However, conditioning on the pyloric rhythm [68] in a model of the pyloric
439 subnetwork model [15] proved to be prohibitive with EPI. The pyloric subnetwork requires many
440 time steps for simulation and many key emergent property statistics (e.g. burst duration and
441 phase gap) are not calculable or easily approximated with differentiable functions. In such cases,
442 SNPE, which does not require differentiability of the emergent property, has proven useful [35].
443 In summary, choice of deep inference technique should consider emergent property complexity and
444 differentiability, dimensionality of parameter space, and the importance of constraining the model
445 behavior predicted by the inferred parameter distribution.

446 In this paper, we demonstrate the value of deep inference for parameter sensitivity analyses at
447 both the local and global level. With these techniques, flexible deep probability distributions are
448 optimized to capture global structure by approximating the full distribution of suitable parame-
449 ters. Importantly, the local structure of this deep probability distribution can be quantified at
450 any parameter choice, offering instant sensitivity measurements after fitting. For example, the
451 global structure captured by EPI revealed two distinct parameter regimes, which had different
452 local structure quantified by the deep probability distribution (see Section 5.5). In comparison,
453 bayesian MCMC is considered a popular approach for capturing global parameter structure [69],
454 but there is no variational approximation (the deep probability distribution in EPI), so sensitiv-
455 ity information is not queryable and sampling remains slow after convergence. Local sensitivity
456 analyses (e.g. [27]) may be performed independently at individual parameter samples, but these
457 methods alone do not capture the full picture in nonlinear, complex distributions. In contrast,
458 deep inference yields a probability distribution that produces a wholistic assessment of parameter
459 sensitivity at the local and global level, which we used in this study to make novel insights into

460 a range of theoretical models. Together, the abilities to condition upon emergent properties, the
461 efficient inference algorithm, and the capacity for parameter sensitivity analyses make EPI a useful
462 method for addressing inverse problems in theoretical neuroscience.

463 **Acknowledgements:**

464 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
465 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
466 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
467 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
468 Fellowship. Helpful conversations were had with Francesca Mastrogiovanni, Srdjan Ostojic, James
469 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

470 **Data availability statement:**

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

473 **Code availability statement:**

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

475 **References**

- 476 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
477 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 478 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
479 21(1):25–45, 1998.
- 480 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 481 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in cog-
482 nition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 483 [5] Timothy O’Leary, Alexander C Sutton, and Eve Marder. Computational models in the age of
484 large datasets. *Current opinion in neurobiology*, 32:87–94, 2015.
- 485 [6] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
486 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
487 models. *PLoS Comput Biol*, 3(10):e189, 2007.

- 488 [7] Kamil Erguler and Michael PH Stumpf. Practical limits for reverse engineering of dynamical
489 systems: a statistical analysis of sensitivity and parameter inferability in systems biology
490 models. *Molecular BioSystems*, 7(5):1593–1602, 2011.
- 491 [8] Brian K Mannakee, Aaron P Ragsdale, Mark K Transtrum, and Ryan N Gutenkunst. Sloppiness
492 and the geometry of parameter space. In *Uncertainty in Biology*, pages 271–299. Springer,
493 2016.
- 494 [9] John J Hopfield. Neural networks and physical systems with emergent collective computational
495 abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 496 [10] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
497 networks. *Physical review letters*, 61(3):259, 1988.
- 498 [11] Andrey V Olypher and Ronald L Calabrese. Using constraints on neuronal activity to reveal
499 compensatory changes in neuronal parameters. *Journal of Neurophysiology*, 98(6):3749–3758,
500 2007.
- 501 [12] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
502 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
503 17(11):4382–4388, 1997.
- 504 [13] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in
505 perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 506 [14] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
507 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 508 [15] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
509 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 510 [16] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
511 model. *PLoS computational biology*, 2(7):e94, 2006.
- 512 [17] Dmitry Fisher, Itsaso Olasagasti, David W Tank, Emre RF Aksay, and Mark S Goldman.
513 A modeling framework for deriving the structural and functional architecture of a short-term
514 memory microcircuit. *Neuron*, 79(5):987–1000, 2013.

- 515 [18] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
516 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
517 sion model. *Neuron*, 82(4):809–821, 2014.
- 518 [19] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
519 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 520 [20] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
521 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
522 2018.
- 523 [21] Christopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state
524 in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.
- 525 [22] Aman B Saleem, Asli Ayaz, Kathryn J Jeffery, Kenneth D Harris, and Matteo Carandini.
526 Integration of visual motion and locomotion in mouse visual cortex. *Nature neuroscience*,
527 16(12):1864–1869, 2013.
- 528 [23] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Church-
529 land. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuro-
530 science*, 22(10):1677–1686, 2019.
- 531 [24] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation
532 in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 533 [25] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
534 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
535 2003.
- 536 [26] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
537 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 538 [27] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
539 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
540 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
541 1929, 2009.

- 542 [28] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
543 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
544 2012.
- 545 [29] Keegan E Hines, Thomas R Middendorf, and Richard W Aldrich. Determination of parameter
546 identifiability in nonlinear biophysical models: A bayesian approach. *Journal of General*
547 *Physiology*, 143(3):401–416, 2014.
- 548 [30] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
549 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 550 [31] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
551 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 552 [32] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
553 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 554 [33] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical
555 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.
- 556 [34] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
557 free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–
558 5533, 2017.
- 559 [35] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
560 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P
561 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
562 dynamics. *bioRxiv*, page 838383, 2019.
- 563 [36] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
564 *International Conference on Machine Learning*, 2015.
- 565 [37] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji
566 Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint*
567 *arXiv:1912.02762*, 2019.
- 568 [38] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
569 networks. *International Conference on Learning Representations*, 2017.

- 570 [39] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
571 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 572 [40] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
573 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 574 [41] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
575 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
576 *Neuron*, 77(5):845–858, 2013.
- 577 [42] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness,
578 and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.
- 579 [43] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
580 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 581 [44] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dy-
582 namics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–
583 1406, 2014.
- 584 [45] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding
585 and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.
- 586 [46] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and vi-
587 sual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,
588 115(3):1399–1409, 2016.
- 589 [47] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
590 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
591 operating regime of cortex. *bioRxiv*, 2020.
- 592 [48] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
593 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
594 routing. *bioRxiv*, page 245613, 2019.
- 595 [49] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
596 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 597 [50] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
598 2009.

- 599 [51] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural
600 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 601 [52] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in*
602 *neurobiology*, 25:156–163, 2014.
- 603 [53] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*
604 *opinion in neurobiology*, 46:1–6, 2017.
- 605 [54] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H
606 Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex
607 embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966,
608 2018.
- 609 [55] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian compu-*
610 *tation*. CRC Press, 2018.
- 611 [56] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference.
612 *Proceedings of the National Academy of Sciences*, 2020.
- 613 [57] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
614 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
615 592, 2009.
- 616 [58] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
617 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,
618 85(2):402–417, 2015.
- 619 [59] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
620 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
621 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 622 [60] Mark M. Churchland, Byron M. Yu, John P. Cunningham, Leo P. Sugrue, Marlene R. Cohen,
623 Greg S. Corrado, William T. Newsome, Andrew M. Clark, Paymon Hosseini, Benjamin B.
624 Scott, David C. Bradley, Matthew A. Smith, Adam Kohn, J. Anthony Movshon, Katherine
625 M. Armstrong, Tirin Moore, Steve W. Chang, Lawrence H. Snyder, Stephen G. Lisberger,
626 Nicholas J. Priebe, Ian M. Finn, David Ferster, Stephen I. Ryu, Gopal Santhanam, Maneesh
627 Sahani, and Krishna V. Shenoy. Stimulus onset quenches neural variability: a widespread
628 cortical phenomenon. *Nat. Neurosci.*, 13(3):369–378, 2010.

- 629 [61] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
630 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
631 5(10):793, 2004.
- 632 [62] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
633 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-*
634 *biology*, 71(1):45–61, 2011.
- 635 [63] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex:
636 From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 637 [64] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
638 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
639 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 640 [65] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate
641 cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
- 642 [66] C Gardiner. Stochastic methods: A Handbook for the Natural and Social Sciences, 2009.
- 643 [67] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain
644 regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.
- 645 [68] Eve Marder and Allen I Selverston. *Dynamic biological networks: the stomatogastric nervous*
646 *system*. MIT press, 1992.
- 647 [69] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
648 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
649 73(2):123–214, 2011.
- 650 [70] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
651 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 652 [71] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
653 Edward Teller. Equation of state calculations by fast computing machines. *The journal of*
654 *chemical physics*, 21(6):1087–1092, 1953.
- 655 [72] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
656 1970.

- 657 [73] Ben Calderhead and Mark Girolami. Statistical analysis of nonlinear dynamical systems using
658 differential geometric sampling methods. *Interface focus*, 1(6):821–835, 2011.
- 659 [74] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-
660 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
661 820, 2011.
- 662 [75] Oksana A Chkrebtii, David A Campbell, Ben Calderhead, Mark A Girolami, et al. Bayesian
663 solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–
664 1267, 2016.
- 665 [76] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
666 A framework for parameter estimation and model selection from experimental data in systems
667 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 668 [77] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
669 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience
670 Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 671 [78] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
672 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with
673 degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 674 [79] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
675 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
676 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 677 [80] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical inference for
678 analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin, Germany*,
679 2019.
- 680 [81] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
681 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
682 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 683 [82] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
684 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on
685 Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

- 686 [83] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
687 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
688 4248. PMLR, 2020.
- 689 [84] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
690 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 691 [85] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
692 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 693 [86] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
694 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
695 2018.
- 696 [87] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
697 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 698 [88] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
699 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 700 [89] Stefan Hengl, Clemens Kreutz, Jens Timmer, and Thomas Maiwald. Data-based identifiability
701 analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618, 2007.
- 702 [90] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density
703 estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.
- 704 [91] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
705 Improved variational inference with inverse autoregressive flow. *Advances in neural information
706 processing systems*, 29:4743–4751, 2016.
- 707 [92] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International
708 Conference on Learning Representations*, 2015.
- 709 [93] Emmanuel Klinger, Dennis Rickert, and Jan Hasenauer. pyabc: distributed, likelihood-free
710 inference. *Bioinformatics*, 34(20):3591–3593, 2018.
- 711 [94] David S Greenberg, Marcel Nonnenmacher, and Jakob H Macke. Automatic posterior trans-
712 formation for likelihood-free inference. *International Conference on Machine Learning*, 2019.

- 713 [95] Daniel P Mossing, Julia Veit, Agostina Palmigiano, Kenneth D. Miller, and Hillel Adesnik.
- 714 Antagonistic inhibitory subnetworks control cooperation and competition across cortical space.
- 715 *bioRxiv*, 2021.

716 **5 Methods**

717 **5.1 Emergent property inference (EPI)**

718 Determining the combinations of model parameters that can produce a desired output is a key part
719 of scientific practice. Solving inverse problems is especially important in neuroscience, since we
720 require detailed circuit models to produce computation of varying levels of complexity. While much
721 machine learning research has focused on how to find latent structure in large-scale neural datasets,
722 less has focused on inverting theoretical circuit models conditioned upon the emergent properties of
723 computation. Here, we introduce a novel method for statistical inference, which finds distributions
724 of parameter solutions that are constrained to produce the desired emergent property. This method
725 seamlessly handles neural circuit models with stochastic nonlinear dynamical generative processes,
726 which are predominant in theoretical neuroscience.

727 Consider model parameterization \mathbf{z} , which is a collection of scientifically meaningful variables that
728 govern the complex simulation of data \mathbf{x} . For example (see Section 3.1), \mathbf{z} may be the electrical
729 conductance parameters of an STG subcircuit, and \mathbf{x} the evolving membrane potentials (the state)
730 of the five neurons. In terms of statistical modeling, this circuit model has an intractable likelihood
731 $p(\mathbf{x} | \mathbf{z})$, which is predicated by the stochastic differential equations that define the model. From a
732 theoretical perspective, we are less concerned about the likelihood of an exemplary dataset \mathbf{x} , but
733 rather the emergent property of intermediate hub frequency (which implies a consistent dataset \mathbf{x}).

734 In the STG example, the statistic $f(\mathbf{x}; \mathbf{z})$ measures hub neuron frequency from the evolution of \mathbf{x}
735 governed by parameters \mathbf{z} . With EPI, we learn distributions of \mathbf{z} constrained to produce intermedi-
736 ate hub frequency: to obey the constraints placed on the mean and variance of $f(\mathbf{x}; \mathbf{z})$. In general,
737 an emergent property \mathcal{X} is defined through the choice of $f(\mathbf{x}; \mathbf{z})$ (which may be one or multiple
738 statistics), and its means $\boldsymbol{\mu}$, and variances $\boldsymbol{\sigma}^2$:

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (13)$$

739 Precisely, the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ must have means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ over the EPI
740 distribution of parameters and the data produced by those parameters. Technically, an emergent
741 property may be a combination of first-, second-, or higher-order moments, but this study focuses
742 on the case written in Equation 13.

