

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

¹Department of Neuroscience, Columbia University,

²Princeton Neuroscience Institute,

³Princeton University,

⁴Allen Institute for Brain Science,

⁵Institute of Neuroscience, Chinese Academy of Sciences,

⁶Howard Hughes Medical Institute,

⁷Department of Statistics, Columbia University

¹ 1 Abstract

² A cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures
³ a hypothesized neural mechanism. Such models are valuable when they give rise to an experimen-
⁴ tally observed phenomenon – whether behavioral or 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 choices of model parameters. A key step is then to identify the model parameters
⁷ consistent with observed phenomena: to solve the inverse problem. To solve challenging inverse
⁸ problems concerning neural datasets, neuroscientists have used statistical inference techniques to
⁹ much success. However, when theorizing circuit models, scientists predominantly focus on repro-
¹⁰ ducing computational properties rather than explaining a given neural dataset well. In this work,
¹¹ we present a novel technique, emergent property inference (EPI), that brings the power and ver-
¹² satility of the modern probabilistic modeling toolkit to theoretical neuroscience. Our method uses
¹³ deep neural networks to learn parameter distributions with complex structure that produce specific
¹⁴ computational properties in circuit models. This methodology is introduced through a motivational
¹⁵ example inferring conductance parameters in a circuit model of the stomatogastric ganglion. Then,
¹⁶ with recurrent neural networks of increasing size, we show that EPI allows precise control over
¹⁷ the behavior of inferred parameters, and that EPI scales better in parameter dimension than al-
¹⁸ ternative techniques. In the remainder of this work, we present novel theoretical findings gained
¹⁹ through the examination of complex parametric structure captured by EPI. In a model of primary
²⁰ visual cortex, we discovered how connectivity with multiple inhibitory subtypes shapes variability

21 in the excitatory population. Finally, in a model of superior colliculus, we identified and charac-
22 terized two distinct regimes of connectivity that facilitate switching between opposite tasks amidst
23 interleaved trials, mechanistically characterized each regime using probabilistic tools afforded by
24 EPI, and found conditions where these circuit models reproduce results from optogenetic silencing
25 experiments. Beyond its scientific contribution, this work illustrates the variety of analyses possible
26 once deep learning is harnessed towards solving theoretical inverse problems.

27 2 Introduction

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

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

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

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

51 data. Related to this point, use of a conventional dataset encourages conventional data likelihoods
52 or loss functions, which focus on some global metric like squared error or evidence, rather than
53 the computation itself. Alternatively, researchers often quantify an *emergent property*: a statistic
54 of data that directly quantifies the computation of interest, wherein the dataset is only implicit.
55 While such a choice may seem esoteric, it is not: the above “gold standard” examples all quantify
56 and focus on some derived feature of the data, rather than the data drawn from the model.

57 An emergent property (EP) is of course a dataset by another name, but it suggests different ap-
58 proach to solving the same inverse problem: here we directly specify the desired emergent property
59 – a statistic of datasets drawn from the model – and the value we wish that property to have,
60 and we set up an optimization program to find the distribution of parameters that produce this
61 computation. This statistical framework is not new: it is intimately connected to the literature
62 on approximate bayesian computation [24–26], parameter sensitivity analyses [27–30], maximum
63 entropy modeling [31–33], and approximate bayesian inference [34, 35]; we detail these connections
64 in Section 5.1.1. However, the adaptation of modern techniques to the problem of theoretical cir-
65 cuit analysis requires recent developments in deep learning for constrained optimization [36], and
66 architectural choices for scalable, flexible generative modeling [37, 38]. We detail our method, which
67 we call emergent property inference (EPI) in Section 3.2.

68 Equipped with this method, we prove out EPI by demonstrating its capabilities and presenting
69 novel theoretical findings from its analysis. First, we show EPI’s ability to handle biologically
70 realistic circuit models using a five-neuron model of the stomatogastric ganglion [39]: a neural
71 circuit whose parametric degeneracy is closely studied [40]. Then, we show EPI’s scalability to
72 high dimensional parameter distributions by inferring connectivities of recurrent neural networks
73 (RNNs) that exhibit stable, yet amplified responses – a hallmark of neural responses throughout
74 the brain [41–43]. In a model of primary visual cortex [44, 45], EPI reveals the how recurrent
75 processing across different neuron-type populations shapes excitatory variability: a finding that
76 we show is analytically intractable. Finally, we investigated the possible connectivities of superior
77 colliculus that allow execution of different tasks on interleaved trials [46]. EPI discovered a rich
78 distribution containing two connectivity regimes with different solution classes. We queried the
79 deep probability distribution learned by EPI to produce a mechanistic understanding of cortical
80 responses in each regime. Intriguingly, the inferred connectivities of each regime reproduced results
81 from optogenetic inactivation experiments in markedly different ways. These theoretical insights
82 afforded by EPI illustrate the value of deep inference for the interrogation of neural circuit models.

83 **3 Results**

84 **3.1 Motivating emergent property inference of theoretical models**

85 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-
86 erty inference. First, one designs or chooses an existing circuit model that, it is hypothesized,
87 captures the computation of interest. To ground this process in a well-known example, consider
88 the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple
89 rhythmic muscle activation patterns for digestion [47]. Despite full knowledge of STG connectivity
90 and a precise characterization of its rhythmic pattern generation, biophysical models of the STG
91 have complicated relationships between circuit parameters and computation [15, 40].

92 A subcircuit model of the STG [39] is shown schematically in Figure 1A. The fast population (f_1
93 and f_2) represents the subnetwork generating the pyloric rhythm and the slow population (s_1 and
94 s_2) represents the subnetwork of the gastric mill rhythm. The two fast neurons mutually inhibit
95 one another, and spike at a greater frequency than the mutually inhibiting slow neurons. The
96 hub neuron couples with either the fast or slow population, or both depending on modulatory
97 conditions. The jagged connections indicate electrical coupling having electrical conductance g_{el} ,
98 smooth connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto
99 the hub neuron, and $g_{synB} = 5nS$ for mutual inhibitory connections. Note that the behavior of this
100 model will be critically dependent on its parameterization – the choices of conductance parameters
101 $\mathbf{z} = [g_{el}, g_{synA}]$.

102 Second, once the model is selected, one must specify what the model should produce. In this
103 STG model, we are concerned with neural spiking frequency, which emerges from the dynamics of
104 the circuit model (Fig. 1B). An emergent property studied by Gutierrez et al. of this stochastic
105 model is the hub neuron firing at an intermediate frequency between the intrinsic spiking rates
106 of the fast and slow populations. This emergent property is shown in Figure 1C at an average
107 frequency of 0.55Hz. Our notion of intermediate hub frequency is not strictly 0.55Hz, but also
108 moderate deviations of this frequency between the fast (.35Hz) and slow (.68Hz) frequencies, which
109 are quantified in the emergent property with variance 0.025^2Hz^2 .

110 Third, the model parameters producing the emergent property are inferred. To infer the STG
111 parameters of intermediate hub frequency with existing methodology, we need an exemplar dataset:
112 experimentally recorded or synthesized. By precisely quantifying the emergent property of interest
113 as a statistical feature of the model, we use EPI to condition directly on this emergent property. EPI

114 learns a probability distribution of model parameters constrained to produce the emergent property.
115 In this last step lies the opportunity for a shift away from a dataset-oriented representation of model
116 output towards that of an implicit dataset, where the only structure is the emergent property of
117 interest.

118 Before presenting technical details (in the following section), let us understand emergent property
119 inference schematically. EPI (Fig. 1D) takes, as input, the model and the specified emergent prop-
120 erty, and as its output, produces the parameter distribution EPI (Fig. 1E). This distribution –
121 represented for clarity as samples from the distribution – is a parameter distribution that produces
122 the emergent property. We can then use this parameter distribution to efficiently generate many
123 parameters producing the emergent property. Most importantly, the resulting parameter distribu-
124 tion can be efficiently queried to quantify its own structure, which informs the central question of
125 this inverse problem: how do model parameters govern the emergent property?

126 3.2 A deep generative modeling approach to emergent property inference

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

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

132 where $C_m=1nF$, 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 [39] (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,

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

140 and that the variance of the hub frequency is moderate

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

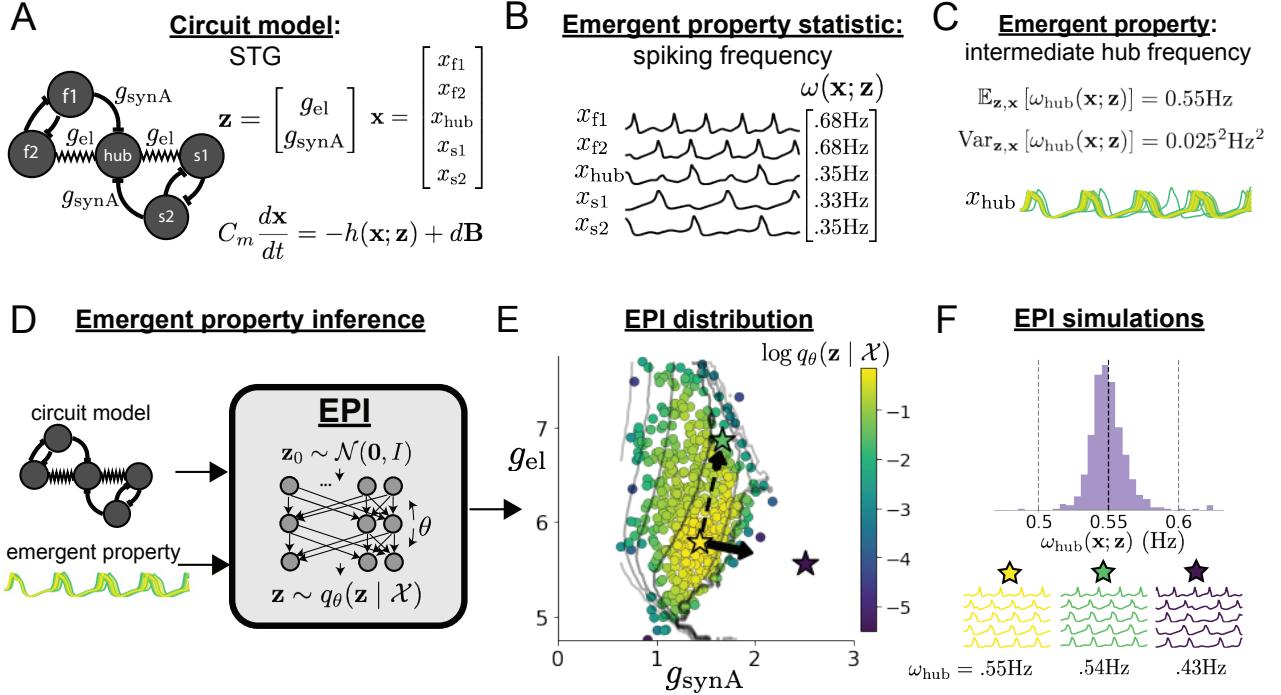


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based subcircuit model of the STG. **B.** Spiking frequency $\omega(\mathbf{x}; \mathbf{z})$ is an emergent property statistic. Simulated at $g_{el} = 4.5\text{nS}$ and $g_{synA} = 3\text{nS}$. **C.** The emergent property of intermediate hub frequency. Simulated activity traces are colored by $\log q_\theta(\mathbf{z} | \mathcal{X})$ of generating parameters. (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, .53,575 Hz (dark to light gray away from mean). Dimension of sensitivity \mathbf{v}_1 (solid) and robustness \mathbf{v}_2 (dashed). **F** (Top) The predictive distribution of EPI. The black and gray dashed lines show the mean and two standard deviations according the emergent property. (Bottom) Simulations at the starred parameter values.

141 In the emergent property of intermediate hub frequency, the statistic of hub neuron frequency is
 142 constrained over the distribution of parameters \mathbf{z} and the distribution of the data \mathbf{x} that those
 143 parameters produce. Formally, the emergent property is the collection of these two constraints

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [\omega_{\text{hub}}(\mathbf{x}; \mathbf{z})] = 0.55\text{Hz}, \quad \text{Var}_{\mathbf{z}, \mathbf{x}} [\omega_{\text{hub}}(\mathbf{x}; \mathbf{z})] = 0.025^2\text{Hz}^2. \quad (4)$$

144 In general, an emergent property is a collection of constraints on first-, second- and higher moments
 145 of one or more statistics (see Sections 3.3, 3.5, and 5.1.5) that together define the computation.

146 Third, we perform emergent property inference: we find a distribution over parameter configu-
 147 rations \mathbf{z} that produces the emergent property; in other words, they obey the constraints intro-
 148 duced in Equation 4. This distribution will be chosen from a family of probability distributions
 149 $\mathcal{Q} = \{q_{\theta}(\mathbf{z}) : \theta \in \Theta\}$, defined by a deep neural network [37, 48, 49] (Figure 1D, EPI box). Deep
 150 probability distributions map a simple random variable \mathbf{z}_0 (drawn from an isotropic gaussian)
 151 through a deep neural network with weights and biases θ to parameters $\mathbf{z} = g_{\theta}(\mathbf{z}_0)$ of a suitably
 152 complicated distribution (see Section 5.1.2 for more details). Many distributions in \mathcal{Q} will respect
 153 the emergent property constraints, so we select the most random (highest entropy) distribution,
 154 which is the same choice made in variational bayesian methods (see Section 5.1.6). In EPI opti-
 155 mization, stochastic gradient steps in θ are taken such that entropy is maximized, and the emergent
 156 property \mathcal{X} is produced (see Section 5.1). The inferred EPI distribution is denoted $q_{\theta}(\mathbf{z} | \mathcal{X})$, since
 157 it is conditioned upon the implicit datasets of emergent property \mathcal{X} in a manner methodologically
 158 equivalent to exemplar data when using variational bayesian approaches.