743 In EPI, deep probability distributions are optimized to learn the inferred distribution. In deep
744 probability distributions, a simple random variable $\mathbf{z}_0 \sim q_0(\mathbf{z}_0)$ (we choose an isotropic gaussian)

745 is mapped deterministically via a sequence of deep neural network layers ($g_1, \dots g_l$) parameterized
 746 by weights and biases θ to the support of the distribution of interest:

$$\mathbf{z} = g_{\theta}(\mathbf{z}_0) = g_l(\dots g_1(\mathbf{z}_0)) \sim q_{\theta}(\mathbf{z}). \quad (14)$$

747 Such deep probability distributions embed the inferred distribution in a deep network. Once op-
 748 timized, this deep network representation has remarkably useful properties: fast sampling and
 749 probability evaluations Importantly, fast probability evaluations confer fast gradient and Hessian
 750 calculations as well.

751 Given this choice of circuit model and emergent property \mathcal{X} , $q_{\theta}(\mathbf{z})$ is optimized via the neural
 752 network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep variational
 753 family \mathcal{Q} producing the emergent property \mathcal{X} :

$$q_{\theta}(\mathbf{z} | \mathcal{X}) = q_{\theta}^*(\mathbf{z}) = \underset{q_{\theta} \in \mathcal{Q}}{\operatorname{argmax}} H(q_{\theta}(\mathbf{z})) \quad (15)$$

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

754 Entropy is chosen as the normative selection principle to match that of variational bayesian methods
 755 (see Section 5.1.3). However, a key difference is that variational bayesian methods do not constrain
 756 the predictions of their inferred parameter distribution. This optimization is executed using the
 757 algorithm of Maximum Entropy Flow Networks (MEFNs) [38].

758 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.
 759 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-
 760 ond, we describe the special class of deep probability distributions used in EPI called normalizing
 761 flows (Section 5.1.2). Then, we establish the known relationship between maximum entropy dis-
 762 tributions and exponential families (Section 5.1.3). Next, we explain the constrained optimization
 763 technique used to solve Equation 15 (Section 5.1.4). Then, we demonstrate the details of this
 764 optimization in a toy example (Section 5.1.5). Finally, we explain how EPI is a form of variational
 765 inference (Section 5.1.6).

766 5.1.1 Related approaches

767 When bayesian inference problems lack conjugacy, scientists use approximate inference methods like
 768 variational inference (VI) [70] and Markov chain Monte Carlo (MCMC) [71, 72]. After optimization,
 769 variational methods return a parameterized posterior distribution, which we can analyze. Also, the
 770 variational approximating distribution class is often chosen such that it permits fast sampling. In

771 contrast MCMC methods only produce samples from the approximated posterior distribution. No
772 parameterized distribution is estimated, and additional samples are always generated with the same
773 sampling complexity. Inference in models defined by systems of differential has been demonstrated
774 with MCMC [69], although this approach requires tractable likelihoods. Advancements have intro-
775 duced sampling [73], likelihood approximation [74], and uncertainty quantification techniques [75]
776 to make MCMC approaches more efficient and expand the class of applicable models.

777 Simulation-based inference [56] is model parameter inference in the absence of a tractable likeli-
778 hood function. The most prevalent approach to simulation-based inference is approximate bayesian
779 computation (ABC) [24], in which satisfactory parameter samples are kept from random prior sam-
780 pling according to a rejection heuristic. The obtained set of parameters do not have a probabilities,
781 and further insight about the model must be gained from examination of the parameter set and
782 their generated activity. Methodological advances to ABC methods have come through the use of
783 Markov chain Monte Carlo (MCMC-ABC) [25] and sequential Monte Carlo (SMC-ABC) [26] sam-
784 pling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still struggles
785 to scale in dimensionality [55] (cf. Fig. 2). Still, this method has enjoyed much success in systems
786 biology [76]. Furthermore, once a parameter set has been obtained by SMC-ABC from a finite set
787 of particles, the SMC-ABC algorithm must be run again from scratch with a new population of
788 initialized particles to obtain additional samples.

789 For scientific model analysis, we seek a parameter distribution represented by an approximating
790 distribution as in variational inference [70]: a variational approximation that once optimized yields
791 fast analytic calculations and samples. For the reasons described above, ABC and MCMC tech-
792 niques are unattractive, since they only produce a set of parameter samples lacking probabilities
793 and have unchanging sampling rate. EPI infers parameters in circuit models using the MEFN [38]
794 algorithm with a deep variational approximation. The deep neural network of EPI (Fig. 1E) de-
795 fines the parametric form (with variational parameters θ) of the deep variational approximation of
796 circuit parameters \mathbf{z} .

797 Since EPI is not conditioning upon exemplary data as in variational bayesian methodology, EPI
798 is not doing established variational inference. In contrast, the EPI distribution is constrained to
799 produce an emergent property. EPI optimization is enabled using stochastic gradient techniques in
800 the spirit of likelihood-free variational inference [34]. The analytic relationship between EPI and
801 variational inference is explained in Section 5.1.6.

802 We note that, during our preparation and early presentation of this work [77, 78], another work

803 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
804 circuits ([35, 79, 80]). We are encouraged by this general problem being recognized by others in the
805 community, and we emphasize that these works offer complementary neuroscientific contributions
806 (different theoretical models of focus) and use different technical methodologies (ours is built on
807 our prior work [38], theirs similarly [81]).

808 The method EPI differs from SNPE in some key ways. SNPE belongs to a “sequential” class
809 of recently developed simulation-based inference methods in which two neural networks are used
810 for posterior inference. This first neural network is a deep probability distribution (normalizing
811 flow) used to estimate the posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE) or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural
812 likelihood (SNL [82])). A recent advance uses an unconstrained neural network to estimate the
813 likelihood ratio (sequential neural ratio estimation (SNRE [83])). In SNL and SNRE, MCMC
814 sampling techniques are used to obtain samples from the approximated posterior. This contrasts
815 with EPI and SNPE, which use deep probability distributions to model parameters, which facilitates
816 immediate measurements of sample probability, gradient, or Hessian for system analysis. The
817 second neural network in this sequential class of methods is the amortizer. This unconstrained
818 deep network maps data \mathbf{x} (or statistics $f(\mathbf{x}; \mathbf{z})$) or model parameters \mathbf{z} to the weights and biases of
819 the first neural network. These methods are optimized on a conditional density (or ratio) estimation
820 objective. The data used to optimize this objective are generated via an adaptive procedure, in
821 which training data pairs $(\mathbf{x}_i, \mathbf{z}_i)$ become sequentially closer to the true data and posterior.

822 The approximating fidelity of the deep probability distribution in sequential approaches is opti-
823 mized to generalize across the training distribution of the conditioning variable. This generalization
824 property of the sequential methods can reduce the accuracy at the singular posterior of interest.
825 Whereas in EPI, the entire expressivity of the deep probability distribution is dedicated to learning
826 a single distribution as well as possible. Amortization is not possible in EPI, since EPI learns
827 an exponential family distribution parameterized by its mean (see Section 5.1.3). Since EPI dis-
828 tributions are defined by the mean μ of their statistics, there is the well-known inverse mapping
829 problem of exponential families [84] that prohibits an amortization based approach. However, we
830 have shown that the same two-network architecture of the sequential simulation-based inference
831 methods can be used for amortized inference in intractable exponential family posteriors using their
832 natural parameterization [85].

833 Finally, one important differentiating factor between EPI and sequential simulation-based infer-
834 ence methods is that EPI leverages gradients $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ during optimization. These gradients can

835 improve convergence time and scalability, as we have shown on an example conditioning low-rank
 836 RNN connectivity on the property of stable amplification (see Section 3.3). With EPI, we prove
 837 out the suggestion that a deep inference technique can improve efficiency by leveraging these model
 838 gradients when they are tractable. Sequential simulation-based inference techniques may be better
 839 suited for scientific problems where $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ is intractable or unavailable, like when there is a non-
 840 differentiable model or it requires lengthy simulations. However, the sequential simulation-based
 841 inference techniques cannot constrain the predictions of the inferred distribution in the manner of
 842 EPI.

843 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in
 844 scientific models. Around a single parameter choice, one can measure the Jacobian. One approach
 845 for this calculation that scales well is EAR [28]. A popular efficient approach for systems of ODEs
 846 has been neural ODE adjoint [86] and its stochastic adaptation [87]. Casting identifiability as a
 847 statistical estimation problem, the profile likelihood works via iterated optimization while holding
 848 parameters fixed [27]. An exciting recent method is capable of recovering the functional form of such
 849 unidentifiabilities away from a point by following degenerate dimensions of the fisher information
 850 matrix [30]. Global structural non-identifiabilities can be found for models with polynomial or
 851 rational dynamics equations using DAISY [88], or through mean optimal transformations [89].
 852 With EPI, we have all the benefits given by a statistical inference method plus the ability to query
 853 the first- or second-order gradient of the probability of the inferred distribution at any chosen
 854 parameter value. The second-order gradient of the log probability (the Hessian), which is directly
 855 afforded by EPI distributions, produces quantified information about parametric sensitivity of the
 856 emergent property in parameter space (see Section 3.2).

857 **5.1.2 Deep probability distributions and normalizing flows**

858 Deep probability distributions are comprised of multiple layers of fully connected neural networks
 859 (Equation 14). When each neural network layer is restricted to be a bijective function, the sample
 860 density can be calculated using the change of variables formula at each layer of the network. For
 861 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

$$p(\mathbf{z}_i) = p(g_i^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial g_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} \right| = p(\mathbf{z}_{i-1}) \left| \det \frac{\partial g_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right|^{-1}. \quad (16)$$

862 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 863 restricting our layers to normalizing flows [36, 37] – bijective functions with fast log determinant

864 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log
865 probability calculation confers efficient optimization of the maximum entropy objective (see Section
866 5.1.4).

867 We use the Real NVP [39] normalizing flow class, because its coupling architecture confers both
868 fast sampling (forward) and fast log probability evaluation (backward). Fast probability evaluation
869 facilitates fast gradient and Hessian evaluation of log probability throughout parameter space.
870 Glow permutations were used in between coupling stages [40]. This is in contrast to autoregressive
871 architectures [90, 91], in which only one of the forward or backward passes can be efficient. In this
872 work, normalizing flows are used as flexible parameter distribution approximations $q_{\theta}(\mathbf{z})$ having
873 weights and biases θ . We specify the architecture used in each application by the number of Real-
874 NVP affine coupling stages, and the number of neural network layers and units per layer of the
875 conditioning functions.

876 When calculating Hessians of log probabilities in deep probability distributions, it is important to
877 consider the normalizing flow architecture. With autoregressive architectures [90, 91], fast sam-
878 pling and fast log probability evaluations are mutually exclusive. That makes these architectures
879 undesirable for EPI, where efficient sampling is important for optimization, and log probability
880 evaluation speed predicates the efficiency of gradient and Hessian calculations. With Real NVP
881 coupling architectures, we get both fast sampling and fast Hessians making both optimization and
882 scientific analysis efficient.

883 5.1.3 Maximum entropy distributions and exponential families

884 EPI is a maximum entropy distribution, which have fundamental links to exponential family dis-
885 tributions. A maximum entropy distribution of form:

$$p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \quad (17)$$

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

886 will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \quad (18)$$

887 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
888 formally hard to identify except in special cases [84].

889 In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equations 2 and 3. The variance constraint is a second moment constraint on $f(\mathbf{x}; \mathbf{z})$

$$\text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \mathbb{E}_{\mathbf{z}, \mathbf{x}} [(f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2] \quad (19)$$

891 As a general maximum entropy distribution (Equation 17), the sufficient statistics vector contains
892 both first and second order moments of $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (20)$$

893 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (21)$$

894 5.1.4 Augmented lagrangian optimization

895 To optimize $q_{\boldsymbol{\theta}}(\mathbf{z})$ in Equation 15, the constrained maximum entropy optimization is executed using
896 the augmented lagrangian method. The following objective is minimized:

$$L(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{opt}}, c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top R(\boldsymbol{\theta}) + \frac{c}{2} \|R(\boldsymbol{\theta})\|^2 \quad (22)$$

897 where average constraint violations $R(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [T(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_{\text{opt}}]]$, $\boldsymbol{\eta}_{\text{opt}} \in \mathbb{R}^m$ are the
898 Lagrange multipliers where $m = |\boldsymbol{\mu}_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})| = 2|f(\mathbf{x}; \mathbf{z})|$, and c is the penalty coefficient. The
899 sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter $\boldsymbol{\mu}_{\text{opt}}$ are determined by the means $\boldsymbol{\mu}$ and variances
900 σ^2 of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 15 (see Section 5.1.6). Specifically,
901 $T(\mathbf{x}; \mathbf{z})$ is a concatenation of the first and second moments, $\boldsymbol{\mu}_{\text{opt}}$ is a concatenation of $\boldsymbol{\mu}$ and σ^2
902 (see section 5.1.3), and the Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of
903 exponential families (see Section 5.1.3). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution
904 are optimized according to Equation 22 using the Adam optimizer with learning rate 10^{-3} [92].

905 The gradient with respect to entropy $H(q_{\boldsymbol{\theta}}(\mathbf{z}))$ can be expressed using the reparameterization trick
906 as an expectation of the negative log density of parameter samples \mathbf{z} over the randomness in the
907 parameterless initial distribution $q_0(\mathbf{z}_0)$:

$$H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \int -q_{\boldsymbol{\theta}}(\mathbf{z}) \log(q_{\boldsymbol{\theta}}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [-\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (23)$$

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

$$\nabla_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\nabla_{\boldsymbol{\theta}} \log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (24)$$

910 The lagrangian parameters η_{opt} are initialized to zero and adapted following each augmented
911 lagrangian epoch, which is a period of optimization with fixed (η_{opt}, c) for a given number of
912 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
913 after each epoch based on constraint error reduction. The penalty coefficient is updated based
914 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
915 $\mathbb{E}[|R(\theta_{k+1})|] > \gamma \mathbb{E}[|R(\theta_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
916 other update rule is $\eta_{\text{opt},k+1} = \eta_{\text{opt},k} + c_k \frac{1}{n} \sum_{i=1}^n (T(\mathbf{x}^{(i)}) - \mu_{\text{opt}})$ given a batch size n . Throughout
917 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
918 according to application.

919 The intention is that c and η_{opt} start at values encouraging entropic growth early in optimization.
920 With each training epoch in which the update rule for c is invoked by unsatisfactory constraint
921 error reduction, the constraint satisfaction terms are increasingly weighted, resulting in a decreased
922 entropy. This encourages the discovery of suitable regions of parameter space, and the subsequent
923 refinement of the distribution to produce the emergent property (see example in Section 5.1.5). The
924 momentum parameters of the Adam optimizer are reset at the end of each augmented lagrangian
925 epoch.

926 Rather than starting optimization from some θ drawn from a randomized distribution, we found
927 that initializing $q_{\theta}(\mathbf{z})$ to approximate an isotropic Gaussian distribution conferred more stable, con-
928 sistent optimization. The parameters of the Gaussian initialization were chosen on an application-
929 specific basis. Throughout the study, we chose isotropic Gaussian initializations with mean μ_{init}
930 at the center of the distribution support and some standard deviation σ_{init} , except for one case,
931 where an initialization informed by random search was used (see Section 5.2).

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

940 When assessing the suitability of EPI for a particular modeling question, there are some important
941 technical considerations. First and foremost, as in any optimization problem, the defined emergent

942 property should always be appropriately conditioned (constraints should not have wildly different
 943 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution
 944 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-
 945 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,
 946 one should consider the computational cost of the gradient calculations. In the best circumstance,
 947 there is a simple, closed form expression (e.g. Section 5.3) for the emergent property statistic given
 948 the model parameters. On the other end of the spectrum, many forward simulation iterations
 949 may be required before a high quality measurement of the emergent property statistic is available
 950 (e.g. Section 5.2). In such cases, backpropagating gradients through the SDE evolution will be
 951 expensive.

952 5.1.5 Example: 2D LDS

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

$$955 \quad \tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (25)$$

955 with

$$955 \quad A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (26)$$

956 To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fixing
 957 $\tau = 1s$), the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ were chosen to contain the oscillatory frequency,
 958 $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscillating system. λ_1 is the eigenvalue of
 959 greatest real part when the imaginary component is zero, and alternatively of positive imaginary
 960 component when the eigenvalues are complex conjugate pairs. To learn the distribution of real
 961 entries of A that produce a band of oscillating systems around 1Hz, we formalized this emergent
 962 property as $\text{real}(\lambda_1)$ having mean zero with variance 0.25^2 , and the oscillation frequency $2\pi\text{imag}(\lambda_1)$
 963 having mean 1Hz with variance $(0.1\text{Hz})^2$:

$$964 \quad \mathbb{E}[T(\mathbf{x})]_{\mathbf{z}, \mathbf{x}} \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \text{imag}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0)^2 \\ (\text{imag}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}_{\text{opt}}. \quad (27)$$

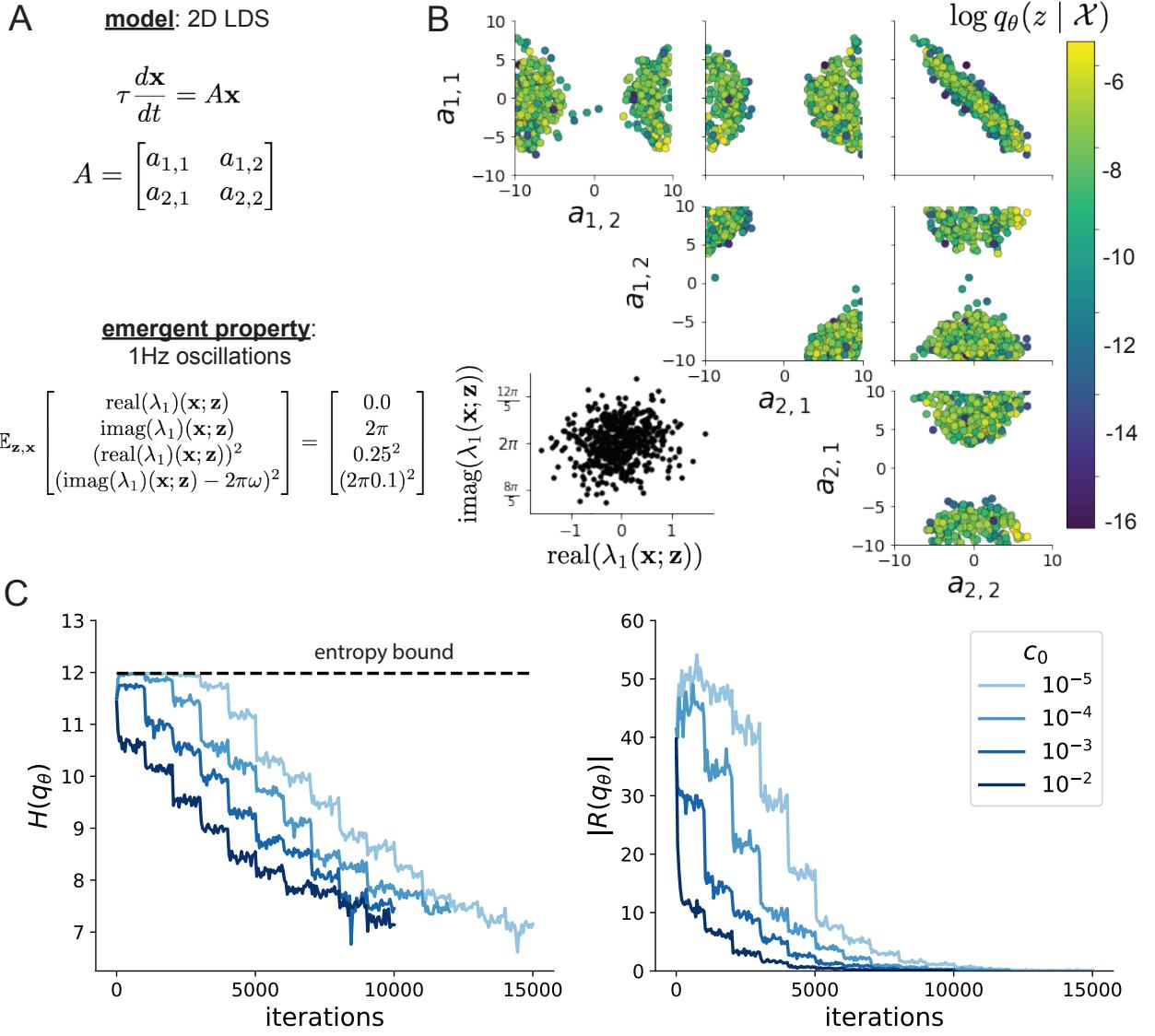


Figure S1: **A.** Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. **B.** The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed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

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

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (28)$$

968 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
 969 does not require simulation, we cannot derive the distribution q_{θ}^* directly. This fact is due to the
 970 formally hard problem of the backward mapping: finding the natural parameters η from the mean
 971 parameters μ of an exponential family distribution [84]. Instead, we used EPI to approximate this
 972 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two
 973 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped
 974 onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

975 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the
 976 parameter distribution. To validate our method, we analytically derived the contours of the prob-
 977 ability density from the emergent property statistics and values. In the a_1 - a_4 plane, the black
 978 line at $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$,
 979 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
 980 of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired
 981 statistical constraints and model degeneracy in the sum of a_1 and a_4 . Intuitively, the parameters
 982 equivalent with respect to emergent property statistic $\text{real}(\lambda_1)$ have similar log densities.

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4-a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (29)$$

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

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \quad (30)$$

987 Since we constrained $\mathbb{E}_{\mathbf{z} \sim q_{\theta}}[\text{imag}(\lambda)] = 2\pi$, we can plot contours of the equation $\text{imag}(\lambda_1)^2 =$
 988 $a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_{\theta}}(|a_1a_4 - E_{q_{\theta}}[a_1a_4]|)$, we show
 989 the contours as $a_1a_4 = 0$ (black), $a_1a_4 = -\sigma_{1,4}$ (black dotted), and $a_1a_4 = -2\sigma_{1,4}$ (grey dotted).
 990 This validates the curved structure of the inferred distribution learned through EPI. We took steps

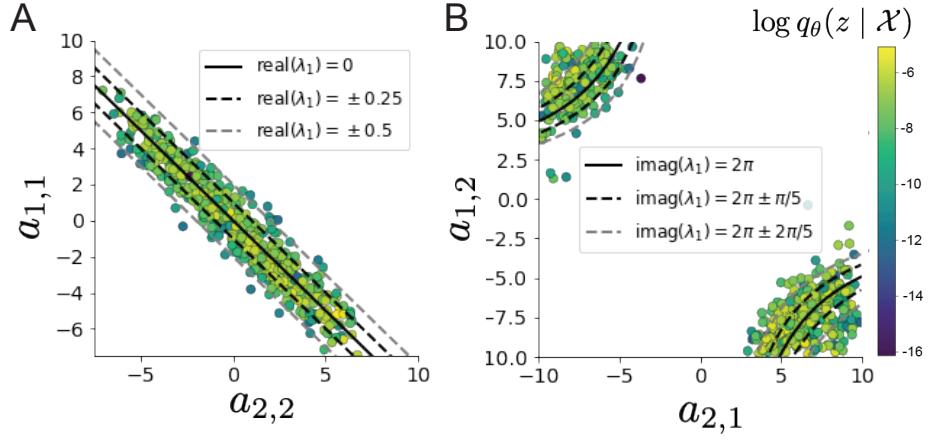


Figure S2: **A.** Probability contours in the a_1-a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. **B.** Probability contours in the a_2-a_3 plane were derived from the emergent property statistic of oscillation frequency $2\pi\text{imag}(\lambda_1)$.

991 in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive values
 992 a_1a_4 in the learned distribution. Subtler combinations of model and emergent property will have
 993 more complexity, further motivating the use of EPI for understanding these systems. As we expect,
 994 the distribution results in samples of two-dimensional linear systems oscillating near 1Hz (Fig. S3).

995 5.1.6 EPI as variational inference

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

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

1000 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
 1001 However, conjugacy is rare in practice, and alternative methods, such as variational inference [70],
 1002 are utilized.

1003 In variational inference, a posterior approximation q_{θ}^* is chosen from within some variational family
 1004 \mathcal{Q}

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

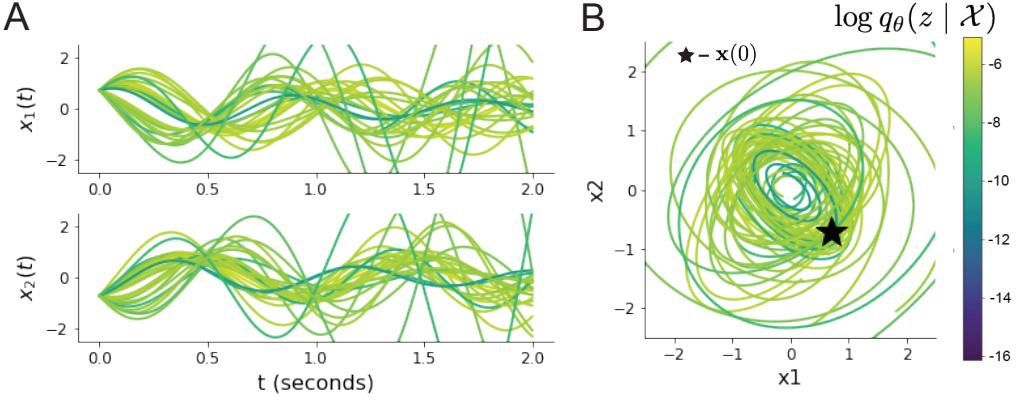


Figure S3: Sampled dynamical systems $\mathbf{z} \sim q_\theta(\mathbf{z})$ and their simulated activity from $\mathbf{x}(t = 0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. **A.** Each dimension of the simulated trajectories throughout time. **B.** The simulated trajectories in phase space.

1005 The KL divergence can be written in terms of entropy of the variational approximation:

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

1006

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

1007 Since the marginal distribution of the data $p(\mathbf{x})$ (or ‘‘evidence’’) is independent of θ , variational
1008 inference is executed by optimizing the remaining expression. This is usually framed as maximizing
1009 the evidence lower bound (ELBO)

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

1010 Now, consider the setting where we have chosen a uniform prior, and stipulate a mean-field gaussian
1011 likelihood on a chosen statistic of the data $f(\mathbf{x}; \mathbf{z})$

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

1012 where $\Sigma_f = \text{diag}(\sigma_f^2)$. The log likelihood is then proportional to a dot product of the natural
1013 parameter of this mean-field gaussian distribution and the first and second moment statistics.