159 EPI produces parameter distributions that can be queried for scientific insight. The modes of
 160 $q_{\theta}(\mathbf{z} | \mathcal{X})$ indicate parameter choices emblematic of the emergent property (Fig. 1E yellow star).
 161 As probability in the EPI distribution decreases, the emergent property deteriorates. Perturbing
 162 \mathbf{z} along a dimension in which $q_{\theta}(\mathbf{z} | \mathcal{X})$ does not change will not disturb the emergent property,
 163 making this parameter combination *robust* with respect to the emergent property. In contrast, if
 164 \mathbf{z} is perturbed along a dimension that strongly decreases $q_{\theta}(\mathbf{z} | \mathcal{X})$, that parameter combination is
 165 deemed *sensitive*. By querying the second order derivative (Hessian) of $\log q_{\theta}(\mathbf{z} | \mathcal{X})$ at a mode, we
 166 can quantitatively identify how sensitive (or robust) each eigenvector is by its eigenvalue; the more
 167 negative, the more sensitive and the closer to zero, the more robust (see Section 5.2.4). Indeed,
 168 samples equidistant from the mode along these EPI-identified dimensions of sensitivity (\mathbf{v}_1 , smaller
 169 eigenvalue) and robustness (\mathbf{v}_2 , greater eigenvalue) (Fig. 1E, arrows) agree with error contours
 170 (Fig. 1E contours) and have diminished or preserved hub frequency, respectively (Fig. 1F activity
 171 traces). This suggests that changes in conductance along the parameter combination described

172 by \mathbf{v}_2 will most preserve hub neuron firing between the intrinsic rates of the pyloric and gastric
 173 mill rhythms. Once an EPI distribution has been inferred, this Hessian calculation requires trivial
 174 computation (see Section 5.1.2).

175 In the following sections, we demonstrate EPI on three neural circuit models across ranges of
 176 biological realism, neural system function, and network scale. First, we demonstrate the superior
 177 scalability of EPI compared to alternative techniques by inferring high-dimensional distributions of
 178 recurrent neural network (RNN) connectivities that exhibit amplified, yet stable responses. Also
 179 in this RNN example, we emphasize that EPI is the only technique that controls the predictions
 180 made by the inferred parameter distribution. Next, in a model of primary visual cortex [44, 45], we
 181 show how EPI captures a curved manifold of parametric degeneracy, revealing how input variability
 182 across neuron types affects the excitatory population. Finally, in a model of superior colliculus [46],
 183 we used EPI to capture multiple parametric regimes of task switching, and queried the dimensions
 184 of sensitivity ($\mathbf{v}_1(\mathbf{z})$) to mechanistically characterize each regime.

185 3.3 Scaling inference of RNN connectivity with EPI

186 Transient amplification is a hallmark of neural activity throughout cortex, and is often thought to be
 187 intrinsically generated by recurrent connectivity in the responding cortical area [41–43]. It has been
 188 shown that to generate such amplified, yet stabilized responses, the connectivity of RNNs must be
 189 non-normal [41, 50], and satisfy additional constraints [51]. In theoretical neuroscience, RNNs are
 190 optimized and then examined to show how dynamical systems could execute a given computation
 191 [52, 53], but such biologically realistic constraints on connectivity are ignored during optimization
 192 for practical reasons. In general, access to distributions of connectivity adhering to theoretical
 193 criteria like stable amplification, chaotic fluctuations [10], or low tangling [54] would add scientific
 194 value and context to existing research with RNNs. Here, we use EPI to learn RNN connectivities
 195 producing stable amplification, and demonstrate the superior scalability and efficiency of EPI to
 196 alternative approaches.

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

$$198 \tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}, \quad (5)$$

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 parameterizations $\mathbf{z} = [\mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{V}_1^\top, \mathbf{V}_2^\top]^\top$ that produce stable
 200 amplification. Two conditions are necessary and sufficient for RNNs to exhibit stable amplification

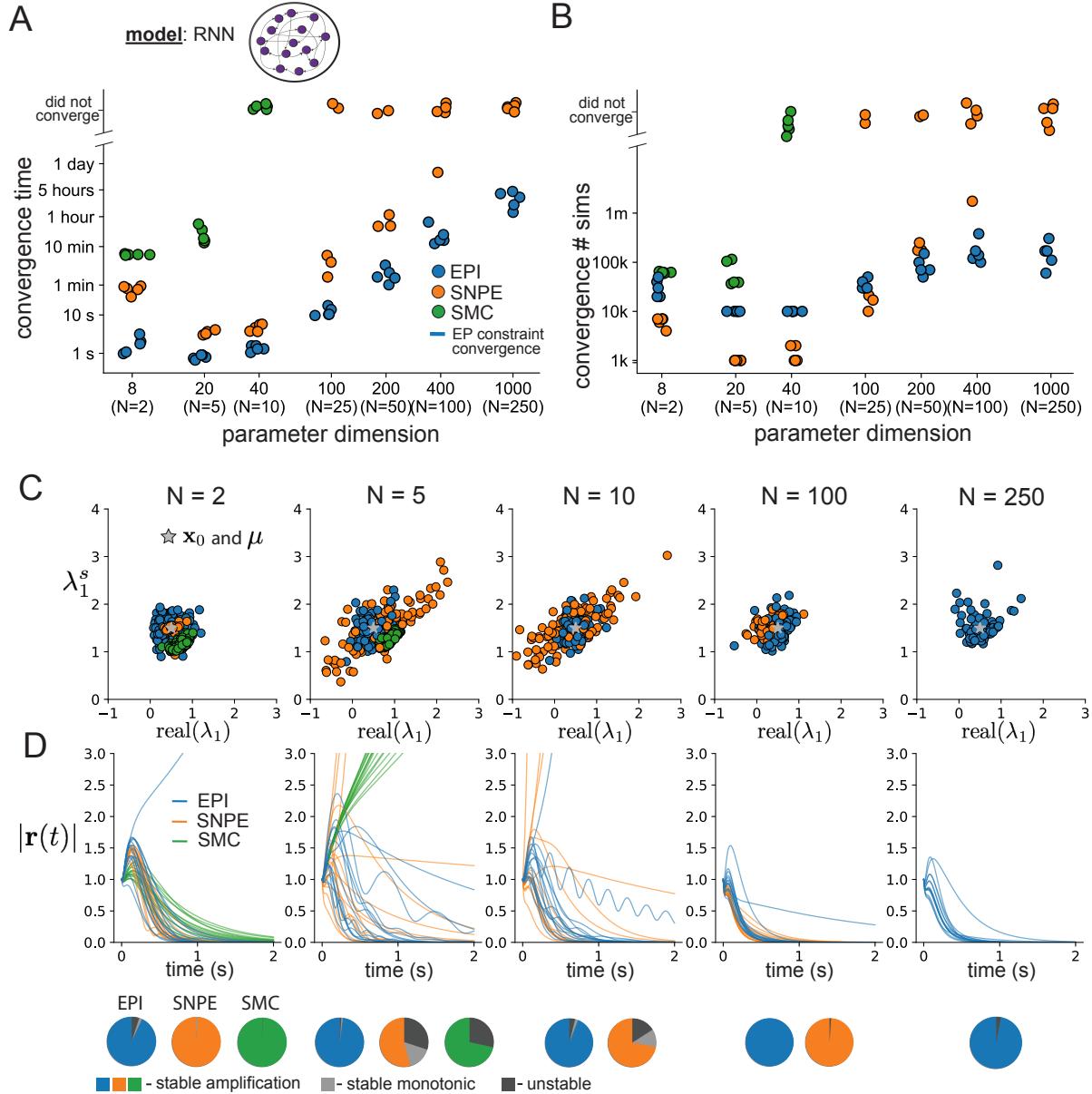


Figure 2: **A.** Wall time of EPI (blue), SNPE (orange), and SMC-ABC (green) to converge on RNN connectivities producing stable amplification. Each dot shows convergence time for an individual random seed. For reference, the mean wall time for EPI to achieve its full constraint convergence (means and 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 $\mathbf{x}_0 = \boldsymbol{\mu}$ (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.

[51]: $\text{real}(\lambda_1) < 1$ and $\lambda_1^s > 1$, where λ_1 is the eigenvalue of W with greatest real part and λ^s is the maximum 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 modest decay rate ($\text{real}(\lambda_1)$ close to its upper bound of 1) and exhibit modest amplification (λ_1^s close to its lower bound of 1). EPI can naturally condition on this emergent property

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}, \end{aligned} \quad (6)$$

under the notion that variance constraints with standard deviation 0.25 predicate that the vast majority of samples (those within two standard deviations) are within the specified ranges.

For comparison, we infer the parameters \mathbf{z} likely to produce stable amplification using two alternative simulation-based inference approaches. We ran sequential Monte Carlo approximate Bayesian computation (SMC-ABC) [26] and sequential neural posterior estimation (SNPE) [35] with exemplary dataset $\mathbf{x}_0 = \boldsymbol{\mu}$. SMC-ABC is a rejection sampling approach that uses SMC techniques to improve efficiency, and SNPE approximates posteriors with deep probability distributions using a two-network architecture (see Section 5.1.1). Unlike EPI, these statistical inference techniques do not constrain the statistics of the predictive distribution, and these predictions of the inferred posteriors are typically affected by model characteristics (e.g. N and g , Fig. 11). To compare the efficiency of these different techniques, we measured the time and number of simulations necessary for the distance of the predictive mean to be less than 0.5 from $\boldsymbol{\mu} = \mathbf{x}_0$ (see Section 5.3).

As the number of neurons N in the RNN is scaled, and thus the dimension of the parameter space $\mathbf{z} \in [-1, 1]^{4N}$, we see that EPI converges at greater speed and at greater dimension than SMC-ABC and SNPE (Fig. 2A). It also becomes most efficient to use EPI in terms of simulation count at $N = 50$ (Fig. 2B). It is well-known that ABC techniques struggle in parameter spaces of modest dimension [55], yet we were careful to assess the scalability of SNPE, which is a more closely related methodology to EPI. Between EPI and SNPE, we closely controlled the number of parameters in deep probability distributions by dimensionality (Fig. 10), and tested more aggressive SNPE hyperparameterizations when SNPE failed to converge (Fig. 12). In this analysis, we see that deep inference techniques EPI and SNPE are far more amenable to inference of high dimensional RNN connectivities than rejection sampling techniques like SMC-ABC, and that EPI outperforms SNPE in both wall time and simulation count.

229 No matter the number of neurons, EPI always produces connectivity distributions with mean and
230 variance of $\text{real}(\lambda_1)$ and λ_1^s according to \mathcal{X} (Fig. 2C, blue). For the dimensionalities in which
231 SMC-ABC is tractable, the inferred parameters are concentrated and offset from the exemplary
232 dataset \mathbf{x}_0 (Fig. 2C, green). When using SNPE, the predictions of the inferred parameters are
233 highly concentrated at some RNN sizes and widely varied in others (Fig. 2C, orange). We see these
234 properties reflected in simulations from the inferred distributions: EPI produces a consistent variety
235 of stable, amplified activity norms $|r(t)|$, SMC-ABC produces a limited variety of responses, and the
236 changing variety of responses from SNPE emphasizes the control of EPI on parameter predictions.
237 EPI outperforms SNPE in high dimensions by using gradient information (from $\nabla_{\mathbf{z}}f(\mathbf{x}; \mathbf{z}) =$
238 $\nabla_{\mathbf{z}}[\text{real}(\lambda_1), \lambda_1^s]^{\top}$) on each iteration of optimization. This agrees with recent speculation that such
239 gradient information could improve the efficiency of simulation-based inference techniques [56].
240 Since gradients of the emergent property statistics are necessary in EPI optimization, gradient
241 tractability is a key criteria when determining the suitability of a simulation-based inference tech-
242 nique. Evidenced by this analysis, EPI is a clear choice for inferring high dimensional parameter
243 distributions when the emergent property gradient is efficiently calculated. This can be invaluable
244 for understanding how RNNs produce complex computations. Even with a high degree of biophys-
245 ical realism and expensive emergent property gradients, EPI was run successfully on intermediate
246 hub frequency in a 5-neuron subcircuit model of the STG (Section 3.1). However, conditioning on
247 the pyloric rhythm [57] in a model of the pyloric subnetwork model [15] proved to be prohibitive
248 with EPI. The pyloric subnetwork requires many time steps for simulation and many key emergent
249 property statistics (e.g. burst duration and phase gap) are not calculable or easily approximated
250 with differentiable functions. In such cases, SNPE, which does not require differentiability of the
251 emergent property has proved to be a powerful approach [35]. In the next two sections, we use EPI
252 for novel scientific insight by examining the structure of inferred distributions.

253 **3.4 EPI reveals how recurrence with multiple inhibitory subtypes governs ex-
254 citatory variability in a V1 model**

255 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output
256 function have succeeded in explaining a host of experimentally documented phenomena in primary
257 visual cortex (V1). In a regime characterized by inhibitory stabilization of strong recurrent excita-
258 tion, these models give rise to paradoxical responses [12], selective amplification [41, 50], surround
259 suppression [58] and normalization [59]. Recent theoretical work [60] shows that stabilized E-I

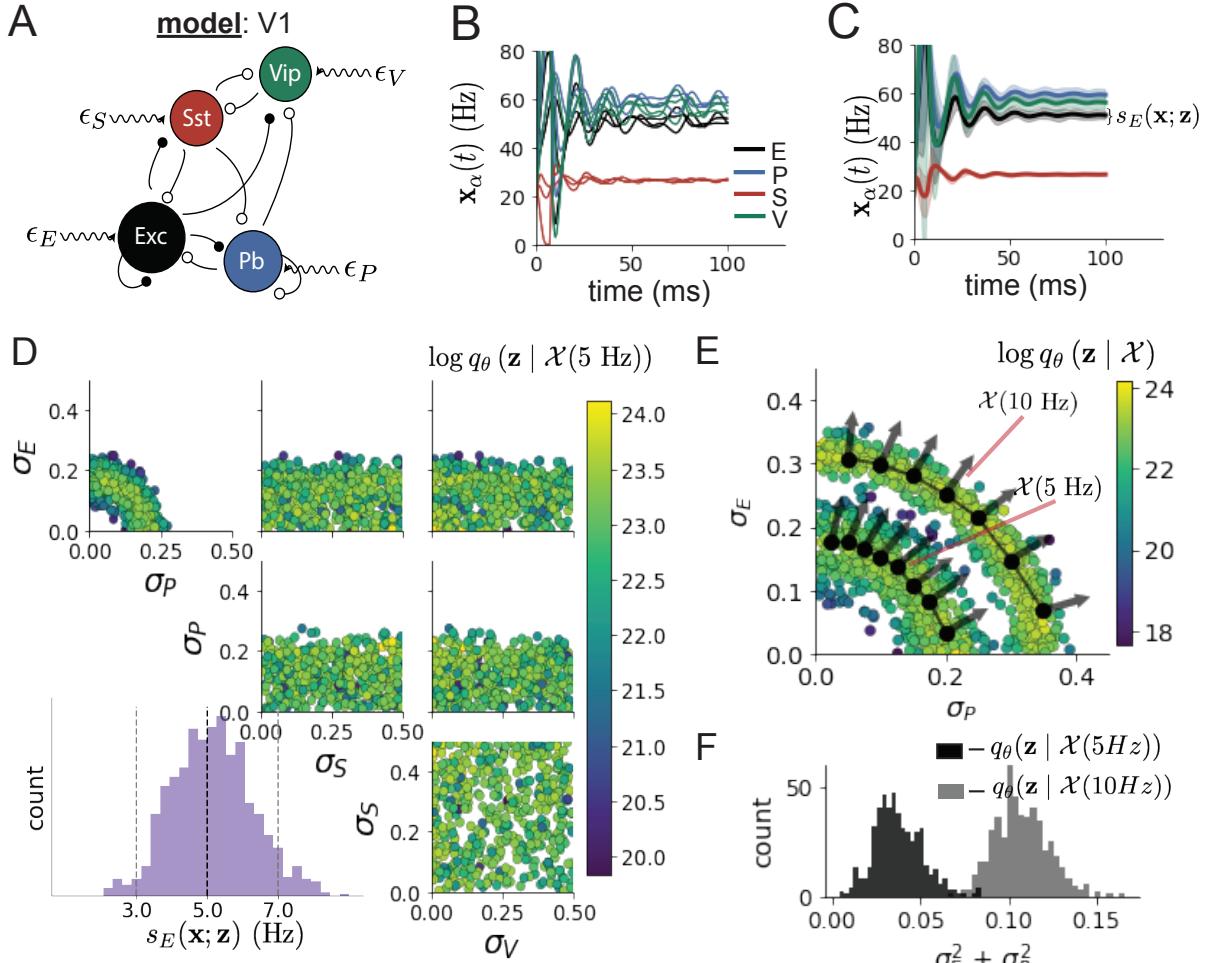


Figure 3: Emergent property inference in the stochastic stabilized supralinear network (SSSN)

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 $q_\theta(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$.