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

1014 where

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

1015

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

1016 The variational objective is then

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

1017 Comparing this to the lagrangian objective (without augmentation) of EPI, we see they are the

1018 same

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

1019 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
 1020 (Equation 20). Thus, EPI is implicitly executing variational inference with a uniform prior and a
 1021 mean-field gaussian likelihood on the emergent property statistics. The mean and variances of the
 1022 mean-field gaussian likelihood are predicated by $\boldsymbol{\eta}_{\text{opt}}$ (Equations 38 and 40), which is adapted after
 1023 each EPI optimization epoch based on \mathcal{X} (see Section 5.1.4). In EPI, the inferred distribution is
 1024 not conditioned on a finite dataset as in variational inference, but rather the emergent property
 1025 \mathcal{X} dictates the likelihood parameterization such that the inferred distribution will produce the
 1026 emergent property. As a note, we could not simply choose $\boldsymbol{\mu}_f$ and $\boldsymbol{\sigma}_f$ directly from the outset, since
 1027 we do not know which of these choices will produce the emergent property \mathcal{X} , which necessitates
 1028 the EPI optimization routine that adapts $\boldsymbol{\eta}_{\text{opt}}$. Accordingly, we replace the notation of $p(\mathbf{z} | \mathbf{x})$
 1029 with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys emergent property \mathcal{X} (see Section
 1030 5.1).

1031

5.2 Stomatogastric ganglion

1032 In Section 3.1 and 3.2, we used EPI to infer conductance parameters in a model of the stomatogastric
 1033 ganglion (STG) [41]. This 5-neuron circuit model represents two subcircuits: that generating the
 1034 pyloric rhythm (fast population) and that generating the gastric mill rhythm (slow population).
 1035 The additional neuron (the IC neuron of the STG) receives inhibitory synaptic input from both
 1036 subcircuits, and can couple to either rhythm dependent on modulatory conditions. There is also
 1037 a parametric regime in which this neuron fires at an intermediate frequency between that of the
 1038 fast and slow populations [41], which we infer with EPI as a motivational example. This model

1039 is not to be confused with an STG subcircuit model of the pyloric rhythm [68], which has been
 1040 statistically inferred in other studies [15, 35].

1041 **5.2.1 STG model**

1042 We analyze how the parameters $\mathbf{z} = [g_{el}, g_{synA}]$ govern the emergent phenomena of intermediate
 1043 hub frequency in a model of the stomatogastric ganglion (STG) [41] shown in Figure 1A with
 1044 activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$, using the same hyperparameter choices as Gutierrez et al.
 1045 Each neuron's membrane potential $x_\alpha(t)$ for $\alpha \in \{f1, f2, \text{hub}, s1, s2\}$ is the solution of the following
 1046 stochastic differential equation:

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

1047 The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization,
 1048 electrical and synaptic currents. Each current component is a function of all membrane potentials
 1049 and the conductance parameters \mathbf{z} . Finally, we include gaussian noise dB to the model of Gutierrez
 1050 et al. so that the model stochastic, although this is not required by EPI.

1051 The capacitance of the cell membrane was set to $C_m = 1nF$. Specifically, the currents are the
 1052 difference in the neuron's membrane potential and that current type's reversal potential multiplied
 1053 by a conductance:

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

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

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

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

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

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

1059 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 1060 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 1061 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 1062 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
 1063 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
 1064 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

1065 Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating
 1066 dynamics dependent on steady-state gating variables M_∞ , N_∞ and H_∞ , respectively:

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

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

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

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

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

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

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

1073 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 1074 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

1075 Finally, there is a synaptic gating variable as well:

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

1076 When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear
 1077 dynamical system. The gaussian noise $d\mathbf{B}$ has variance $(1 \times 10^{-12})^2$ A², and introduces variability
 1078 in frequency at each parameterization \mathbf{z} .

1079 5.2.2 Hub frequency calculation

1080 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for
 1081 $T = 300$ time steps of $dt = 25\text{ms}$. The chosen dt and T were the most computationally convenient
 1082 choices yielding accurate frequency measurement. We used a basis of complex exponentials with
 1083 frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

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

1084 To measure spiking frequency, we processed simulated membrane potentials with a relu (spike
 1085 extraction) and low-pass filter with averaging window of size 20, then took the frequency with the

1086 maximum absolute value of the complex exponential basis coefficients of the processed time-series.
 1087 The first 20 temporal samples of the simulation are ignored to account for initial transients.
 1088 To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in$
 1089 $\mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in$
 1090 $\{f1, f2, \text{hub}, s1, s2\}$. The soft-argmax is then calculated using temperature parameter $\beta_\psi = 100$

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

1091 where $i = [0, 1, \dots, 100]$. The frequency is then calculated as

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

1092 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
 1093 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
 1094 frequency, where the mean was chosen to be 0.55Hz,(Equation 2) and variance was chosen to be
 1095 0.025^2 Hz 2 (Equation 3).

1096 5.2.3 EPI details for the STG model

1097 As a maximum entropy distribution, $T(\mathbf{x}; \mathbf{z})$ is comprised of both these first and second moments
 1098 of the hub neuron frequency (as in Equations 20 and 21)

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

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

1099
 1100 Throughout optimization, the augmented lagrangian parameters η and c , were updated after each
 1101 epoch of 5,000 iterations(see Section 5.1.4). The optimization converged after five epochs (Fig. S4).
 1102 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-
 1103 layer neural networks of 25 units per layer. The normalizing flow architecture mapped $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$
 1104 to a support of $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}] \in [4, 8] \times [0.01, 4]$, initialized to a gaussian approximation of samples
 1105 returned by a preliminary ABC search. We did not include $g_{\text{synA}} < 0.01$, for numerical stability.
 1106 EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an
 1107 augmented lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The architecture
 1108 converged with criteria $N_{\text{test}} = 100$.

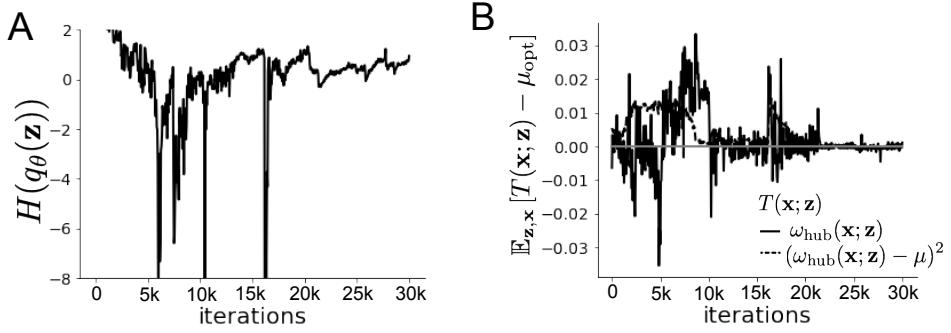


Figure S4: EPI optimization of the STG model producing network syncing. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented lagrangian epoch.

1109 **5.2.4 Hessian sensitivity vectors**

1110 To quantify the second-order structure of the EPI distribution, we evaluated the Hessian of the log
1111 probability $\frac{\partial^2 \log q(\mathbf{z}|\mathcal{X})}{\partial \mathbf{z} \partial \mathbf{z}^\top}$. The eigenvector of this Hessian with most negative eigenvalue is defined as
1112 the sensitivity dimension \mathbf{v}_1 , and all subsequent eigenvectors are ordered by increasing eigenvalue.
1113 These eigenvalues are quantifications of how fast the emergent property deteriorates via the param-
1114 eter combination of their associated eigenvector. In Figure 1D, the sensitivity dimension v_1 (solid)
1115 and the second eigenvector of the Hessian v_2 (dashed) are shown evaluated at the mode of the
1116 distribution. Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D
1117 parameter space were chosen to have positive g_{synA} . The length of the arrows is inversely propor-
1118 tional to the square root of the absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$. For
1119 the same magnitude perturbation away from the mode, intermediate hub frequency only diminishes
1120 along the sensitivity dimension \mathbf{v}_1 (Fig. 1E-F).

1121 **5.3 Scaling EPI for stable amplification in RNNs**

1122 **5.3.1 Rank-2 RNN model**

1123 We examined the scaling properties of EPI by learning connectivities of RNNs of increasing size
1124 that exhibit stable amplification. Rank-2 RNN connectivity was modeled as $W = UV^\top$, where
1125 $U = [\mathbf{U}_1 \quad \mathbf{U}_2] + g\chi^{(W)}$, $V = [\mathbf{V}_1 \quad \mathbf{V}_2] + g\chi^{(V)}$, and $\chi_{i,j}^{(W)}, \chi_{i,j}^{(V)} \sim \mathcal{N}(0, 1)$. This RNN model has
1126 dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}. \quad (62)$$

1127 In this analysis, we inferred connectivity parameterizations $\mathbf{z} = [\mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{V}_1^\top, \mathbf{V}_2^\top]^\top \in [-1, 1]^{(4N)}$
1128 that produced stable amplification using EPI, SMC-ABC [26], and SNPE [35] (see Section Related
1129 Methods).

1130 **5.3.2 Stable amplification**

1131 For this RNN model to be stable, all real eigenvalues of W must be less than 1: $\text{real}(\lambda_1) < 1$,
1132 where λ_1 denotes the greatest real eigenvalue of W . For a stable RNN to amplify at least one input
1133 pattern, the symmetric connectivity $W^s = \frac{W + W^\top}{2}$ must have an eigenvalue greater than 1: $\lambda_1^s > 1$,
1134 where λ^s is the maximum eigenvalue of W^s . These two conditions are necessary and sufficient for
1135 stable amplification in RNNs [51].

1136 **5.3.3 EPI details for RNNs**

1137 We defined the emergent property of stable amplification with means of these eigenvalues (0.5
1138 and 1.5, respectively) that satisfy these conditions. To complete the emergent property definition,
1139 we chose variances (0.25^2) about those means such that samples rarely violate the eigenvalue
1140 constraints. In terms of the EPI optimization variables, this is written as

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \lambda_1^s(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0.5)^2 \\ (\lambda_1^s(\mathbf{x}; \mathbf{z}) - 1.5)^2 \end{bmatrix}, \quad (63)$$

1141

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.25^2 \\ 0.25^2 \end{bmatrix}. \quad (64)$$

1142 Gradients of maximum eigenvalues of Hermitian matrices like W^s are available with modern auto-
1143 matic differentiation tools. To differentiate through the $\text{real}(\lambda_1)$, we solved the following equation
1144 for eigenvalues of rank-2 matrices using the rank reduced matrix $W^r = V^\top U$

$$\lambda_{\pm} = \frac{\text{Tr}(W^r) \pm \sqrt{\text{Tr}(W^r)^2 - 4\text{Det}(W^r)}}{2}. \quad (65)$$

1145 For EPI in Fig. 2, we used a real NVP architecture with three coupling layers of affine transfor-
1146 mations parameterized by two-layer neural networks of 100 units per layer. The initial distribution

1147 was a standard isotropic gaussian $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to the support of $\mathbf{z}_i \in [-1, 1]$. We used
 1148 an augmented lagrangian coefficient of $c_0 = 10^3$, a batch size $n = 200$, $\beta = 4$, and chose to use
 1149 500 iterations per augmented lagrangian epoch and emergent property constraint convergence was
 1150 evaluated at $N_{\text{test}} = 200$ (Fig. 2B blue line, and Fig. 2C-D blue).

1151 5.3.4 Methodological comparison

1152 We compared EPI to two alternative simulation-based inference techniques, since the likelihood
 1153 of these eigenvalues given \mathbf{z} is not available. Approximate bayesian computation (ABC) [24] is a
 1154 rejection sampling technique for obtaining sets of parameters \mathbf{z} that produce activity \mathbf{x} close to some
 1155 observed data \mathbf{x}_0 . Sequential Monte Carlo approximate bayesian computation (SMC-ABC) is the
 1156 state-of-the-art ABC method, which leverages SMC techniques to improve sampling speed. We ran
 1157 SMC-ABC with the pyABC package [93] to infer RNNs with stable amplification: connectivities
 1158 having eigenvalues within an ϵ -defined l_2 distance of

$$\mathbf{x}_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}. \quad (66)$$

1159 SMC-ABC was run with a uniform prior over $\mathbf{z} \in [-1, 1]^{(4N)}$, a population size of 1,000 particles
 1160 with simulations parallelized over 32 cores, and a multivariate normal transition model.

1161 SNPE, the next approach in our comparison, is far more similar to EPI. Like EPI, SNPE treats pa-
 1162 rameters in mechanistic models with deep probability distributions, yet the two learning algorithms
 1163 are categorically different. SNPE uses a two-network architecture to approximate the posterior dis-
 1164 tribution of the model conditioned on observed data \mathbf{x}_0 . The amortizing network maps observations
 1165 \mathbf{x}_i to the parameters of the deep probability distribution. The weights and biases of the parameter
 1166 network are optimized by sequentially augmenting the training data with additional pairs $(\mathbf{z}_i, \mathbf{x}_i)$
 1167 based on the most recent posterior approximation. This sequential procedure is important to get
 1168 training data \mathbf{z}_i to be closer to the true posterior, and \mathbf{x}_i to be closer to the observed data. For
 1169 the deep probability distribution architecture, we chose a masked autoregressive flow with affine
 1170 couplings (the default choice), three transforms, 50 hidden units, and a normalizing flow mapping
 1171 to the support as in EPI. This architectural choice closely tracked the size of the architecture used
 1172 by EPI (Fig. S5). As in SMC-ABC, we ran SNPE with $\mathbf{x}_0 = \mu$. All SNPE optimizations were
 1173 run for a limit of 1.5 days on a Tesla V100 GPU, or until two consecutive rounds resulted in a
 1174 validation log probability lower than the maximum observed for that random seed.

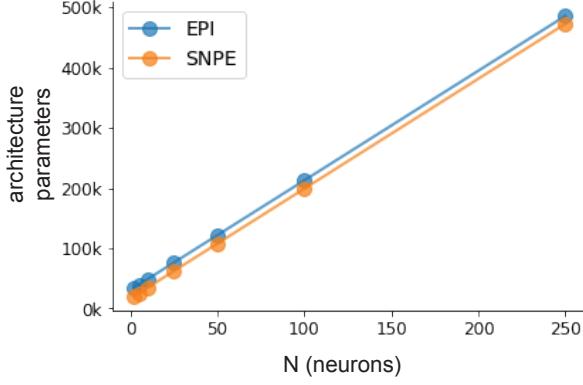


Figure S5: Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size (N).

1175 To compare the efficiency of these algorithms for inferring RNN connectivity distributions producing
 1176 stable amplification, we develop a convergence criteria that can be used across methods. While EPI
 1177 has its own hypothesis testing convergence criteria for the emergent property, it would not make
 1178 sense to use this criteria on SNPE and SMC-ABC which do not constrain the means and variances
 1179 of their predictions. Instead, we consider EPI and SNPE to have converged after completing its
 1180 most recent optimization epoch (EPI) or round (SNPE) in which the distance

$$d(q_{\theta}(\mathbf{z})) = \|\mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}\|_2 \quad (67)$$

1181 is less than 0.5. We consider SMC-ABC to have converged once the population produces samples
 1182 within the $\epsilon = 0.5$ ball ensuring stable amplification.

1183 When assessing the scalability of SNPE, it is important to check that alternative hyperparameter-
 1184izations could not yield better performance. Key hyperparameters of the SNPE optimization are
 1185 the number of simulations per round n_{round} , the number of atoms used in the atomic proposals of
 1186 the SNPE-C algorithm [94], and the batch size n . To match EPI, we used a batch size of $n = 200$
 1187 for $N \leq 25$, however we found $n = 1,000$ to be helpful for SNPE in higher dimensions. While
 1188 $n_{\text{round}} = 1,000$ yielded SNPE convergence for $N \leq 25$, we found that a substantial increase to
 1189 $n_{\text{round}} = 25,000$ yielded more consistent convergence at $N = 50$ (Fig. S6A). By increasing n_{round} ,
 1190 we also necessarily increase the duration of each round. At $N = 100$, we tried two hyperparameter
 1191 modifications. As suggested in [94], we increased n_{atom} by an order of magnitude to improve gra-
 1192 dient quality, but this had little effect on the optimization (much overlap between same random
 1193 seeds) (Fig. S6B). Finally, we increased n_{round} by an order of magnitude, which yielded conver-
 1194 gence in one case, but no others. We found no way to improve the convergence rate of SNPE

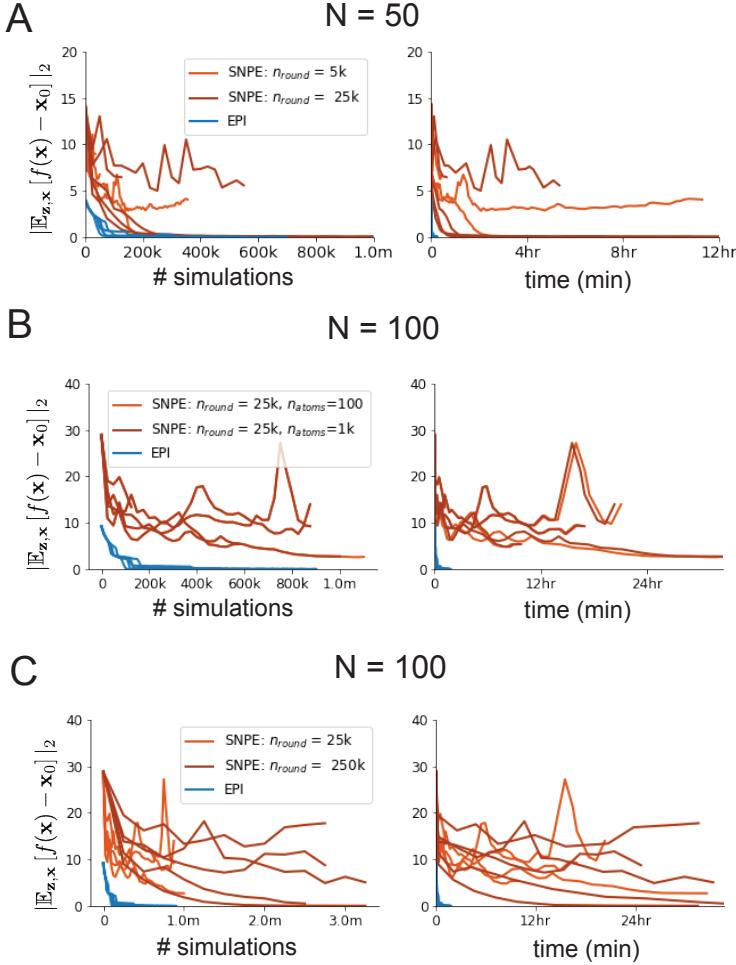


Figure S6: SNPE convergence was enabled by increasing n_{round} , not n_{atom} . **A.** Difference of mean predictions \mathbf{x}_0 throughout optimization at $N = 50$ with by simulation count (left) and wall time (right) of SNPE with $n_{\text{round}} = 5,000$ (light orange), SNPE with $n_{\text{round}} = 25,000$ (dark orange), and EPI (blue). Each line shows an individual random seed. **B.** Same conventions as A at $N = 100$ of SNPE with $n_{\text{atom}} = 100$ (light orange) and $n_{\text{atom}} = 1,000$ (dark orange). **C.** Same conventions as A at $N = 100$ of SNPE with $n_{\text{round}} = 25,000$ (light orange) and $n_{\text{round}} = 250,000$ (dark orange).

1195 without making more aggressive hyperparameter choices requiring high numbers of simulations. In
1196 Figure 2C-D, we show samples from the random seed resulting in emergent property convergence
1197 at greatest entropy (EPI), the random seed resulting in greatest validation log probability (SNPE),
1198 and the result of all converged random seeds (SMC).

1199 **5.3.5 Effect of RNN parameters on EPI and SNPE inferred distributions**

1200 To clarify the difference in objectives of EPI and SNPE, we show their results on RNN models
1201 with different numbers of neurons N and random strength g . The parameters inferred by EPI
1202 consistently produces the same mean and variance of $\text{real}(\lambda_1)$ and λ_1^s , while those inferred by
1203 SNPE change according to the model definition (Fig. S7A). For $N = 2$ and $g = 0.01$, the SNPE
1204 posterior has greater concentration in eigenvalues around \mathbf{x}_0 than at $g = 0.1$, where the model has
1205 greater randomness (Fig. S7B top, orange). At both levels of g when $N = 2$, the posterior of SNPE
1206 has lower entropy than EPI at convergence (Fig. S7B top). However at $N = 10$, SNPE results in
1207 a predictive distribution of more widely dispersed eigenvalues (Fig. S7A bottom), and an inferred
1208 posterior with greater entropy than EPI (Fig. S7B bottom). We highlight these differences not
1209 to focus on an insightful trend, but to emphasize that these methods optimize different objectives
1210 with different implications.

1211 Note that SNPE converges when its validation log probability has saturated after several rounds
1212 of optimization (Fig. S7C), and that EPI converges after several epochs of its own optimization
1213 to enforce the emergent property constraints (Fig. S7D blue). Importantly, as SNPE optimizes
1214 its posterior approximation, the predictive means change, and at convergence may be different
1215 than \mathbf{x}_0 (Fig. S7D orange, left). It is sensible to assume that predictions of a well-approximated
1216 SNPE posterior should closely reflect the data on average (especially given a uniform prior and
1217 a low degree of stochasticity), however this is not a given. Furthermore, no aspect of the SNPE
1218 optimization controls the variance of the predictions (Fig. S7D orange, right).

1219 **5.4 Primary visual cortex**

1220 **5.4.1 V1 model**

1221 E-I circuit models, rely on the assumption that inhibition can be studied as an indivisible unit,
1222 despite ample experimental evidence showing that inhibition is instead composed of distinct ele-
1223 ments [63]. In particular three types of genetically identified inhibitory cell-types – parvalbumin

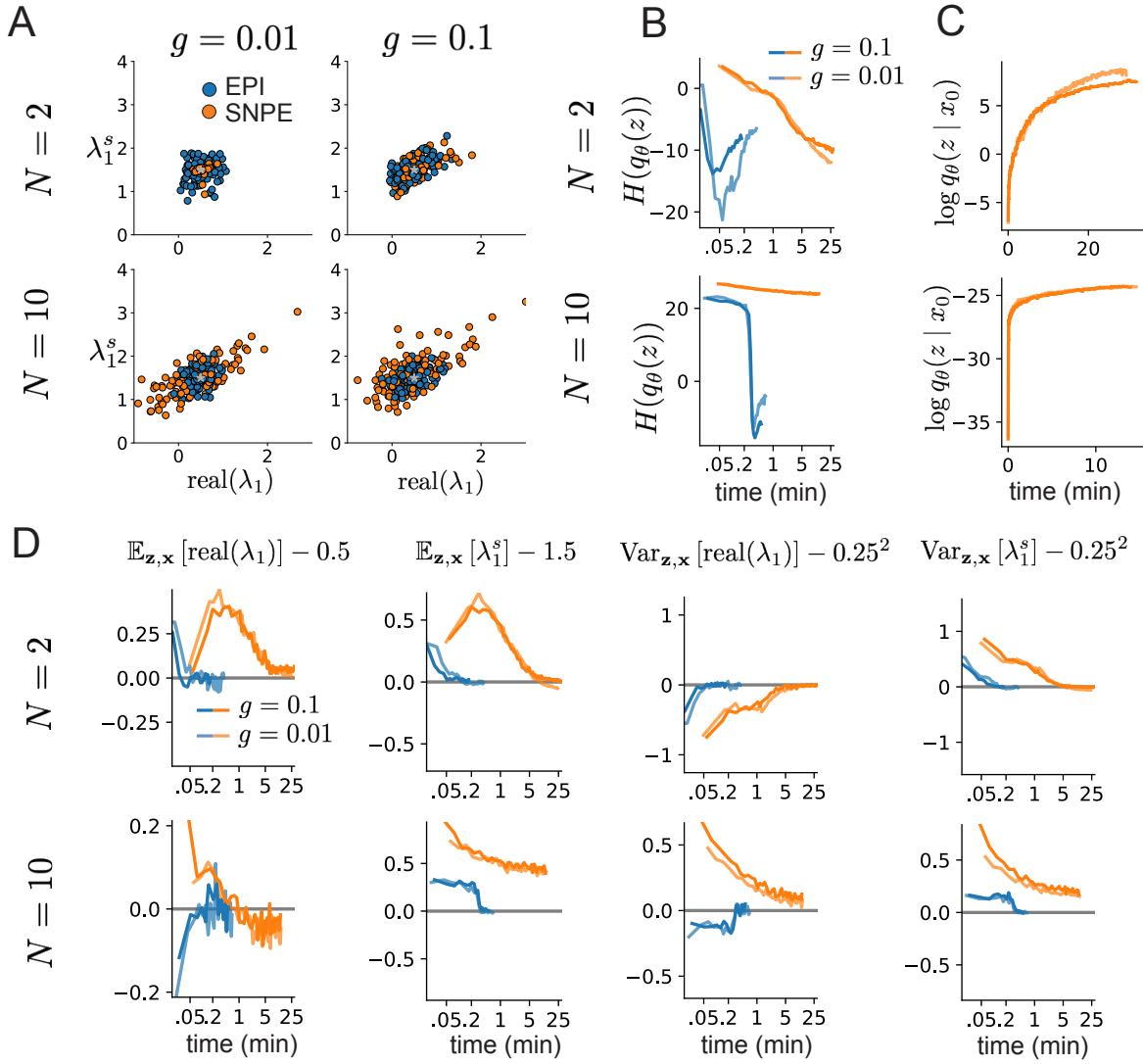


Figure S7: Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A.** Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with $N = 2$ (top), $N = 10$ (bottom), $g = 0.01$ (left), and $g = 0.1$ (right). **B.** Entropy of parameter distribution approximations throughout optimization with $N = 2$ (top), $N = 10$ (bottom), $g = 0.1$ (dark shade), and $g = 0.01$ (light shade). **C.** Validation log probabilities throughout SNPE optimization. Same conventions as B. **D.** Adherence to EPI constraints. Same conventions as B.

1224 (P), somatostatin (S), VIP (V) – compose 80% of GABAergic interneurons in V1 [61–63], and follow
 1225 specific connectivity patterns (Fig. 3A) [64], which lead to cell-type specific computations [47, 95].
 1226 Currently, how the subdivision of inhibitory cell-types, shapes correlated variability by reconfigur-
 1227 ing recurrent network dynamics is not understood.