260 models reproduce the effect of variability suppression [61]. However, the way in which E-I recur-
 261 rence structures variability into shared and private dimensions [62] is not understood. Furthermore,
 262 experimental evidence shows that inhibition is composed of distinct elements – parvalbumin (P),
 263 somatostatin (S), VIP (V) – composing 80% of GABAergic interneurons in V1 [63–65], and that
 264 these inhibitory cell types follow specific connectivity patterns (Fig. 3A) [66]. Here, we use EPI
 265 on a model of V1 with biologically realistic connectivity to show how the structure of input across
 266 neuron types affects the variability of the excitatory population – the population largely responsible
 267 for projecting to other brain areas [67].

268 We considered response variability of a nonlinear dynamical V1 circuit model (Fig. 3A) with a state
 269 comprised of each neuron-type population’s rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population receives
 270 recurrent input $W\mathbf{x}$, where W is the effective connectivity matrix (see Section 5.4) and an external
 271 input with mean \mathbf{h} , which determines population rate via supralinear nonlinearity $\phi(\cdot) = [\cdot]_+^2$. The
 272 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
 273 has a slower dynamical timescale $\tau_{\text{noise}} > \tau$ than the population rate, allowing fluctuations around
 274 a stimulus-dependent steady-state (Fig. 3B). This model is the stochastic stabilized supralinear
 275 network (SSSN) [60]

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \epsilon). \quad (7)$$

276 generalized to have multiple inhibitory neuron types, and introduces stochasticity to previous four
 277 neuron-type models of V1 [44]. Stochasticity and inhibitory multiplicity introduce substantial com-
 278 plexity to the mathematical treatment of this problem (see Section 5.4.4) motivating the treatment
 279 of this model with EPI. Here, we consider fixed weights W and input \mathbf{h} [45], and study the effect
 280 of input variability $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on excitatory variability.

281 We quantify levels y of E-population variability with the emergent property

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

282 where $s_E(\mathbf{x}; \mathbf{z})$ is the standard deviation of the stochastic E -population response about its steady
 283 state (Fig. 3C). In the following analyses, we compare levels of 5Hz and 10Hz, and select 1 Hz²
 284 variance such that the two emergent properties do not overlap in $s_E(\mathbf{z}; \mathbf{x})$.

285 First, we ran EPI to obtain parameter distribution $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ producing E-population vari-
 286 ability around 5 Hz (Fig. 3D). From the marginal distribution of σ_E and σ_P (Fig. 3D, top-left), we
 287 can see that $s_E(\mathbf{x}; \mathbf{z})$ is sensitive to various combinations of σ_E and σ_P . Alternatively, both σ_S and

288 σ_V are degenerate with respect to $s_E(\mathbf{x}; \mathbf{z})$ evidenced by the high variability in those dimensions
289 (Fig. 3D, bottom-right). Together, these observations imply a curved path with respect to $s_E(\mathbf{x}; \mathbf{z})$
290 of 5 Hz, which is indicated by the modes along σ_P (Fig. 3E).

291 Figure 3E suggests a quadratic relationship in E-population fluctuations and the standard deviation
292 of E- and P-population input; as the square of either σ_E or σ_P increases, the other compensatorily
293 decreases to preserve the level of $s_E(\mathbf{x}; \mathbf{z})$. This quadratic relationship is preserved at greater level
294 of E-population variability $\mathcal{X}(10 \text{ Hz})$ (Fig. 3E). Indeed, the sum of squares of σ_E and σ_P is larger
295 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
296 σ_S and σ_V are not significantly different in the two EPI distributions (Fig. 15, $p = .40$), in which
297 parameters were bounded from 0 to 0.5. The strong interactive influence of E- and P-population
298 input variability on excitatory variability is intriguing, since this circuit exhibits a paradoxical effect
299 in the P-population (and no other inhibitory types) (Fig. 15), meaning that the E-population is
300 P-stabilized. Future research may uncover a link between the population of network stabilization
301 and compensatory interactions governing excitatory variability.

302 EPI revealed the quadratic dependence of excitatory variability on input variability to the E- and
303 P-populations, as well as its independence to input from the other two inhibitory populations. In
304 a simplified model ($\tau = \tau_{\text{noise}}$), it can be shown that surfaces of equal variance are ellipsoids as
305 a function of σ . Nevertheless, the sensitive and degenerate parameters are challenging to predict
306 mathematically, since the covariance matrix depends on the steady-state solution of the network
307 [60,68], and terms in the covariance expression increase quadratically with each additional neuron-
308 type population (see also Section 5.4.4). This emphasizes the value of streamlined methods for
309 gaining understanding about theoretical models when mathematical analysis becomes onerous or
310 impractical. While we have just shown that EPI can be used to investigate fundamental aspects
311 of sensory computation, in the next two sections, we use the probabilistic tools of EPI to identify
312 and characterize two distinct parametric regimes of a neural circuit executing a computation, and
313 then relate these insights to behavioral experiments.

314 3.5 EPI identifies two regimes of rapid task switching

315 It has been shown that rats can learn to switch from one behavioral task to the next on randomly
316 interleaved trials [69], and an important question is what types of neural connectivity produce this
317 computation. In this experimental setup, rats were explicitly cued on each trial to either orient
318 towards a visual stimulus in the Pro (P) task or orient away from the stimulus in the Anti (A)

task (Fig. 4A). Neural recordings in superior colliculus (SC) exhibited two populations of neurons that represented task context (Pro or Anti). Furthermore, Pro/Anti neurons in each hemisphere were strongly correlated with the animal’s decision [46]. These results motivated a model of SC that is a four-population dynamical system with functionally-defined neuron-types. Here, our goal is to understand how connectivity in this circuit model governs the ability to perform rapid task switching: to respond with satisfactory accuracy in both tasks on randomly interleaved trials.

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

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

The model responds to the side with greater Pro neuron activation; e.g. the response is left if $x_{LP} > x_{RP}$ at the end of the trial. Here, we use EPI to determine the network connectivity $\mathbf{z} = [sW, vW, dW, hW]^\top$ that produces rapid task switching.

Rapid task switching is formalized mathematically as an emergent property with two statistics: 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 .75 in each task with variance .075²

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .75 \\ .75 \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} .075^2 \\ .075^2 \end{bmatrix}. \end{aligned} \tag{10}$$

75% accuracy is a realistic level of performance in each task, and with the chosen variance, inferred models will not exhibit fully random responses (50%), nor perfect performance (100%).

The EPI inferred distribution (Fig. 4C) produces Pro and Anti task accuracies (Fig. 4C, middle-left) consistent with rapid task switching (Equation 10). This parameter distribution has rich structure, that is not captured well by simple linear correlations (Fig. 17). Specifically, the shape of the EPI distribution is sharply bent, matching ground truth structure indicated by brute-force

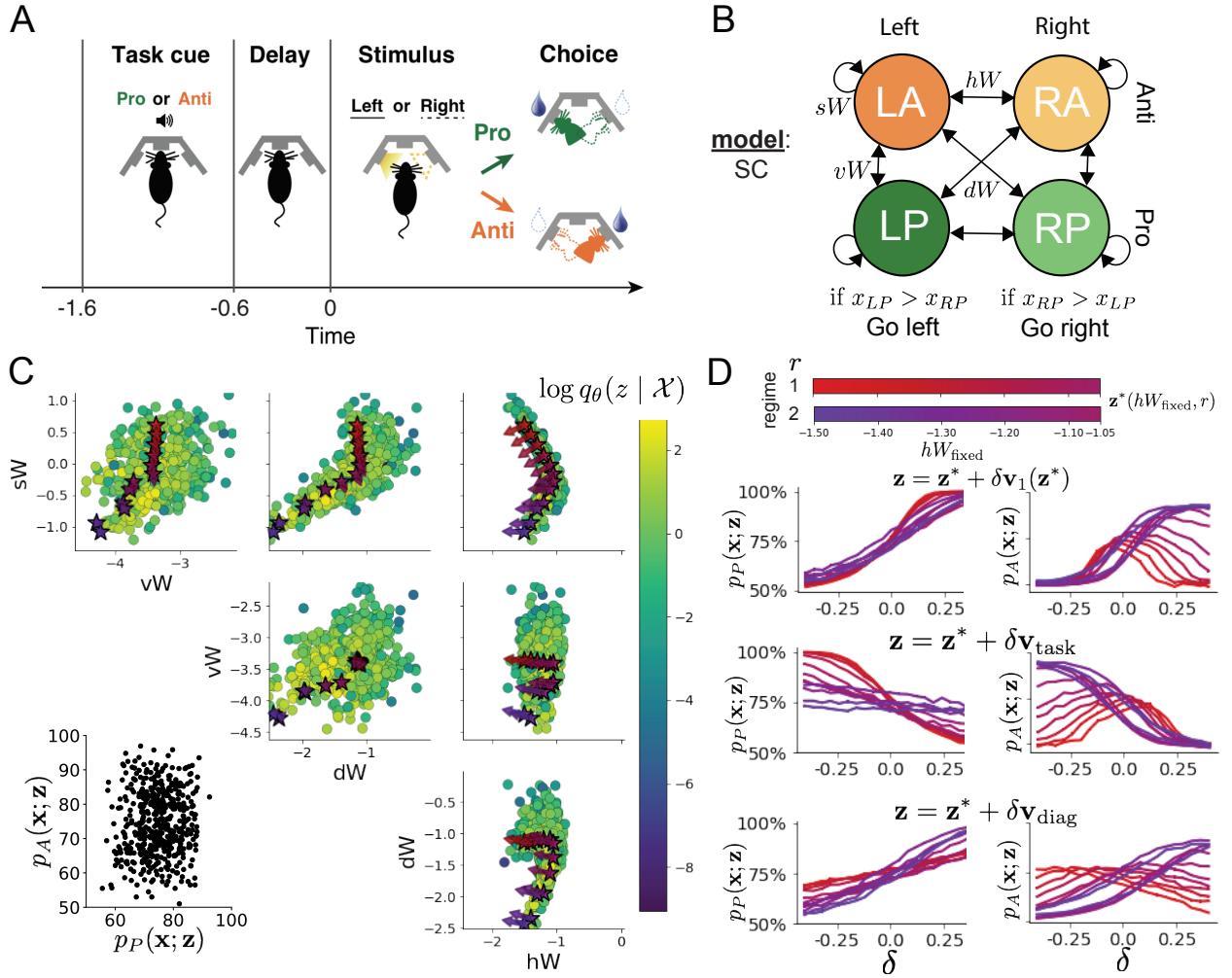


Figure 4: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - Left Pro, RP - Right Pro, LA - Left Anti, RA - Right Anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. **C.** The EPI inferred distribution of rapid task switching networks. Stars indicate modes \mathbf{z}^* whose color indicates connectivity regime (see legend Fig 4D). Sensitivity vectors $\mathbf{v}_1(\mathbf{z}^*)$ are shown by arrows. (Bottom-left) EPI predictive distribution of task accuracies. **D.** The connectivity regimes have different responses to perturbation. (Top) Mean and standard error ($N_{\text{test}} = 25$) of accuracy with respect to perturbation along the sensitivity dimension of each mode \mathbf{z}^* . (Middle) Same with perturbation in the dimension of increasing λ_{task} (\mathbf{v}_{task}). (Bottom) Same with perturbation in the dimension of increasing λ_{diag} (\mathbf{v}_{diag}).

344 sampling (Fig. 24) This is most saliently pointed out in the marginal distribution of $sW-hW$ (Fig.
345 4C top-right), where anticorrelation between sW and hW switches to correlation with decreasing
346 sW . The two regimes produce different types of responses in the Pro and Anti tasks (Fig. SC2).
347 Not only has EPI captured this complicated distribution of connectivities producing rapid task
348 switching, we can query the EPI distribution $q_{\theta}(\mathbf{z} | \mathcal{X})$ to understand these two parametric regimes
349 of SC connectivity.

350 To distinguish these two regimes, we use the EPI distribution to identify two sets of modes. By
351 fixing hW to different values and doing gradient ascent on $\log q_{\theta}(\mathbf{z} | \mathcal{X})$, we arrive at two solutions
352 $\mathbf{z}^*(hW_{\text{fixed}}, r)$ where regime $r \in [1, 2]$, and regime 1 is that of greater sW (see Section 5.5.4). As
353 hW_{fixed} increases, the modes coalesce to intermediate parameters reflecting a transition between
354 the two sets of modes (Fig. 20 top). By using EPI to connect these two regimes through this
355 transitional region of parameter space, we can explore what distinguishes the two regimes by
356 stepping from the prototypical connectivity of regime 1 to that of regime 2.