1228 In the stochastic stabilized supralinear network [59], population rate responses \mathbf{x} to mean input \mathbf{h} ,
 1229 recurrent input $W\mathbf{x}$ and slow noise $\boldsymbol{\epsilon}$ are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \quad (68)$$

1230 where the noise is an Ornstein-Uhlenbeck process $\boldsymbol{\epsilon} \sim OU(\tau_{\text{noise}}, \boldsymbol{\sigma})$

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (69)$$

1231 with $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$. The noisy process is parameterized as

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (70)$$

1232 so that $\boldsymbol{\sigma}$ parameterizes the variance of the noisy input in the absence of recurrent connectivity
 1233 ($W = \mathbf{0}$). As contrast $c \in [0, 1]$ increases, input to the E- and P-populations increases relative to
 1234 a baseline input $\mathbf{h} = \mathbf{h}_b + c\mathbf{h}_c$. Connectivity (W_{fit}) and input ($\mathbf{h}_{b,\text{fit}}$ and $\mathbf{h}_{c,\text{fit}}$) parameters were fit
 1235 using the deterministic V1 circuit model [47]

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (71)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (72)$$

1236 and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (73)$$

1237 To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equiv-
 1238 alence class

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (74)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (75)$$

1239 and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (76)$$

1240 Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5\text{ms}$ from an initial condition
 1241 drawn from $\mathbf{x}(0) \sim U[10\text{Hz}, 25\text{Hz}]$. Standard deviation of the E-population $s_E(\mathbf{x}; \mathbf{z})$ is calculated
 1242 as the square root of the temporal variance from $t_{ss} = 75\text{ms}$ to $Tdt = 100\text{ms}$ averaged over 100
 1243 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[\sqrt{\mathbb{E}_{t>t_{ss}} [(x_E(t) - \mathbb{E}_{t>t_{ss}} [x_E(t)])^2]} \right] \quad (77)$$

1244 5.4.2 EPI details for the V1 model

1245 For EPI in Figures 3D-E and S8, we used a real NVP architecture with three Real NVP coupling
 1246 layers and two-layer neural networks of 50 units per layer. The normalizing flow architecture
 1247 mapped $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ to a support of $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V] \in [0.0, 0.5]^4$. EPI optimization was run
 1248 using three different random seeds for architecture initialization $\boldsymbol{\theta}$ with an augmented lagrangian
 1249 coefficient of $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the
 1250 architectures converging with criteria $N_{\text{test}} = 100$ at greatest entropy across three random seeds.

1251 Optimization details are shown in Figure S9. The sums of squares of each pair of parameters are
 1252 shown for each EPI distribution in Figure S10.

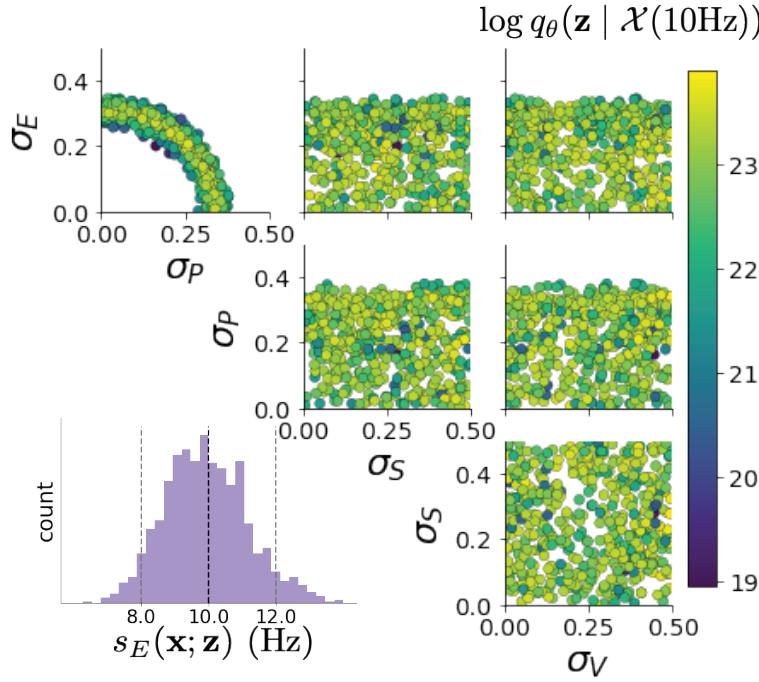


Figure S8: EPI inferred distribution for $\mathcal{X}(10\text{Hz})$.

1253 5.4.3 Sensitivity analyses

1254 In Fig. 3E, we visualize the modes of $q_{\theta}(\mathbf{z} \mid \mathcal{X})$ throughout the σ_E - σ_P marginal. Specifically, we
 1255 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_{\theta}(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (78)$$

1256 At each mode \mathbf{z}^* , we calculated the Hessian and visualized the sensitivity dimension in the direction
 1257 of positive σ_E .

1258 5.4.4 Testing for the paradoxical effect

1259 The paradoxical effect occurs when a populations steady state rate is decreased (or increased)
 1260 when an increase (decrease) in current is applied to that population [12]. To see which, if any,
 1261 populations exhibited a paradoxical effect, we examined responses to changes in input (Fig. S11).

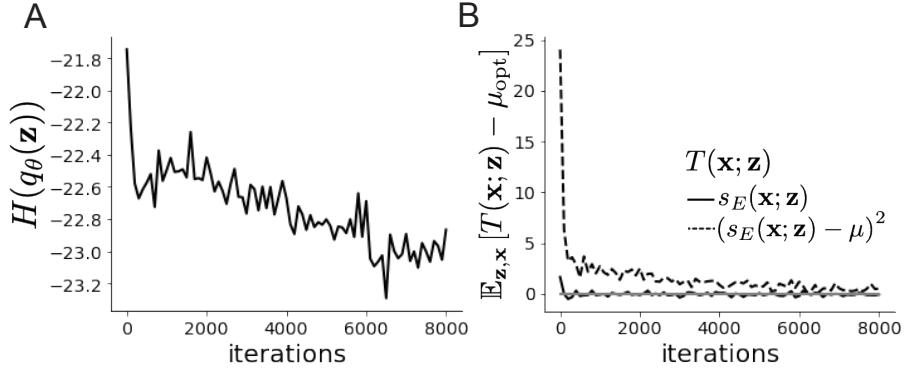


Figure S9: EPI optimization $q_\theta(\mathbf{z} | \mathcal{X}(5\text{Hz}))$ **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 8,000 iterations following the fourth augmented lagrangian epoch.

1262 Input magnitudes were chosen so that the effect is salient (0.002 for E and P, but 0.02 for S and
1263 V). Only the P-population exhibited the paradoxical effect at this connectivity W and input \mathbf{h} .

1264 **5.4.5 Primary visual cortex: Mathematical intuition and challenges**

1265 The dynamical system that we are working with can be written as

$$\begin{aligned} dx &= \frac{1}{\tau}(-x + f(Wx + h + \epsilon))dt \\ d\epsilon &= -\frac{dt}{\tau_{\text{noise}}} \epsilon + \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}} \Sigma_\epsilon dW \end{aligned} \tag{79}$$

1266 Where in this paper we chose

$$\Sigma_\epsilon = \tau_{\text{noise}} \begin{bmatrix} \tilde{\sigma}_E & 0 & 0 & 0 \\ 0 & \tilde{\sigma}_P & 0 & 0 \\ 0 & 0 & \tilde{\sigma}_S & 0 \\ 0 & 0 & 0 & \tilde{\sigma}_V \end{bmatrix} \tag{80}$$

1267 where $\tilde{\sigma}_\alpha$ is the reparameterized standard deviation of the noise for population α from Equation
1268 70.

1269 In order to compute this covariance, we define $v = \omega x + h + \epsilon$ and $S = I - \omega f'(v)$, to re-write Eq.
1270 (79) as an 8-dimensional system:

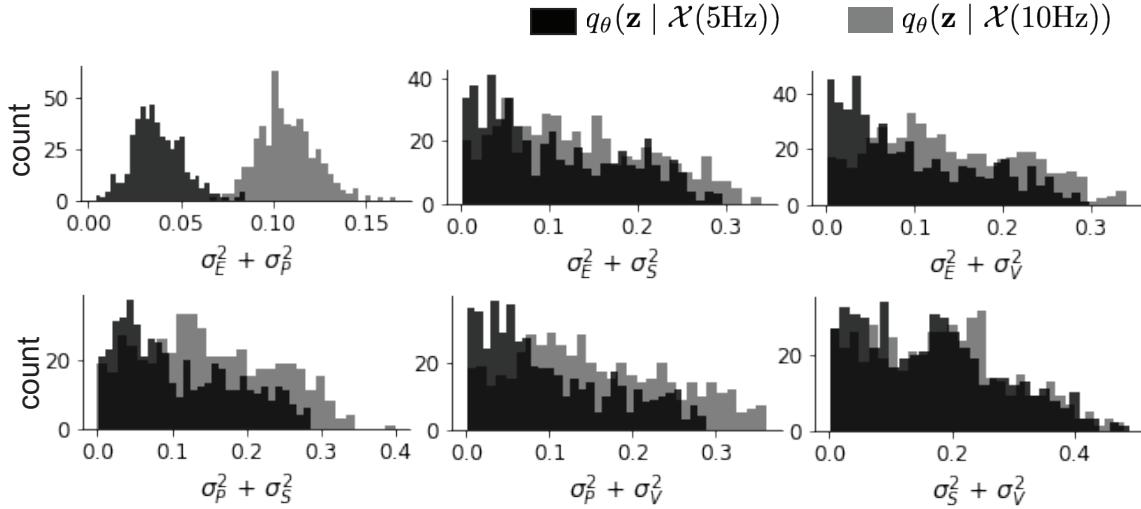


Figure S10: EPI predictive distributions of the sum of squares of each pair of noise parameters.

$$d \begin{pmatrix} \delta v \\ \epsilon \end{pmatrix} = - \begin{pmatrix} S & -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I \\ 0 & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} \begin{pmatrix} \delta v \\ \epsilon \end{pmatrix} dt + \begin{pmatrix} 0 & \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}} \Sigma_\epsilon \\ 0 & \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}} \Sigma_\epsilon \end{pmatrix} d\mathbf{W} \quad (81)$$

1271 Where $d\mathbf{W}$ is a vector with the private noise of each variable. The $d\mathbf{W}$ term is multiplied by a
 1272 non-diagonal matrix is because the noise that the voltage receives is the exact same than the one
 1273 that comes from the OU process and not another process. The solution of this problem is given by
 1274 the Lyapunov Equation [59, 66]:

$$\begin{pmatrix} S & -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I \\ 0 & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} \begin{pmatrix} \Lambda_v & \Lambda_c \\ \Lambda_c^T & \Lambda_\epsilon \end{pmatrix} + \begin{pmatrix} \Lambda_v & \Lambda_c \\ \Lambda_c^T & \Lambda_\epsilon \end{pmatrix} \begin{pmatrix} S^T & 0 \\ -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} = \begin{pmatrix} \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon & \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon \\ \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon & \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon \end{pmatrix} \quad (82)$$

1275 To obtain an equation for Λ_v , we solve this block matrix multiplication:

$$S\Lambda_v + \Lambda_v S^T = \frac{2\Lambda_\epsilon}{\tau_{\text{noise}}} + \frac{\tau_{\text{noise}}^2 - \tau^2}{(\tau \tau_{\text{noise}})^2} \left(\left(\frac{1}{\tau_{\text{noise}}} I + S \right)^{-1} \Lambda_\epsilon + \Lambda_\epsilon \left(\frac{1}{\tau_{\text{noise}}} I + S^T \right)^{-1} \right) \quad (83)$$

Which is another Lyapunov Equation, now in 4 dimensions. In the simplest case in which $\tau_{\text{noise}} = \tau$, the voltage is directly driven by white noise, and Λ_v can be expressed in powers of S and S^T . Because S satisfies its own polynomial equation (Cayley Hamilton theorem), there will be 4 coefficients for the expansion of S and 4 for S^T , resulting in 16 coefficients that define Λ_v for a

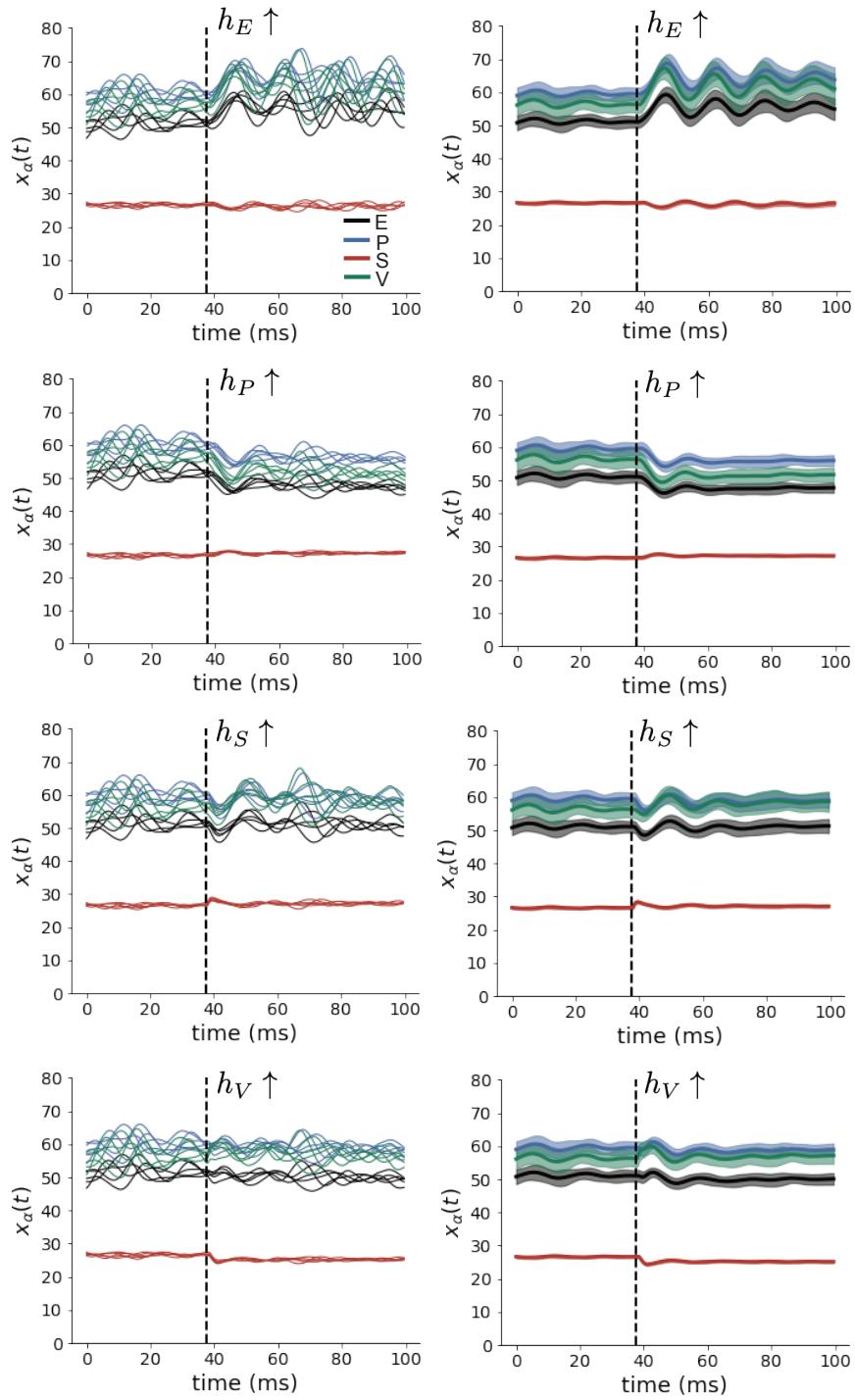


Figure S11: (Left) SSSN simulations for small increases in neuron-type population input. (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

given S . Due to symmetry arguments [66], in this case the diagonal elements of the covariance matrix of the voltage will have the form:

$$\Lambda_{v_{ii}} = \sum_{i=\{E,P,S,V\}} g_i(S) \sigma_{ii}^2 \quad (84)$$

1276 These coefficients $g_i(S)$ are complicated functions of the Jacobian of the system. Although expres-
 1277 sions for these coefficients can be found explicitly, only numerical evaluation of those expressions
 1278 determine which components of the noisy input are going to strongly influence the variability of ex-
 1279 citatory population. Showing the generality of this dependence in more complicated noise scenarios
 1280 (e.g. $\tau_{\text{noise}} > \tau$ as in Section 3.4), is the focus of current research.

1281 5.5 Superior colliculus

1282 5.5.1 SC model

1283 The ability to switch between two separate tasks throughout randomly interleaved trials, or “rapid
 1284 task switching,” has been studied in rats, and midbrain superior colliculus (SC) has been show to
 1285 play an important in this computation [67]. Neural recordings in SC exhibited two populations of
 1286 neurons that simultaneously represented both task context (Pro or Anti) and motor response (con-
 1287 tralateral or ipsilateral to the recorded side), which led to the distinction of two functional classes:
 1288 the Pro/Contra and Anti/Ipsi neurons [48]. Given this evidence, Duan et al. proposed a model
 1289 with four functionally-defined neuron-type populations: two in each hemisphere corresponding to
 1290 the Pro/Contra and Anti/Ipsi populations. We study how the connectivity of this neural circuit
 1291 governs rapid task switching ability.

1292 The four populations of this model are denoted as left Pro (LP), left Anti (LA), right Pro (RP)
 1293 and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh \left(\frac{u_\alpha - a}{b} \right) + \frac{1}{2} \right), \quad (85)$$

1294 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-
 1295 earity. We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (86)$$

1296 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (87)$$

1297 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2 . These
1298 hyperparameter values are motivated by modeling choices and results from [48].

1299 The weight matrix has 4 parameters for self sW , vertical vW , horizontal hW , and diagonal dW
1300 connections:

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (88)$$

1301 We study the role of parameters $\mathbf{z} = [sW, vW, hW, dW]^\top$ in rapid task switching.

1302 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (89)$$

1303 There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (90)$$

1304 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (91)$$

1305 rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (92)$$

1306

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (93)$$

1307 a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (94)$$

1308 and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases}. \quad (95)$$

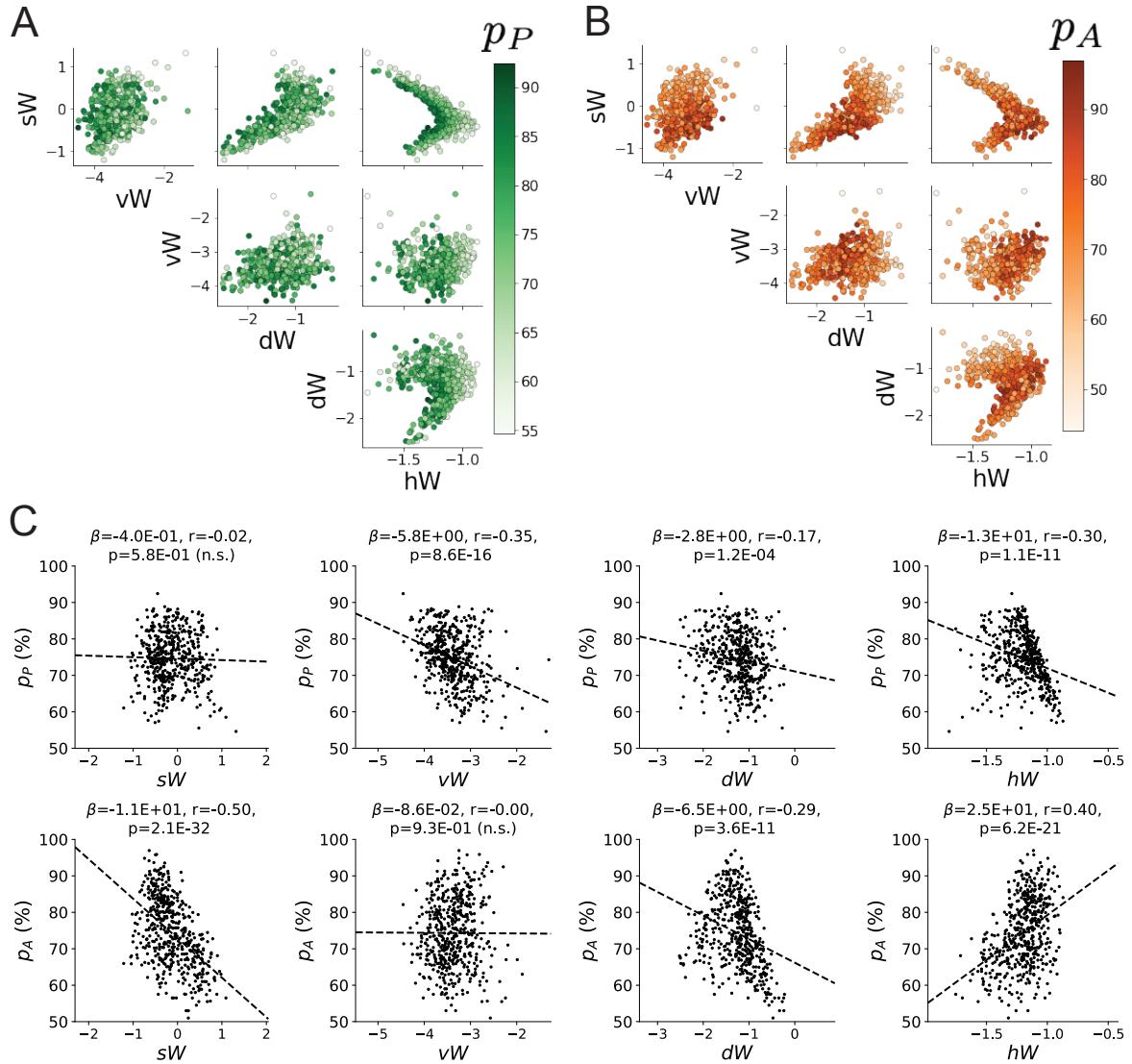


Figure S12: **A.** Same pairplot as Fig. 4C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

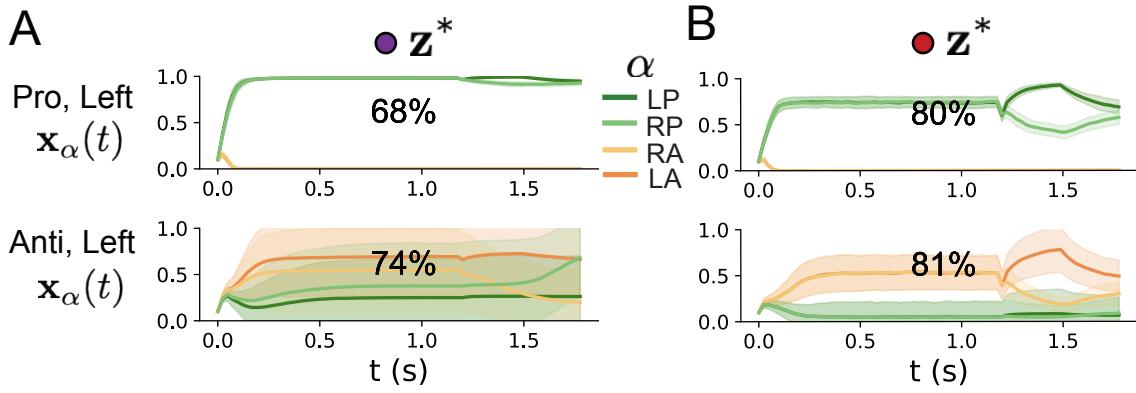


Figure S13: **A.** Simulations in network regime 1: $\mathbf{z}^*(sW = -0.75)$. **B.** Simulations in network regime 2: $\mathbf{z}^*(sW = 0.75)$.

¹³⁰⁹ The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{\text{P,bias}} = 0.5$, $I_{\text{P,rule}} = 0.6$, $I_{\text{A,rule}} = 0.6$,
¹³¹⁰ $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

¹³¹¹ 5.5.2 Task accuracy calculation

¹³¹² The accuracies of each Pro and Anti tasks are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (96)$$

¹³¹³ and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (97)$$

¹³¹⁴ given that the stimulus is on the left side and Θ approximates the Heaviside step function. Our
¹³¹⁵ accuracy calculation only considers one stimulus presentation (Left), since the model is left-right
¹³¹⁶ symmetric. The accuracy is averaged over 200 independent trials, and the Heaviside step function
¹³¹⁷ is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta_{\Theta}\mathbf{x}), \quad (98)$$

¹³¹⁸ where $\beta_{\Theta} = 100$.

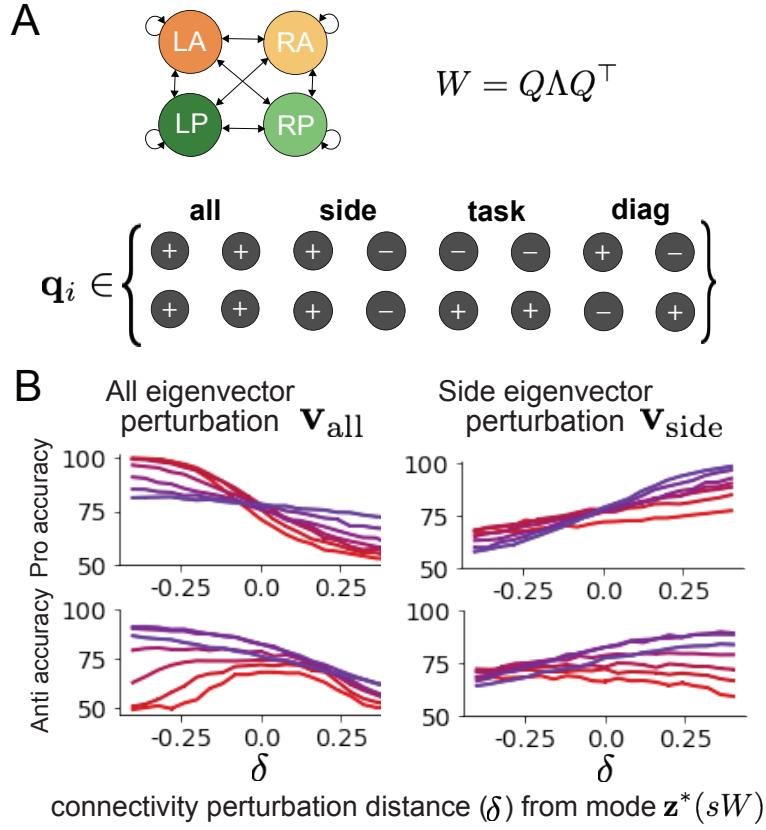


Figure S14: **A.** Invariant eigenvectors of connectivity matrix W . **B.** Accuracies for connectivity perturbations when changing λ_{all} and λ_{side} (λ_{task} and λ_{diag} shown in Fig. 4D).