357 Although the connectivities gradually coalesce to the transitional part of parameter space, the
358 sensitivity dimensions $\mathbf{v}_1(\mathbf{z})$ are categorically different across regimes (Fig. 20 bottom). The
359 sensitivity dimension identifies the parameter combination which causes the emergent property to
360 diminish with the shortest perturbation. Since the two regimes have different $\mathbf{v}_1(\mathbf{z})$, this suggests
361 they have different pathologies in their connectivity. By perturbing connectivity in each regime
362 along the sensitivity dimension, we can get a sense of the differing nature of these pathologies.

363 When perturbing connectivity along the sensitivity dimension, Pro accuracy monotonically in-
364 creases in both regimes (Fig. 4D, top-left). However, there is a stark difference between regimes in
365 Anti accuracy. Anti accuracy falls in either direction of \mathbf{v}_1 in regime 1, yet monotonically increases
366 along with Pro accuracy in regime 2 (Fig. 4D, top-right). These distinct pathologies of rapid task
367 switching are caused by distinct connectivity changes ($\mathbf{v}_1(\mathbf{z}^*(r=1))$ vs $\mathbf{v}_1(\mathbf{z}^*(r=2))$) and explain
368 the sharp change in local structure of the EPI distribution.

369 To further examine the two regimes, we can perturb \mathbf{z} in the same way along dimensions that inde-
370 pendently change the eigenvalues of the connectivity matrix (which has constant eigenvectors with
371 respect to \mathbf{z}). These eigenvalues λ_{all} , λ_{side} , λ_{task} , and λ_{diag} correspond to connectivity eigenmodes
372 with intuitive roles in processing in this task (Fig. 19A). For example, greater λ_{task} will strengthen
373 internal representations of task, while greater λ_{diag} will amplify dominance of Pro and Anti pairs in
374 opposite hemispheres (Section 5.5.6). Perturbation analyses reveal that decreasing λ_{task} has a very
375 similar effect on Anti accuracy as perturbations along the sensitivity dimension (Fig. 4D, middle).

376 This suggests that there is a carefully tuned strength of task representation in connectivity regime
377 1, which if disturbed results in random Anti trial responses. Finally, we recognize that increasing
378 λ_{diag} has opposite effects on Anti accuracy in each regime (Fig. 4D, bottom). In the next section,
379 we build on these mechanistic characterizations of each regime by examining their resilience to
380 optogenetic silencing.

381 **3.6 EPI inferred SC connectivities reproduce results from optogenetic inacti-
382 vation experiments**

383 During the delay period of this task, the circuit must prepare to execute the correct task according
384 to the presented cue. Experimental results from Duan et al. found that bilateral optogenetic
385 inactivation of SC during the delay period consistently decreased performance in the Anti task, but
386 had no effect on the Pro task (Fig. 5A). This suggests that SC maintains a representation of task
387 throughout the delay period, which is important for correct execution of the Anti task. Network
388 connectivities inferred by EPI exhibited this same effect in simulation at high optogenetic strengths
389 γ (Fig. 5B) (see Section 5.5.7).

390 The mean increase in Anti error is closest to the experimentally measured value of 7% at $\gamma = 0.675$
391 (Fig. 5B, black dot). At this level of optogenetic strength, only regime 1 exhibits an increase in
392 Anti error with delay period silencing (Fig. 5C, left). The connectivities in regime 2 are thus more
393 resilient to delay period silencing during Anti trials than regime 1. In regime 1, greater λ_{task} and
394 λ_{diag} decrease Anti error (Fig. 5C, right). In other words, these anticorrelations show that stronger
395 task representations and diagonal amplification make the SC model more resilient to delay period
396 silencing in the Anti task. All correlations of connectivity eigenvalue with Anti error degrade in
397 regime 2, where there is no effect of delay period silencing on Anti error (Fig. 5C, right).

398 At about $\gamma = 0.85$ (Fig. 5B, gray dot), the Anti error saturates, while Pro error remains at zero
399 Following delay period inactivation at this optogenetic strength, there are strong similarities in
400 the responses of Pro and Anti trials during the choice period (Fig. 5D, left). We interpreted
401 these similarities to suggest that delay period inactivation at this saturated level flips the internal
402 representation of task (from Anti to Pro) in the circuit model. This would explain why the Anti
403 error saturates at 50%: the average Anti accuracy in EPI inferred connectivities is 75%, but is 25%
404 when the internal representation is flipped during delay period silencing. This hypothesis prescribes
405 a model of Anti accuracy during delay period silencing of $p_{A,\text{opto}} = 100\% - p_P$, which is fit closely
406 across both regimes of the EPI inferred connectivities (Fig. 5D, right). Similarities between Pro

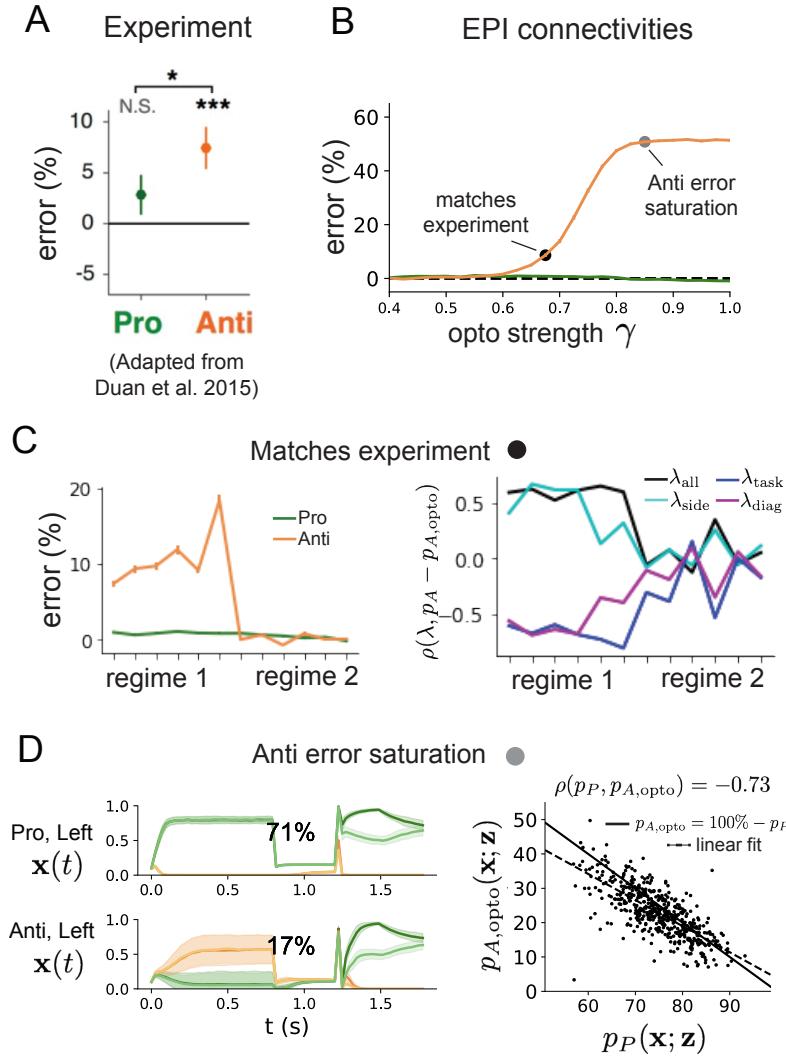


Figure 5: **A.** Experimental effect of delay period silencing on Pro and Anti task accuracy in rats. **B.** Mean and standard error (bars) of task error induced by delay period inactivation of varying optogenetic strength across the EPI distribution. **C.** (Left) Mean and standard error of Pro and 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 $\gamma = 0.85$. (Right) Anti accuracy following delay period inactivation at $\gamma = 0.85$ versus accuracy in the Pro task across connectivities in the EPI distribution.

407 and Anti trial responses were not present at the experiment-matching level of $\gamma = 0.675$ (Fig. 22
408 left) and neither was anti-correlation in p_P and $p_{A,\text{opto}}$ (Fig. 22 right).

409 In summary, the connectivity inferred by EPI to perform rapid task switching replicated results
410 from optogenetic silencing experiments. We found that at levels of optogenetic strength matching
411 experimental levels of Anti error, only one regime actually exhibited the effect. This suggests that
412 one regime is less resilient to optogenetic perturbation, and perhaps more biologically realistic.
413 Finally, we mechanistically characterized the pathology in Anti error that occurs in both regimes
414 when optogenetic strength is increased to high levels. The probabilistic tools afforded by EPI
415 yielded this insight: we identified two regimes and the continuum of connectivities between them
416 by taking gradients of parameter probabilities in the EPI distribution, we identified sensitivity
417 dimensions by measuring the Hessian of the EPI distribution, and we obtained many parameter
418 samples at each step along the continuum (in 7.36 seconds with the EPI distribution rather than
419 4.2 days with brute force methods, see Section 5.5).

420 4 Discussion

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

431 Methodology for statistical inference in circuit models has evolved considerably in recent years.
432 Early work used rejection sampling techniques [24–26], but more recently developed methodology
433 employs deep learning to improve efficiency or provide deep, flexible distribution approximations.
434 SNPE [35], developed along with EPI (see Section 5.1.1), has been used for posterior inference of
435 parameters in circuit models conditioned upon exemplar data used to represent computation. Like
436 SNPE, EPI is a deep inference technique, but it infers parameter distributions that only produce

437 the computation of interest (see Section 3.3).

438 Exemplary data versus emergent properties aside, EPI has better scaling properties than SNPE
439 when emergent property gradients are tractable (Section 3.3). However, SNPE has its own relative
440 advantages. SNPE is effective when circuit model simulations are lengthy or nondifferentiable. For
441 example, SNPE can infer the STG parameters that produce the pyloric rhythm [35], while EPI
442 cannot. Thus, while it is nice to infer parameter distributions with constrained emergent properties
443 with EPI, SNPE is most appropriate when emergent property gradients are intractable.

444 The scientific analyses of Sections 3.4 and 3.5 derived theoretical findings by querying the structure
445 of the inferred distribution of EPI. By measuring the dimensions in which probability decreases
446 fastest, the dimensions of *sensitivity*, we gain valuable understanding of how model parameters
447 govern the emergent property. A rich literature on parameter sensitivity analyses in biological
448 models presents several methods towards this scientific approach [27, 28, 30, 70]. The value offered
449 by EPI (and other deep inference methods like SNPE), is that the once the flexible deep probability
450 distributions are fit to the parameter distribution, this distribution’s structure can be quantified at
451 any parameter choice, offering instantly available sensitivity measurements. Together, the ability
452 to condition upon emergent properties, the efficient inference algorithm, and the capacity for pa-
453 rameter sensitivity analyses make EPI a powerful new method for addressing inverse problems in
454 theoretical neuroscience.

455 **Acknowledgements:**

456 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
457 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
458 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
459 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
460 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James
461 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

462 **Data availability statement:**

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

465 **Code availability statement:**

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

467 **References**

- 468 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
469 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 470 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
471 21(1):25–45, 1998.
- 472 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 473 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in cog-
474 nition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 475 [5] Timothy O’Leary, Alexander C Sutton, and Eve Marder. Computational models in the age of
476 large datasets. *Current opinion in neurobiology*, 32:87–94, 2015.
- 477 [6] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
478 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
479 models. *PLoS Comput Biol*, 3(10):e189, 2007.
- 480 [7] Kamil Erguler and Michael PH Stumpf. Practical limits for reverse engineering of dynamical
481 systems: a statistical analysis of sensitivity and parameter inferability in systems biology
482 models. *Molecular BioSystems*, 7(5):1593–1602, 2011.
- 483 [8] Brian K Mannakee, Aaron P Ragsdale, Mark K Transtrum, and Ryan N Gutenkunst. Sloppi-
484 ness and the geometry of parameter space. In *Uncertainty in Biology*, pages 271–299. Springer,
485 2016.
- 486 [9] John J Hopfield. Neural networks and physical systems with emergent collective computational
487 abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 488 [10] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
489 networks. *Physical review letters*, 61(3):259, 1988.
- 490 [11] Andrey V Olypher and Ronald L Calabrese. Using constraints on neuronal activity to reveal
491 compensatory changes in neuronal parameters. *Journal of Neurophysiology*, 98(6):3749–3758,
492 2007.

- 493 [12] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
494 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
495 17(11):4382–4388, 1997.
- 496 [13] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in
497 perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 498 [14] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
499 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 500 [15] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
501 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 502 [16] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
503 model. *PLoS computational biology*, 2(7):e94, 2006.
- 504 [17] Dmitry Fisher, Itsaso Olasagasti, David W Tank, Emre RF Aksay, and Mark S Goldman.
505 A modeling framework for deriving the structural and functional architecture of a short-term
506 memory microcircuit. *Neuron*, 79(5):987–1000, 2013.
- 507 [18] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
508 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
509 sion model. *Neuron*, 82(4):809–821, 2014.
- 510 [19] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
511 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 512 [20] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
513 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
514 2018.
- 515 [21] Christopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state
516 in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.
- 517 [22] Aman B Saleem, Asli Ayaz, Kathryn J Jeffery, Kenneth D Harris, and Matteo Carandini.
518 Integration of visual motion and locomotion in mouse visual cortex. *Nature neuroscience*,
519 16(12):1864–1869, 2013.

- 520 [23] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Church-
521 land. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuro-*
522 *science*, 22(10):1677–1686, 2019.
- 523 [24] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation
524 in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 525 [25] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
526 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
527 2003.
- 528 [26] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
529 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 530 [27] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
531 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
532 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
533 1929, 2009.
- 534 [28] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
535 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
536 2012.
- 537 [29] Keegan E Hines, Thomas R Middendorf, and Richard W Aldrich. Determination of parameter
538 identifiability in nonlinear biophysical models: A bayesian approach. *Journal of General
539 Physiology*, 143(3):401–416, 2014.
- 540 [30] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
541 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 542 [31] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
543 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 544 [32] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
545 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 546 [33] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical
547 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.

- 548 [34] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
549 free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–
550 5533, 2017.
- 551 [35] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
552 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P
553 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
554 dynamics. *bioRxiv*, page 838383, 2019.
- 555 [36] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
556 networks. *International Conference on Learning Representations*, 2017.
- 557 [37] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
558 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 559 [38] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
560 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 561 [39] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
562 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
563 *Neuron*, 77(5):845–858, 2013.
- 564 [40] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, robustness,
565 and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238, 2001.
- 566 [41] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
567 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 568 [42] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dy-
569 namics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–
570 1406, 2014.
- 571 [43] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding
572 and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.
- 573 [44] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and vi-
574 sual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,
575 115(3):1399–1409, 2016.

- 576 [45] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
577 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
578 operating regime of cortex. *bioRxiv*, 2020.
- 579 [46] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
580 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
581 routing. *bioRxiv*, page 245613, 2018.
- 582 [47] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
583 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 584 [48] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
585 *International Conference on Machine Learning*, 2015.
- 586 [49] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density
587 estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.
- 588 [50] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
589 2009.
- 590 [51] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural
591 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 592 [52] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in*
593 *neurobiology*, 25:156–163, 2014.
- 594 [53] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*
595 *opinion in neurobiology*, 46:1–6, 2017.
- 596 [54] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H
597 Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex
598 embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966,
599 2018.
- 600 [55] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian compu-*
601 *tation*. CRC Press, 2018.
- 602 [56] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference.
603 *Proceedings of the National Academy of Sciences*, 2020.

- 604 [57] Eve Marder and Allen I Selverston. *Dynamic biological networks: the stomatogastric nervous*
605 *system.* MIT press, 1992.
- 606 [58] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
607 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
608 592, 2009.
- 609 [59] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
610 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,
611 85(2):402–417, 2015.
- 612 [60] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
613 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
614 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 615 [61] Mark M. Churchland, Byron M. Yu, John P. Cunningham, Leo P. Sugrue, Marlene R. Cohen,
616 Greg S. Corrado, William T. Newsome, Andrew M. Clark, Paymon Hosseini, Benjamin B.
617 Scott, David C. Bradley, Matthew A. Smith, Adam Kohn, J. Anthony Movshon, Katherine
618 M. Armstrong, Tirin Moore, Steve W. Chang, Lawrence H. Snyder, Stephen G. Lisberger,
619 Nicholas J. Priebe, Ian M. Finn, David Ferster, Stephen I. Ryu, Gopal Santhanam, Maneesh
620 Sahani, and Krishna V. Shenoy. Stimulus onset quenches neural variability: a widespread
621 cortical phenomenon. *Nat. Neurosci.*, 13(3):369–378, 2010.
- 622 [62] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn.
623 Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.
- 624 [63] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
625 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
626 5(10):793, 2004.
- 627 [64] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
628 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neuro-
629 biology*, 71(1):45–61, 2011.
- 630 [65] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex:
631 From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.

- 632 [66] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
633 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
634 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 635 [67] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate
636 cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
- 637 [68] C Gardiner. Stochastic methods: A Handbook for the Natural and Social Sciences, 2009.
- 638 [69] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and midbrain
639 regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503, 2015.
- 640 [70] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
641 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 642 [71] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
643 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 644 [72] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
645 Edward Teller. Equation of state calculations by fast computing machines. *The journal of
646 chemical physics*, 21(6):1087–1092, 1953.
- 647 [73] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
648 1970.
- 649 [74] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
650 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
651 73(2):123–214, 2011.
- 652 [75] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-
653 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
654 820, 2011.
- 655 [76] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
656 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience
657 Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 658 [77] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
659 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with
660 degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.

- 661 [78] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
662 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
663 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 664 [79] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical inference for
665 analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin, Germany*,
666 2019.
- 667 [80] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
668 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
669 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 670 [81] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
671 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on
672 Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 673 [82] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
674 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
675 4248. PMLR, 2020.
- 676 [83] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
677 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 678 [84] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
679 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 680 [85] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
681 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
682 2018.
- 683 [86] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
684 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 685 [87] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji
686 Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint
687 arXiv:1912.02762*, 2019.

- 688 [88] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
689 Improved variational inference with inverse autoregressive flow. *Advances in neural information*
690 *processing systems*, 29:4743–4751, 2016.
- 691 [89] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International*
692 *Conference on Learning Representations*, 2015.
- 693 [90] Emmanuel Klinger, Dennis Rickert, and Jan Hasenauer. pyabc: distributed, likelihood-free
694 inference. *Bioinformatics*, 34(20):3591–3593, 2018.
- 695 [91] David S Greenberg, Marcel Nonnenmacher, and Jakob H Macke. Automatic posterior trans-
696 formation for likelihood-free inference. *International Conference on Machine Learning*, 2019.
- 697 [92] Daniel P Mossing, Julia Veit, Agostina Palmigiano, Kenneth D. Miller, and Hillel Adesnik.
698 Antagonistic inhibitory subnetworks control cooperation and competition across cortical space.
699 *bioRxiv*, 2021.

700 **5 Methods**

701 **5.1 Emergent property inference (EPI)**

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

711 Consider model parameterization \mathbf{z} , which is a collection of scientifically meaningful variables that
712 govern the complex simulation of data \mathbf{x} . For example (see Section 3.1), \mathbf{z} may be the electrical
713 conductance parameters of an STG subcircuit, and \mathbf{x} the evolving membrane potentials (the state)
714 of the five neurons. In terms of statistical modeling, this circuit model has an intractable likelihood
715 $p(\mathbf{x} | \mathbf{z})$, which is predicated by the stochastic differential equations that define the model. From a
716 theoretical perspective, we are less concerned about the likelihood of an exemplary dataset \mathbf{x} , but
717 rather the emergent property of intermediate hub frequency (which implies a consistent dataset \mathbf{x}).

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

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (11)$$

723 Precisely, the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ must have means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ over the EPI
724 distribution of parameters and the data produced by those parameters. Technically, an emergent
725 property may be a combination of first-, second-, or higher-order moments, but this study focuses
726 on the case written in Equation 11.

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

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

$$\mathbf{z} = g_\theta(\mathbf{z}_0) = g_l(\dots g_1(\mathbf{z}_0)) \sim q_\theta(\mathbf{z}). \quad (12)$$

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

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

$$\begin{aligned} q_\theta(\mathbf{z} | \mathcal{X}) &= q_\theta^*(\mathbf{z}) = \operatorname{argmax}_{q_\theta \in \mathcal{Q}} H(q_\theta(\mathbf{z})) \\ \text{s.t. } \mathcal{X} &: \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \end{aligned} \quad (13)$$

738 Entropy is chosen as the normative selection principle to match that of variational Bayesian methods
 739 (see Section 5.1.3). However, a key difference is that variational Bayesian methods do not constrain
 740 the predictions of their inferred parameter distribution. This optimization is executed using the
 741 algorithm of Maximum Entropy Flow Networks (MEFNs) [36].

742 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.
 743 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-
 744 ond, we describe the special class of deep probability distributions used in EPI called normalizing
 745 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-
 746 tion 13 (Section 5.1.4). Then, we demonstrate the details of this optimization in a toy example
 747 (Section 5.1.5). Finally, we establish the known relationship between maximum entropy distri-
 748 butions and exponential families (Section 5.1.3), which is used to explain how EPI is a form of
 749 variational inference (Section 5.1.6).

750 5.1.1 Related approaches

751 When bayesian inference problems lack conjugacy, scientists use approximate inference methods like
 752 variational inference (VI) [71] and Markov chain Monte Carlo (MCMC) [72, 73]. After optimization,
 753 variational methods return a parameterized posterior distribution, which we can analyze. Also, the
 754 variational approximating distribution class is often chosen such that it permits fast sampling. In

755 contrast MCMC methods only produce samples from the approximated posterior distribution. No
756 parameterized distribution is estimated, and additional samples are always generated with the same
757 sampling complexity. Inference in models defined by systems of differential has been demonstrated
758 with MCMC [74], although this approach requires tractable likelihoods. Advancements have lever-
759 aged structure in stochastic differential equation models to improve likelihood approximations, thus
760 expanding the domain of applicable models [75].

761 Simulation-based inference [56] is model parameter inference in the absence of a tractable likeli-
762 hood function. The most prevalent approach to simulation-based inference is approximate Bayesian
763 computation (ABC) [24], in which satisfactory parameter samples are kept from random prior sam-
764 pling according to a rejection heuristic. The obtained set of parameters do not have a probabilities,
765 and further insight about the model must be gained from examination of the parameter set and
766 their generated activity. Methodological advances to ABC methods have come through the use of
767 Markov chain Monte Carlo (MCMC-ABC) [25] and sequential Monte Carlo (SMC-ABC) [26] sam-
768 pling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still struggles
769 to scale in dimensionality (cf. Fig. 2). Furthermore, once a parameter set has been obtained by
770 SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again from scratch
771 with a new population of initialized particles to obtain additional samples.

772 For scientific model analysis, we seek a parameter distribution represented by an approximating
773 distribution as in variational inference [71] : a variational approximation that once optimized
774 yields fast analytic calculations and samples. For the reasons described above, ABC and MCMC
775 techniques are unattractive, since they only produce a set of parameter samples lacking probabilities
776 and have unchanging sampling rate. EPI infers parameters in circuit models using the MEFN [36]
777 algorithm with a deep variational approximation. The deep neural network of EPI (Fig. 1E)
778 defines the parametric form (with variational parameters θ) of the deep variational approximation
779 of circuit parameters \mathbf{z} .

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

785 We note that, during our preparation and early presentation of this work [76, 77], another work
786 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural

787 circuits ([35, 78, 79]). We are encouraged by this general problem being recognized by others in the
788 community, and we emphasize that these works offer complementary neuroscientific contributions
789 (different theoretical models of focus) and use different technical methodologies (ours is built on
790 our prior work [36], theirs similarly [80]).

791 The method EPI differs from SNPE in some key ways. SNPE belongs to a “sequential” class
792 of recently developed simulation-based inference methods in which two neural networks are used
793 for posterior inference. This first neural network is a deep probability distribution (normalizing
794 flow) used to estimate the posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE) or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural
795 likelihood (SNL [81])). A recent advance uses an unconstrained neural network to estimate the
796 likelihood ratio (sequential neural ratio estimation (SNRE [82])). In SNL and SNRE, MCMC
797 sampling techniques are used to obtain samples from the approximated posterior. This contrasts
798 with EPI and SNPE, which use deep probability distributions to model parameters, which facilitates
799 immediate measurements of sample probability, gradient, or Hessian for system analysis. The
800 second neural network in this sequential class of methods is the amortizer. This unconstrained
801 deep network maps data \mathbf{x} (or statistics $f(\mathbf{x}; \mathbf{z})$) or model parameters \mathbf{z} to the weights and biases of
802 the first neural network. These methods are optimized on a conditional density (or ratio) estimation
803 objective. The data used to optimize this objective are generated via an adaptive procedure, in
804 which training data pairs $(\mathbf{x}_i, \mathbf{z}_i)$ become sequentially closer to the true data and posterior.

805 The approximating fidelity of the deep probability distribution in sequential approaches is opti-
806 mized to generalize across the training distribution of the conditioning variable. This generalization
807 property of the sequential methods can reduce the accuracy at the singular posterior of interest.
808 Whereas in EPI, the entire expressivity of the deep probability distribution is dedicated to learning
809 a single distribution as well as possible. Amortization is not possible in EPI, since EPI learns
810 an exponential family distribution parameterized by its mean (see Section 5.1.3). Since EPI dis-
811 tributions are defined by the mean μ of their statistics, there is the well-known inverse mapping
812 problem of exponential families [83] that prohibits an amortization based approach. However, we
813 have shown that the same two-network architecture of the sequential simulation-based inference
814 methods can be used for amortized inference in intractable exponential family posteriors using their
815 natural parameterization [84].

816 Finally, one important differentiating factor between EPI and sequential simulation-based infer-
817 ence methods is that EPI leverages gradients $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ during optimization. These gradients can
818 improve convergence time and scalability, as we have shown on an example conditioning low-rank

819 RNN connectivity on the property of stable amplification (see Section 3.3). With EPI, we prove
 820 out the suggestion that a deep inference technique can improve efficiency by leveraging these model
 821 gradients when they are tractable. Sequential simulation-based inference techniques may be better
 822 suited for scientific problems where $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ is intractable or unavailable, like when there is a non-
 823 differentiable model or it requires lengthy simulations. However, the sequential simulation-based
 824 inference techniques cannot constrain the predictions of the inferred distribution in the manner of
 825 EPI.

826 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in
 827 scientific models. Around a single parameter choice, one can measure the Jacobian. One approach
 828 for this calculation that scales well is EAR [28]. A popular efficient approach for systems of ODEs
 829 has been neural ODE adjoint [85] and its stochastic adaptation [86]. Casting identifiability as a
 830 statistical estimation problem, the profile likelihood works via iterated optimization while holding
 831 parameters fixed [27]. An exciting recent method is capable of recovering the functional form of such
 832 unidentifiabilities away from a point by following degenerate dimensions of the fisher information
 833 matrix [30]. Global structural non-identifiabilities can be found for models with polynomial or
 834 rational dynamics equations using DAISY [70]. With EPI, we have all the benefits given by a
 835 statistical inference method plus the ability to query the first- or second-order gradient of the
 836 probability of the inferred distribution at any chosen parameter value. The second-order gradient
 837 of the log probability (the Hessian), which is directly afforded by EPI distributions, produces
 838 quantified information about parametric sensitivity of the emergent property in parameter space
 839 (see Section 3.2).

840 **5.1.2 Deep probability distributions and normalizing flows**

841 Deep probability distributions are comprised of multiple layers of fully connected neural networks
 842 (Equation 12). When each neural network layer is restricted to be a bijective function, the sample
 843 density can be calculated using the change of variables formula at each layer of the network. For
 844 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

$$p(\mathbf{z}_i) = p(g_i^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial g_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} \right| = p(\mathbf{z}_{i-1}) \left| \det \frac{\partial g_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right|^{-1}. \quad (14)$$

845 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 846 restricting our layers to normalizing flows [48, 87] – bijective functions with fast log determinant
 847 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log

848 probability calculation confers efficient optimization of the maximum entropy objective (see Section
849 5.1.4).

850 We use the Real NVP [37] normalizing flow class, because its coupling architecture confers both
851 fast sampling (forward) and fast log probability evaluation (backward). Fast probability evaluation
852 facilitates fast gradient and Hessian evaluation of log probability throughout parameter space.
853 Glow permutations were used in between coupling stages [38]. This is in contrast to autoregressive
854 architectures [49, 88], in which only one of the forward or backward passes can be efficient. In this
855 work, normalizing flows are used as flexible parameter distribution approximations $q_{\theta}(\mathbf{z})$ having
856 weights and biases θ . We specify the architecture used in each application by the number of Real-
857 NVP affine coupling stages, and the number of neural network layers and units per layer of the
858 conditioning functions.