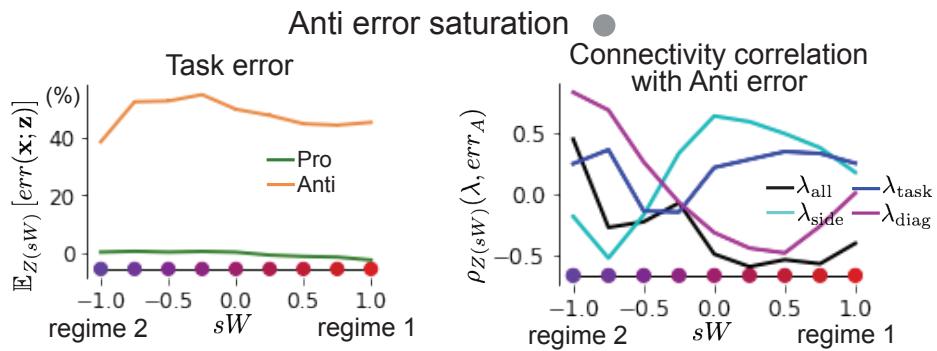


Figure S15: (Left) Mean and standard error of Pro and Anti error from regime 1 to regime 2 at $\gamma = 0.85$. (Right) Correlations of connectivity eigenvalues with Anti error from regime 1 to regime 2 at $\gamma = 0.85$.

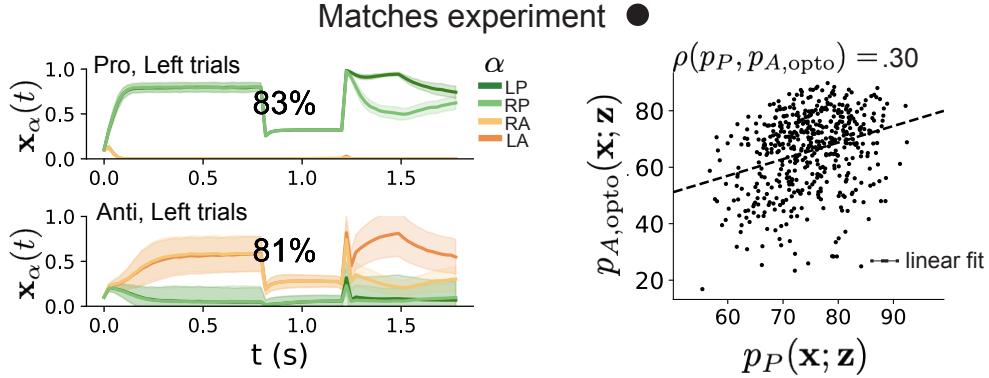


Figure S16: (Left) Mean and standard deviation (shading) of responses of the SC model at the mode of the EPI distribution to delay period inactivation at $\gamma = 0.675$. Accuracy in Pro (top) and Anti (bottom) task is shown as a percentage. (Right) Anti accuracy following delay period inactivation at $\gamma = 0.675$ versus accuracy in the Pro task across connectivities in the EPI distribution.

1319 5.5.3 EPI details for the SC model

1320 Writing the EPI distribution as a maximum entropy distribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these
1321 first and second moments of the accuracy in each task (as in Equations 20 and 21)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \\ (p_P(\mathbf{x}; \mathbf{z}) - .75)^2 \\ (p_A(\mathbf{x}; \mathbf{z}) - .75)^2 \end{bmatrix}, \quad (99)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} .75 \\ .75 \\ .075^2 \\ .075^2 \end{bmatrix}. \quad (100)$$

1323 Throughout optimization, the augmented lagrangian parameters η and c , were updated after each
1324 epoch of 2,000 iterations (see Section 5.1.4). The optimization converged after ten epochs (Fig.
1325 S16).

1326 For EPI in Fig. 4C, we used a real NVP architecture with three coupling layers of affine transfor-
1327 mations parameterized by two-layer neural networks of 50 units per layer. The initial distribution
1328 was a standard isotropic gaussian $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to a support of $\mathbf{z}_i \in [-5, 5]$. We used an
1329 augmented lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution
1330 was the greatest EPI distribution to converge across 5 random seeds with criteria $N_{\text{test}} = 25$.

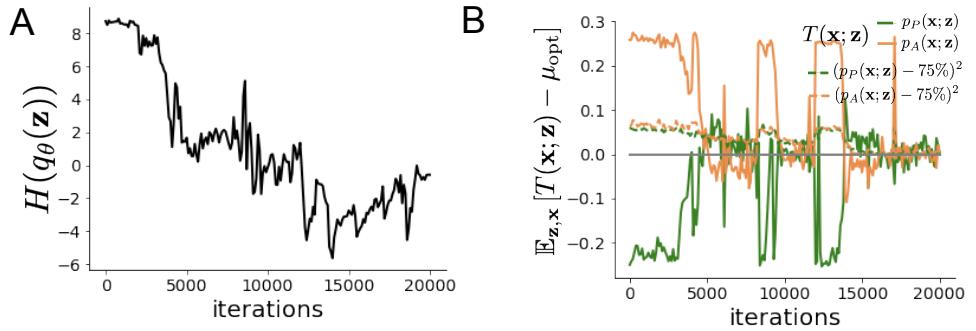


Figure S17: EPI optimization of the SC model producing rapid task switching. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented lagrangian epoch.

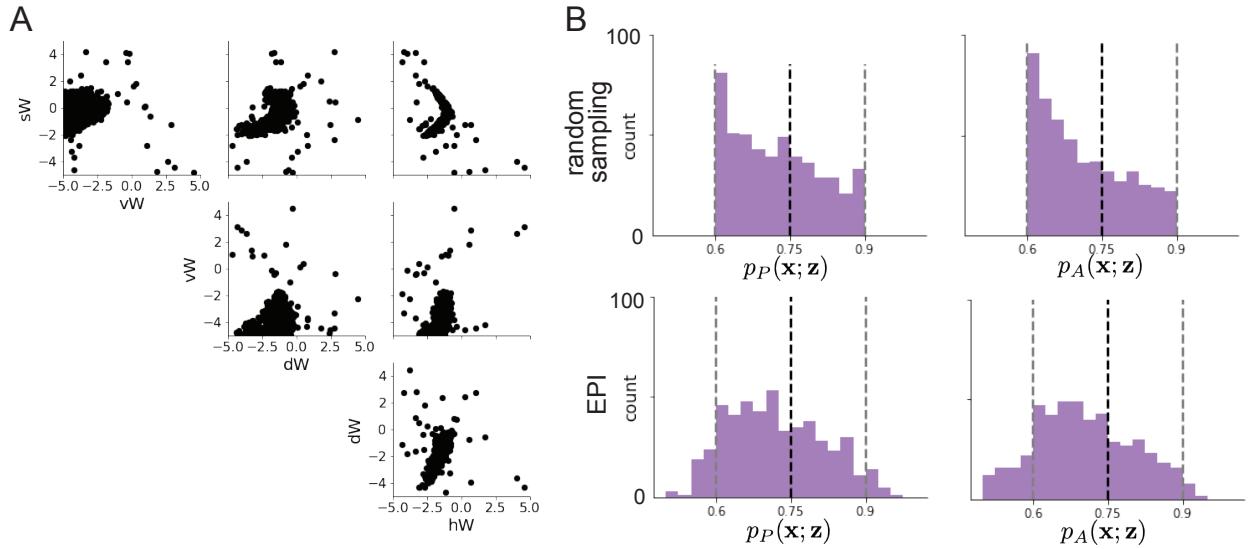


Figure S18: **A.** Rapid task switching SC connectivities obtained from random sampling. **B.** Task accuracies of the inferred distributions from random sampling (top) and EPI (bottom).

1331 The bend in the EPI distribution is not a spurious result of the EPI optimization. The structure
1332 discovered by EPI matches the shape of the set of points returned from brute-force random sampling
1333 (Fig. S18A) These connectivities were sampled from a uniform distribution over the range of each
1334 connectivity parameter, and all parameters producing accuracy in each task within the range of
1335 60% to 90% were kept. This set of connectivities will not match the distribution of EPI exactly,
1336 since it is not conditioned on the emergent property. For example the parameter set returned by
1337 the brute-force search is biased towards lower accuracies (Fig. S18B).

1338 5.5.4 Mode identification with EPI

1339 We found one mode of the EPI distribution for fixed values of sW from 1 to -1 in steps of 0.25.
1340 To begin, we chose an initial parameter value from 500 parameter samples $\mathbf{z} \sim q_{\theta}(\mathbf{z} \mid \mathcal{X})$ that
1341 had closest sW value to 1. We then optimized this estimate of the mode (for fixed sW) using
1342 probability gradients of the deep probability distribution for 500 steps of gradient ascent with a
1343 learning rate of 5×10^{-3} . The next mode (at $sW = 0.75$) was found using the previous mode as
1344 the initialization. This and all subsequent optimizations used 200 steps of gradient ascent with a
1345 learning rate of 1×10^{-3} , except at $sW = -1$ where a learning rate of 5×10^{-4} was used. During all
1346 mode identification optimizations, the learning rate was reduced by half (decay = 0.5) after every
1347 100 iterations.

1348 5.5.5 Sample grouping by mode

1349 For the analyses in Figure 5C and Figure S15, we obtained parameters for each step along the
1350 continuum between regimes 1 and 2 by sampling from the EPI distribution. Each sample was
1351 assigned to the closest mode $\mathbf{z}^*(sW)$ (Equation 12). Sampling continued until 500 samples were
1352 assigned to each mode, which took 2.67 seconds (5.34ms/sample-per-mode). It took 9.59 minutes
1353 to obtain just 5 samples for each mode with brute force sampling requiring accuracies between 60%
1354 and 90% in each task (115s/sample-per-mode). This corresponds to a sampling speed increase of
1355 roughly 21,500 once the EPI distribution has been learned.

1356 5.5.6 Sensitivity analysis

1357 At each mode, we measure the sensitivity dimension (that of most negative eigenvalue in the Hessian
1358 of the EPI distribution) $\mathbf{v}_1(\mathbf{z}^*)$. To resolve sign degeneracy in eigenvectors, we chose $\mathbf{v}_1(\mathbf{z}^*)$ to have

1359 negative element in hW . This tells us what parameter combination rapid task switching is most
 1360 sensitive to at this parameter choice in the regime.

1361 **5.5.7 Connectivity eigendecomposition and processing modes**

1362 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-
 1363 tion of the connectivity matrices $W = Q\Lambda Q^{-1}$, which results in the same eigenmodes \mathbf{q}_i for all W
 1364 parameterized by \mathbf{z} (Fig. S14A). These eigenvectors are always the same, because the connectivity
 1365 matrix is symmetric and the model also assumes symmetry across hemispheres, but the eigenvalues
 1366 of connectivity (or degree of eigenmode amplification) change with \mathbf{z} . These basis vectors have in-
 1367 tuitive roles in processing for this task, and are accordingly named the *all* eigenmode - all neurons
 1368 co-fluctuate, *side* eigenmode - one side dominates the other, *task* eigenmode - the Pro or Anti pop-
 1369 ulations dominate the other, and *diag* mode - Pro- and Anti-populations of opposite hemispheres
 1370 dominate the opposite pair. Due to the parametric structure of the connectivity matrix, the pa-
 1371 rameters \mathbf{z} are a linear function of the eigenvalues $\boldsymbol{\lambda} = [\lambda_{\text{all}}, \lambda_{\text{side}}, \lambda_{\text{task}}, \lambda_{\text{diag}}]^T$ associated with these
 1372 eigenmodes.

$$\mathbf{z} = A\boldsymbol{\lambda} \quad (101)$$

$$1373 \quad A = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}. \quad (102)$$

1374 We are interested in the effect of raising or lowering the amplification of each eigenmode in the
 1375 connectivity matrix by perturbing individual eigenvalues λ . To test this, we calculate the unit
 1376 vector of changes in the connectivity \mathbf{z} that result from a change in the associated eigenvalues

$$\mathbf{v}_a = \frac{\frac{\partial \mathbf{z}}{\partial \lambda_a}}{\left| \frac{\partial \mathbf{z}}{\partial \lambda_a} \right|_2}, \quad (103)$$

1377 where

$$\frac{\partial \mathbf{z}}{\partial \lambda_a} = A\mathbf{e}_a, \quad (104)$$

1378 and e.g. $\mathbf{e}_{\text{all}} = [1, 0, 0, 0]^T$. So \mathbf{v}_a is the normalized column of A corresponding to eigenmode
 1379 a . The parameter dimension \mathbf{v}_a ($a \in \{\text{all}, \text{side}, \text{task}, \text{and diag}\}$) that increases the eigenvalue of
 1380 connectivity λ_a is \mathbf{z} -invariant (Equation 104) and $\mathbf{v}_a \perp \mathbf{v}_{b \neq a}$. By perturbing \mathbf{z} along \mathbf{v}_a , we
 1381 can examine how model function changes by directly modulating the connectivity amplification of
 1382 specific eigenmodes, which having interpretable roles in processing in each task.

1383 **5.5.8 Modeling optogenetic silencing.**

1384 We tested whether the inferred SC model connectivities could reproduce experimental effects of
1385 optogenetic inactivation in rats [48]. During periods of simulated optogenetic inactivation, activity
1386 was decreased proportional to the optogenetic strength $\gamma \in [0, 1]$

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (105)$$

1387 Delay period inactivation was from $0.8 < t < 1.2$.