859 When calculating Hessians of log probabilities in deep probability distributions, it is important to
860 consider the normalizing flow architecture. With autoregressive architectures [49, 88], fast sam-
861 pling and fast log probability evaluations are mutually exclusive. That makes these architectures
862 undesirable for EPI, where efficient sampling is important for optimization, and log probability
863 evaluation speed predicates the efficiency of gradient and Hessian calculations. With Real NVP
864 coupling architectures, we get both fast sampling and fast Hessians making both optimization and
865 scientific analysis efficient.

866 5.1.3 Maximum entropy distributions and exponential families

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

$$p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \quad (15)$$

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

869 will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \quad (16)$$

870 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
871 formally hard to identify except in special cases [83].

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

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

$$\text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \mathbb{E}_{\mathbf{z}, \mathbf{x}} \left[(f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \right] \quad (17)$$

874 As a general maximum entropy distribution (Equation 15), the sufficient statistics vector contains
 875 both first and second order moments of $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (18)$$

876 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (19)$$

877 5.1.4 Augmented Lagrangian optimization

878 To optimize $q_{\boldsymbol{\theta}}(\mathbf{z})$ in Equation 13, the constrained maximum entropy optimization is executed using
 879 the augmented Lagrangian method. The following objective is minimized:

$$L(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{opt}}, c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top R(\boldsymbol{\theta}) + \frac{c}{2} \|R(\boldsymbol{\theta})\|^2 \quad (20)$$

880 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
 881 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
 882 sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter $\boldsymbol{\mu}_{\text{opt}}$ are determined by the means $\boldsymbol{\mu}$ and variances
 883 σ^2 of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 13 (see Section 5.1.6). Specifically,
 884 $T(\mathbf{x}; \mathbf{z})$ is a concatenation of the first and second moments, $\boldsymbol{\mu}_{\text{opt}}$ is a concatenation of $\boldsymbol{\mu}$ and σ^2
 885 (see section 5.1.3), and the Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of
 886 exponential families (see Section 5.1.3). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution
 887 are optimized according to Equation 20 using the Adam optimizer with learning rate 10^{-3} [89].

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

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

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

$$\nabla_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\nabla_{\boldsymbol{\theta}} \log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (22)$$

893 The lagrangian parameters η_{opt} are initialized to zero and adapted following each augmented
894 Lagrangian epoch, which is a period of optimization with fixed (η_{opt}, c) for a given number of
895 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
896 after each epoch based on constraint error reduction. The penalty coefficient is updated based
897 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
898 $\mathbb{E}[|R(\theta_{k+1})|] > \gamma \mathbb{E}[|R(\theta_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
899 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
900 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
901 according to application.

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

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

915 To assess whether the EPI distribution $q_{\theta}(\mathbf{z})$ produces the emergent property, we assess whether
916 each individual constraint on the means and variances of $f(\mathbf{x}; \mathbf{z})$ is satisfied. We consider the EPI
917 to have converged when a null hypothesis test of constraint violations $R(\theta)_i$ being zero is accepted
918 for all constraints $i \in \{1, \dots, m\}$ at a significance threshold $\alpha = 0.05$. This significance threshold is
919 adjusted through Bonferroni correction according to the number of constraints m . The p-values for
920 each constraint are calculated according to a two-tailed nonparametric test, where 200 estimations
921 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
922 Lagrangian epoch.

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

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

935 5.1.5 Example: 2D LDS

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

$$\tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (23)$$

938 with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (24)$$

939 To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fixing
 940 $\tau = 1s$), the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ were chosen to contain the oscillatory frequency,
 941 $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscillating system. λ_1 is the eigenvalue of
 942 greatest real part when the imaginary component is zero, and alternatively of positive imaginary
 943 component when the eigenvalues are complex conjugate pairs. To learn the distribution of real
 944 entries of A that produce a band of oscillating systems around 1Hz, we formalized this emergent
 945 property as $\text{real}(\lambda_1)$ having mean zero with variance 0.25^2 , and the oscillation frequency $2\pi\text{imag}(\lambda_1)$
 946 having mean 1Hz with variance $(0.1\text{Hz})^2$:

$$\mathbb{E}[T(\mathbf{x})]_{\mathbf{z}, \mathbf{x}} \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \text{imag}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0)^2 \\ (\text{imag}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}_{\text{opt.}} \quad (25)$$

947

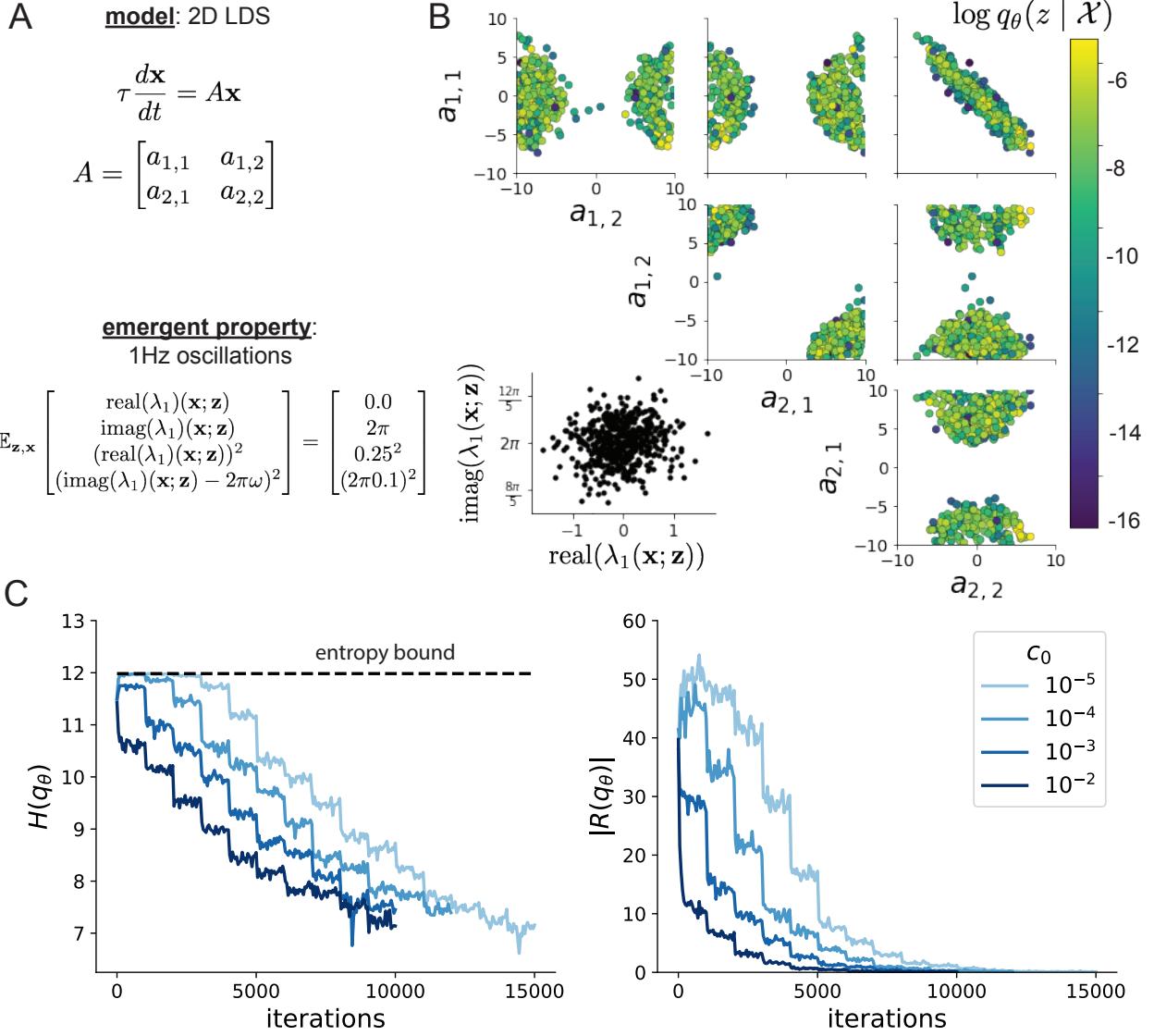


Figure 6: (LDS1): **A.** Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. **B.** The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed lines indicate the parameter axes. **C.** Entropy throughout the optimization. At the beginning of each augmented Lagrangian epoch (2,000 iterations), the entropy dipped due to the shifted optimization manifold where emergent property constraint satisfaction is increasingly weighted. **D.** Emergent property moments throughout optimization. At the beginning of each augmented Lagrangian epoch, the emergent property moments adjust closer to their constraints.

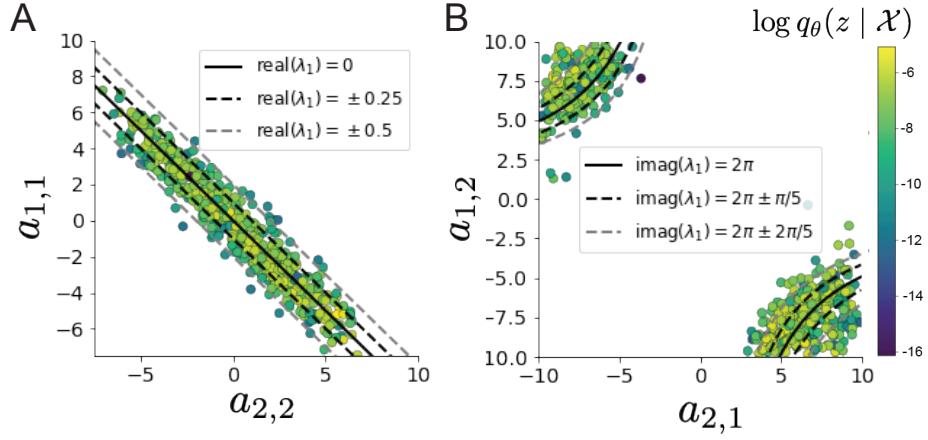


Figure 7: (LDS2): **A.** Probability contours in the a_1 - a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. **B.** Probability contours in the a_2 - a_3 plane were derived from the emergent property statistic of oscillation frequency $2\pi\text{imag}(\lambda_1)$.

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

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (26)$$

951 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
 952 does not require simulation, we cannot derive the distribution q_θ^* directly. This fact is due to the
 953 formally hard problem of the backward mapping: finding the natural parameters η from the mean
 954 parameters μ of an exponential family distribution [83]. Instead, we used EPI to approximate this
 955 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two
 956 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped
 957 onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

958 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the
 959 parameter distribution. To validate our method, we analytically derived the contours of the prob-
 960 ability density from the emergent property statistics and values. In the a_1 - a_4 plane, the black
 961 line at $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$,
 962 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
 963 of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired
 964 statistical constraints and model degeneracy in the sum of a_1 and a_4 . Intuitively, the parameters
 965 equivalent with respect to emergent property statistic $\text{real}(\lambda_1)$ have similar log densities.

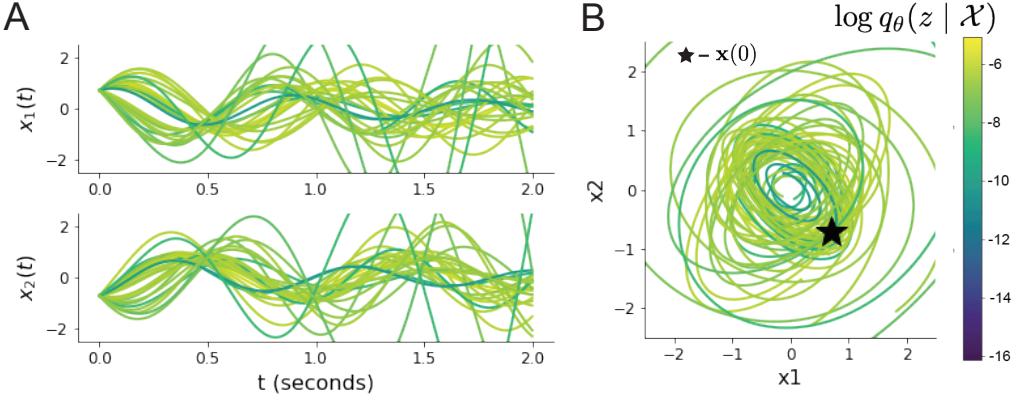


Figure 8: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\theta}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. **A.** Each dimension of the simulated trajectories throughout time. **B.** The simulated trajectories in phase space.

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1 a_4 - a_2 a_3}{\tau}}, & \text{if } a_1 a_4 < a_2 a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (27)$$

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

$$\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 \quad (28)$$

970 Since we constrained $\mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\text{imag}(\lambda)] = 2\pi$, we can plot contours of the equation $\text{imag}(\lambda_1)^2 =$
971 $a_1 a_4 - a_2 a_3 = (2\pi)^2$ for various $a_1 a_4$ (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [|a_1 a_4 - E_{q_{\theta}}[a_1 a_4]|]$, we show
972 the contours as $a_1 a_4 = 0$ (black), $a_1 a_4 = -\sigma_{1,4}$ (black dotted), and $a_1 a_4 = -2\sigma_{1,4}$ (grey dotted).
973 This validates the curved structure of the inferred distribution learned through EPI. We took steps
974 in negative standard deviation of $a_1 a_4$ (dotted and gray lines), since there are few positive values
975 $a_1 a_4$ in the learned distribution. Subtler combinations of model and emergent property will have
976 more complexity, further motivating the use of EPI for understanding these systems. As we expect,
977 the distribution results in samples of two-dimensional linear systems oscillating near 1Hz (Fig. S3).

978 **5.1.6 EPI as variational inference**

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

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \quad (29)$$

983 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
 984 However, conjugacy is rare in practice, and alternative methods, such as variational inference [71],
 985 are utilized.

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

$$q_{\theta}^*(\mathbf{z}) = \operatorname{argmin}_{q_{\theta} \in \mathcal{Q}} KL(q_{\theta}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})). \quad (30)$$

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

$$KL(q_{\theta}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})) = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(q_{\theta}(\mathbf{z}))] - \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(p(\mathbf{z} | \mathbf{x}))] \quad (31)$$

$$= -H(q_{\theta}) - \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z})) - \log(p(\mathbf{x}))] \quad (32)$$

990 Since the marginal distribution of the data $p(\mathbf{x})$ (or “evidence”) is independent of θ , variational
 991 inference is executed by optimizing the remaining expression. This is usually framed as maximizing
 992 the evidence lower bound (ELBO)

$$\operatorname{argmin}_{q_{\theta} \in \mathcal{Q}} KL(q_{\theta} || p(\mathbf{z} | \mathbf{x})) = \operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}) + \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z}))]. \quad (33)$$

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

$$p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) | \boldsymbol{\mu}_f, \Sigma_f), \quad (34)$$

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

$$\log p(\mathbf{x} | \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \quad (35)$$

997 where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\boldsymbol{\mu}_f}{\boldsymbol{\sigma}_f^2} \\ \frac{-1}{2\boldsymbol{\sigma}_f^2} \end{bmatrix}, \text{ and} \quad (36)$$

998

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \quad (37)$$

999 The variational objective is then

$$\underset{q_{\theta} \in Q}{\operatorname{argmax}} H(q_{\theta}) + \boldsymbol{\eta}_f^{\top} \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [T(\mathbf{x}; \mathbf{z})] \quad (38)$$

1000 Comparing this to the Lagrangian objective (without augmentation) of EPI, we see they are the
1001 same

$$\begin{aligned} q_{\theta}^*(\mathbf{z}) &= \underset{q_{\theta} \in Q}{\operatorname{argmin}} -H(q_{\theta}) + \boldsymbol{\eta}_{\text{opt}}^{\top} (\mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}_{\text{opt}}) \\ &= \underset{q_{\theta} \in Q}{\operatorname{argmin}} -H(q_{\theta}) + \boldsymbol{\eta}_{\text{opt}}^{\top} \mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})]. \end{aligned} \quad (39)$$

1002 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
1003 (Equation 18). Thus, EPI is implicitly executing variational inference with a uniform prior and a
1004 mean-field gaussian likelihood on the emergent property statistics. The mean and variances of the
1005 mean-field gaussian likelihood are predicated by $\boldsymbol{\eta}_{\text{opt}}$ (Equations 36 and 38), which is adapted after
1006 each EPI optimization epoch based on \mathcal{X} (see Section 5.1.4). In EPI, the inferred distribution is
1007 not conditioned on a finite dataset as in variational inference, but rather the emergent property
1008 \mathcal{X} dictates the likelihood parameterization such that the inferred distribution will produce the
1009 emergent property. As a note, we could not simply choose $\boldsymbol{\mu}_f$ and $\boldsymbol{\sigma}_f$ directly from the outset, since
1010 we do not know which of these choices will produce the emergent property \mathcal{X} , which necessitates
1011 the EPI optimization routine that adapts $\boldsymbol{\eta}_{\text{opt}}$. Accordingly, we replace the notation of $p(\mathbf{z} \mid \mathbf{x})$
1012 with $p(\mathbf{z} \mid \mathcal{X})$ conceptualizing an inferred distribution that obeys emergent property \mathcal{X} (see Section
1013 5.1).1014

5.2 Stomatogastric ganglion

1015 In Section 3.1 and 3.2, we used EPI to infer conductance parameters in a model of the stomatogastric
1016 ganglion (STG) [39]. This 5-neuron circuit model represents two subcircuits: that generating the
1017 pyloric rhythm (fast population) and that generating the gastric mill rhythm (slow population).
1018 The additional neuron (the IC neuron of the STG) receives inhibitory synaptic input from both
1019 subcircuits, and can couple to either rhythm dependent on modulatory conditions. There is also
1020 a parametric regime in which this neuron fires at an intermediate frequency between that of the
1021 fast and slow populations [39], which we infer with EPI as a motivational example. This model

1022 is not to be confused with an STG subcircuit model of the pyloric rhythm [57], which has been
 1023 statistically inferred in other studies [15, 35].

1024 **5.2.1 STG model**

1025 We analyze how the parameters $\mathbf{z} = [g_{el}, g_{synA}]$ govern the emergent phenomena of intermediate
 1026 hub frequency in a model of the stomatogastric ganglion (STG) [39] shown in Figure 1A with
 1027 activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$, using the same hyperparameter choices as Gutierrez et al.
 1028 Each neuron's membrane potential $x_\alpha(t)$ for $\alpha \in \{f1, f2, \text{hub}, s1, s2\}$ is the solution of the following
 1029 stochastic differential equation:

$$C_m \frac{dx_\alpha}{dt} = -[h_{leak}(\mathbf{x}; \mathbf{z}) + h_{Ca}(\mathbf{x}; \mathbf{z}) + h_K(\mathbf{x}; \mathbf{z}) + h_{hyp}(\mathbf{x}; \mathbf{z}) + h_{elec}(\mathbf{x}; \mathbf{z}) + h_{syn}(\mathbf{x}; \mathbf{z})] + dB. \quad (40)$$

1030 The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization,
 1031 electrical and synaptic currents. Each current component is a function of all membrane potentials
 1032 and the conductance parameters \mathbf{z} . Finally, we include gaussian noise dB to the model of Gutierrez
 1033 et al. so that the model stochastic, although this is not required by EPI.

1034 The capacitance of the cell membrane was set to $C_m = 1nF$. Specifically, the currents are the
 1035 difference in the neuron's membrane potential and that current type's reversal potential multiplied
 1036 by a conductance:

$$h_{leak}(\mathbf{x}; \mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (41)$$

$$h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{post} - x_\alpha^{pre}) \quad (42)$$

$$h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (43)$$

$$h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (44)$$

$$h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (45)$$

$$h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (46)$$

1042 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 1043 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 1044 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 1045 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
 1046 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
 1047 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

1048 Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating
 1049 dynamics dependent on steady-state gating variables M_∞ , N_∞ and H_∞ , respectively:

$$M_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_1}{v_2} \right) \right) \quad (47)$$

$$\frac{dN}{dt} = \lambda_N (N_\infty - N) \quad (48)$$

$$N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_3}{v_4} \right) \right) \quad (49)$$

$$\lambda_N = \phi_N \cosh \left(\frac{x_\alpha - v_3}{2v_4} \right) \quad (50)$$

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (51)$$

$$H_\infty = \frac{1}{1 + \exp \left(\frac{x_\alpha + v_5}{v_6} \right)} \quad (52)$$

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp \left(\frac{-x_\alpha + v_7}{v_8} \right)} \right). \quad (53)$$

1056 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 1057 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

1058 Finally, there is a synaptic gating variable as well:

$$S_\infty = \frac{1}{1 + \exp \left(\frac{v_{th} - x_\alpha}{v_9} \right)}. \quad (54)$$

1059 When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear
 1060 dynamical system. The gaussian noise $d\mathbf{B}$ has variance $(1 \times 10^{-12})^2$ A², and introduces variability
 1061 in frequency at each parameterization \mathbf{z} .

1062 5.2.2 Hub frequency calculation

1063 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for
 1064 $T = 300$ time steps of $dt = 25\text{ms}$. The chosen dt and T were the most computationally convenient
 1065 choices yielding accurate frequency measurement. We used a basis of complex exponentials with
 1066 frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

$$\Phi = [0.0, 0.01, \dots, 1.0]^\top .. \quad (55)$$

1067 To measure spiking frequency, we processed simulated membrane potentials with a relu (spike
 1068 extraction) and low-pass filter with averaging window of size 20, then took the frequency with the

1069 maximum absolute value of the complex exponential basis coefficients of the processed time-series.
 1070 The first 20 temporal samples of the simulation are ignored to account for initial transients.
 1071 To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in$
 1072 $\mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in$
 1073 $\{f1, f2, \text{hub}, s1, s2\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta |X_\alpha| \odot i), \quad (56)$$

1074 where $i = [0, 1, \dots, 100]$. The frequency is then calculated as

$$\omega_\alpha = 0.01\psi_\alpha \text{Hz}. \quad (57)$$

1075 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
 1076 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
 1077 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be $(0.025\text{Hz})^2$
 1078 (Equation 4).

1079 **5.2.3 EPI details for the STG model**

1080 As a maximum entropy distribution, $T(\mathbf{x}; \mathbf{z})$ is comprised of both these first and second moments
 1081 of the hub neuron frequency (as in Equations 18 and 19)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \quad (58)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \quad (59)$$

1082 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 1083 epoch of 5,000 iterations(see Section 5.1.4). The optimization converged after five epochs (Fig. S4).

1085 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-
 1086 layer neural networks of 25 units per layer. The normalizing flow architecture mapped $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, I)$
 1087 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
 1088 returned by a preliminary ABC search. We did not include $g_{\text{synA}} < 0.01$, for numerical stability.
 1089 EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an
 1090 augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The architecture
 1091 converged with criteria $N_{\text{test}} = 100$.

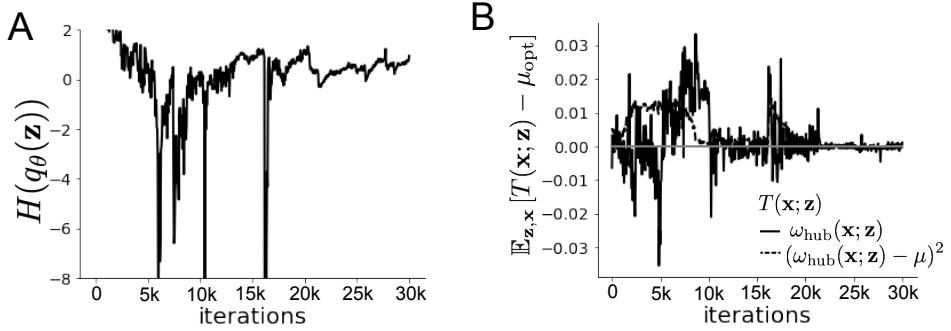


Figure 9: (STG1): EPI optimization of the STG model producing network syncing. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

1092 5.2.4 Hessian sensitivity vectors

1093 To quantify the second-order structure of the EPI distribution, we evaluated the Hessian of the log
 1094 probability $\frac{\partial^2 \log q(\mathbf{z}|\mathcal{X})}{\partial \mathbf{z} \mathbf{z}^\top}$. The eigenvector of this Hessian with most negative eigenvalue is defined as
 1095 the sensitivity dimension \mathbf{v}_1 , and all subsequent eigenvectors are ordered by increasing eigenvalue.
 1096 These eigenvalues are quantifications of how fast the emergent property deteriorates via the param-
 1097 eter combination of their associated eigenvector. In Figure 1D, the sensitivity dimension v_1 (solid)
 1098 and the second eigenvector of the Hessian v_2 (dashed) are shown evaluated at the mode of the
 1099 distribution. Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D
 1100 parameter space were chosen to have positive g_{synA} . The length of the arrows is inversely propor-
 1101 tional to the square root of the absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$. For
 1102 the same magnitude perturbation away from the mode, intermediate hub frequency only diminishes
 1103 along the sensitivity dimension \mathbf{v}_1 (Fig. 1E-F).

1104 5.3 Scaling EPI for stable amplification in RNNs

1105 5.3.1 Rank-2 RNN model

1106 We examined the scaling properties of EPI by learning connectivities of RNNs of increasing size
 1107 that exhibit stable amplification. Rank-2 RNN connectivity was modeled as $W = UV^\top$, where
 1108 $U = [\mathbf{U}_1 \ \mathbf{U}_2] + g\chi^{(W)}$, $V = [\mathbf{V}_1 \ \mathbf{V}_2] + g\chi^{(V)}$, and $\chi_{i,j}^{(W)}, \chi_{i,j}^{(V)} \sim \mathcal{N}(0, 1)$. This RNN model has

1109 dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}. \quad (60)$$

1110 In this analysis, we inferred connectivity parameterizations $\mathbf{z} = [\mathbf{U}_1^\top, \mathbf{U}_2^\top, \mathbf{V}_1^\top, \mathbf{V}_2^\top]^\top \in [-1, 1]^{(4N)}$
1111 that produced stable amplification using EPI, SMC-ABC [26], and SNPE [35] (see Section Related
1112 Methods).

1113 **5.3.2 Stable amplification**

1114 For this RNN model to be stable, all real eigenvalues of W must be less than 1: $\text{real}(\lambda_1) < 1$,
1115 where λ_1 denotes the greatest real eigenvalue of W . For a stable RNN to amplify at least one input
1116 pattern, the symmetric connectivity $W^s = \frac{W+W^\top}{2}$ must have an eigenvalue greater than 1: $\lambda_1^s > 1$,
1117 where λ^s is the maximum eigenvalue of W^s . These two conditions are necessary and sufficient for
1118 stable amplification in RNNs [51].

1119 **5.3.3 EPI details for RNNs**

1120 We defined the emergent property of stable amplification with means of these eigenvalues (0.5
1121 and 1.5, respectively) that satisfy these conditions. To complete the emergent property definition,
1122 we chose variances (0.25^2) about those means such that samples rarely violate the eigenvalue
1123 constraints. In terms of the EPI optimization variables, this is written as

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \lambda_1^s(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0.5)^2 \\ (\lambda_1^s(\mathbf{x}; \mathbf{z}) - 1.5)^2 \end{bmatrix}, \quad (61)$$

1124

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.25^2 \\ 0.25^2 \end{bmatrix}. \quad (62)$$

1125 Gradients of maximum eigenvalues of Hermitian matrices like W^s are available with modern auto-
1126 matic differentiation tools. To differentiate through the $\text{real}(\lambda_1)$, we solved the following equation
1127 for eigenvalues of rank-2 matrices using the rank reduced matrix $W^r = V^\top U$

$$\lambda_{\pm} = \frac{\text{Tr}(W^r) \pm \sqrt{\text{Tr}(W^r)^2 - 4\text{Det}(W^r)}}{2}. \quad (63)$$

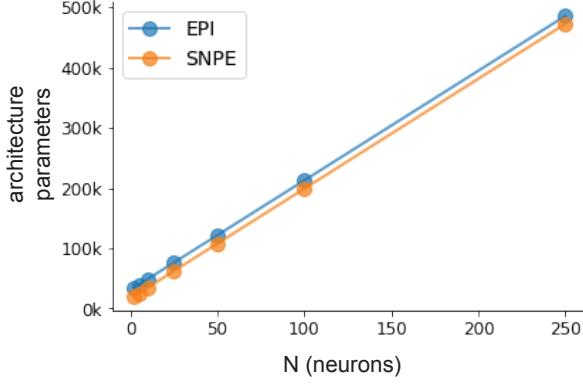


Figure 10: (RNN1): Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size (N).

1128 For EPI in Fig. 2, we used a real NVP architecture with three coupling layers of affine transformations parameterized by two-layer neural networks of 100 units per layer. The initial distribution
 1129 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
 1130 an augmented Lagrangian coefficient of $c_0 = 10^3$, a batch size $n = 200$, $\beta = 4$, and chose to use
 1131 500 iterations per augmented Lagrangian epoch and emergent property constraint convergence was
 1132 evaluated at $N_{\text{test}} = 200$ (Fig. 2B blue line, and Fig. 2C-D blue).
 1133

1134 5.3.4 Methodological comparison

1135 We compared EPI to two alternative simulation-based inference techniques, since the likelihood
 1136 of these eigenvalues given \mathbf{z} is not available. Approximate Bayesian computation (ABC) [24] is a
 1137 rejection sampling technique for obtaining sets of parameters \mathbf{z} that produce activity \mathbf{x} close to some
 1138 observed data \mathbf{x}_0 . Sequential Monte Carlo approximate Bayesian computation (SMC-ABC) is the
 1139 state-of-the-art ABC method, which leverages SMC techniques to improve sampling speed. We ran
 1140 SMC-ABC with the pyABC package [90] to infer RNNs with stable amplification: connectivities
 1141 having eigenvalues within an ϵ -defined l_2 distance of

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}. \quad (64)$$

1142 SMC-ABC was run with a uniform prior over $\mathbf{z} \in [-1, 1]^{(4N)}$, a population size of 1,000 particles
 1143 with simulations parallelized over 32 cores, and a multivariate normal transition model.

1144 SNPE, the next approach in our comparison, is far more similar to EPI. Like EPI, SNPE treats pa-

1145 rameters in mechanistic models with deep probability distributions, yet the two learning algorithms
1146 are categorically different. SNPE uses a two-network architecture to approximate the posterior dis-
1147 tribution of the model conditioned on observed data \mathbf{x}_0 . The amortizing network maps observations
1148 \mathbf{x}_i to the parameters of the deep probability distribution. The weights and biases of the parameter
1149 network are optimized by sequentially augmenting the training data with additional pairs $(\mathbf{z}_i, \mathbf{x}_i)$
1150 based on the most recent posterior approximation. This sequential procedure is important to get
1151 training data \mathbf{z}_i to be closer to the true posterior, and \mathbf{x}_i to be closer to the observed data. For
1152 the deep probability distribution architecture, we chose a masked autoregressive flow with affine
1153 couplings (the default choice), three transforms, 50 hidden units, and a normalizing flow mapping
1154 to the support as in EPI. This architectural choice closely tracked the size of the architecture used
1155 by EPI (Fig. 10). As in SMC-ABC, we ran SNPE with $\mathbf{x}_0 = \mu$. All SNPE optimizations were
1156 run for a limit of 1.5 days on a Tesla V100 GPU, or until two consecutive rounds resulted in a
1157 validation log probability lower than the maximum observed for that random seed.

1158 To clarify the difference in objectives of EPI and SNPE, we show their results on RNN models
1159 with different numbers of neurons N and random strength g . The parameters inferred by EPI
1160 consistently produces the same mean and variance of $\text{real}(\lambda_1)$ and λ_1^s , while those inferred by
1161 SNPE change according to the model definition (Fig. 11A). For $N = 2$ and $g = 0.01$, the SNPE
1162 posterior has greater concentration in eigenvalues around \mathbf{x}_0 than at $g = 0.1$, where the model has
1163 greater randomness (Fig. 11B top, orange). At both levels of g when $N = 2$, the posterior of SNPE
1164 has lower entropy than EPI at convergence (Fig. 11B top). However at $N = 10$, SNPE results in
1165 a predictive distribution of more widely dispersed eigenvalues (Fig. 11A bottom), and an inferred
1166 posterior with greater entropy than EPI (Fig. 11B bottom). We highlight these differences not
1167 to focus on an insightful trend, but to emphasize that these methods optimize different objectives
1168 with different implications.

1169 Note that SNPE converges when it's validation log probability has saturated after several rounds
1170 of optimization (Fig. 11C), and that EPI converges after several epochs of its own optimization
1171 to enforce the emergent property constraints (Fig. 11D blue). Importantly, as SNPE optimizes
1172 its posterior approximation, the predictive means change, and at convergence may be different
1173 than \mathbf{x}_0 (Fig. 11D orange, left). It is sensible to assume that predictions of a well-approximated
1174 SNPE posterior should closely reflect the data on average (especially given a uniform prior and
1175 a low degree of stochasticity), however this is not a given. Furthermore, no aspect of the SNPE
1176 optimization controls the variance of the predictions (Fig. 11D orange, right).

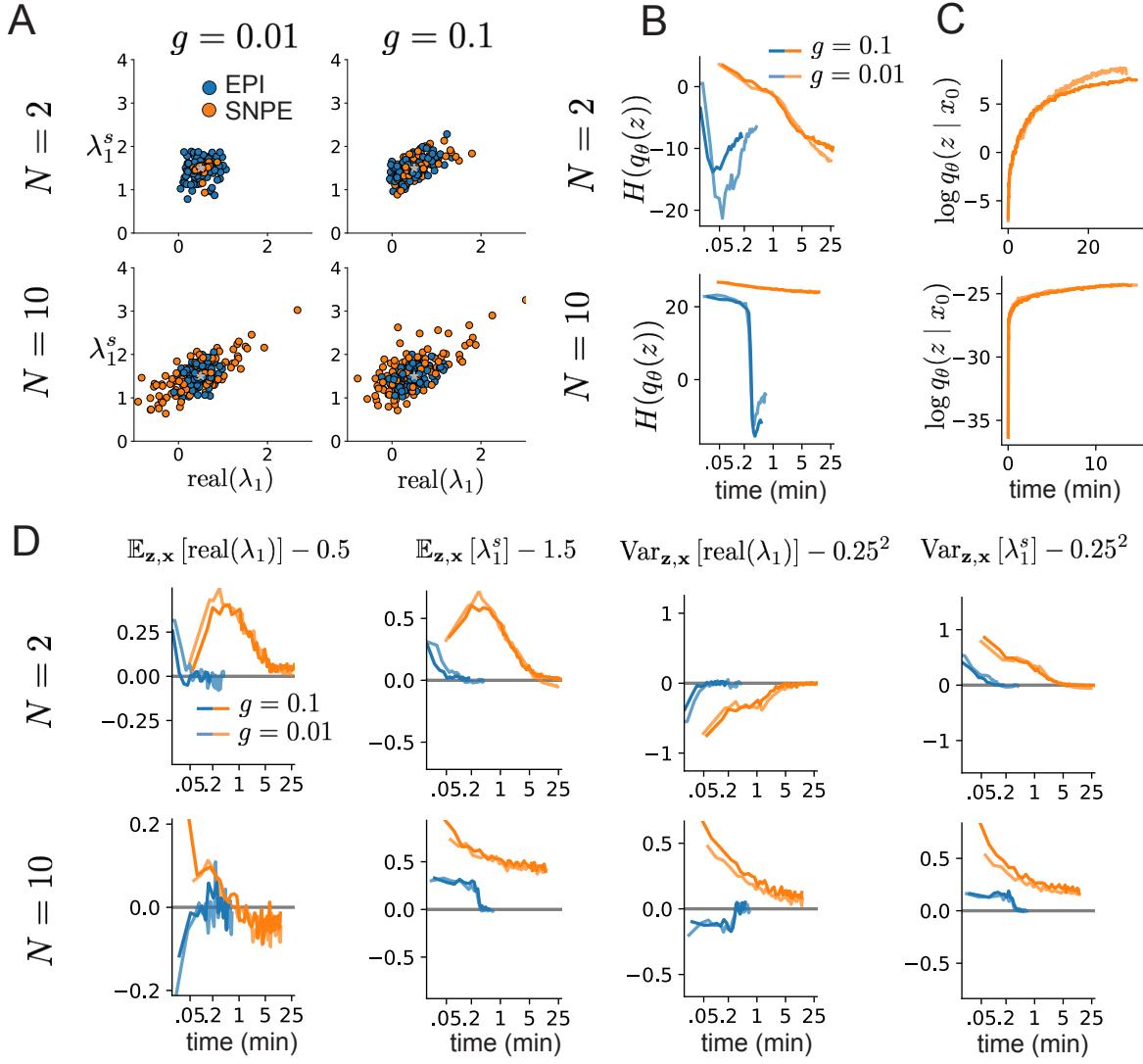


Figure 11: (RNN2): Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A.** Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with $N = 2$ (top), $N = 10$ (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.

1177 To compare the efficiency of these algorithms for inferring RNN connectivity distributions producing
1178 stable amplification, we develop a convergence criteria that can be used across methods. While EPI
1179 has its own hypothesis testing convergence criteria for the emergent property, it would not make
1180 sense to use this criteria on SNPE and SMC-ABC which do not constrain the means and variances
1181 of their predictions. Instead, we consider EPI and SNPE to have converged after completing its
1182 most recent optimization epoch (EPI) or round (SNPE) in which the distance

$$d(q_\theta(z)) = \|\mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}\|_2 \quad (65)$$

1183 is less than 0.5. We consider SMC-ABC to have converged once the population produces samples
1184 within the $\epsilon = 0.5$ ball ensuring stable amplification.

1185 When assessing the scalability of SNPE, it is important to check that alternative hyperparameter-
1186izations could not yield better performance. Key hyperparameters of the SNPE optimization are
1187 the number of simulations per round n_{round} , the number of atoms used in the atomic proposals of
1188 the SNPE-C algorithm [91], and the batch size n . To match EPI, we used a batch size of $n = 200$
1189 for $N \leq 25$, however we found $n = 1,000$ to be helpful for SNPE in higher dimensions. While
1190 $n_{\text{round}} = 1,000$ yielded SNPE convergence for $N \leq 25$, we found that a substantial increase to
1191 $n_{\text{round}} = 25,000$ yielded more consistent convergence at $N = 50$ (Fig. 12A). By increasing n_{round} ,
1192 we also necessarily increase the duration of each round. At $N = 100$, we tried two hyperparameter
1193 modifications. As suggested in [91], we increased n_{atom} by an order of magnitude to improve gra-
1194 dient quality, but this had little effect on the optimization (much overlap between same random
1195 seeds) (Fig. 12B). Finally, we increased n_{round} by an order of magnitude, which yielded convergence
1196 in one case, but no others. We found no way to improve the convergence rate of SNPE without
1197 making more aggressive hyperparameter choices requiring high numbers of simulations.

1198 In Figure 2C-D, we show samples from the random seed resulting in emergent property convergence
1199 at greatest entropy (EPI), the random seed resulting in greatest validation log probability (SNPE),
1200 and the result of all converged random seeds (SMC).

1201 5.4 Primary visual cortex

1202 5.4.1 V1 model

1203 E-I circuit models, rely on the assumption that inhibition can be studied as an indivisible unit,
1204 despite ample experimental evidence showing that inhibition is instead composed of distinct ele-
1205 ments [65]. In particular three types of genetically identified inhibitory cell-types – parvalbumin

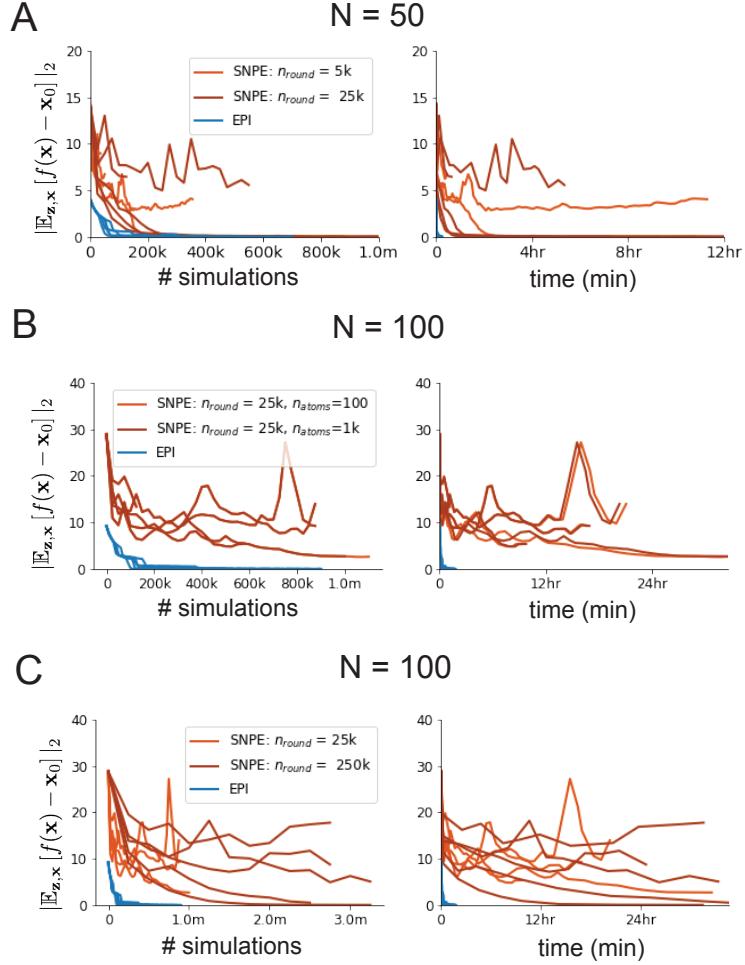


Figure 12: (RNN3): SNPE convergence was enabled by increasing n_{round} , not n_{atom} . **A.** Difference of mean predictions \mathbf{x}_0 throughout optimization at $N = 50$ with by simulation count (left) and wall time (right) of SNPE with $n_{\text{round}} = 5,000$ (light orange), SNPE with $n_{\text{round}} = 25,000$ (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).

1206 (P), somatostatin (S), VIP (V) – compose 80% of GABAergic interneurons in V1 [63–65], and follow
1207 specific connectivity patterns (Fig. 3A) [66], which lead to cell-type specific computations [45, 92].
1208 Currently, how the subdivision of inhibitory cell-types, shapes correlated variability by reconfigur-
1209 ing recurrent network dynamics is not understood.

1210 In the stochastic stabilized supralinear network [60], population rate responses \mathbf{x} to mean input \mathbf{h} ,
1211 recurrent input $W\mathbf{x}$ and slow noise $\boldsymbol{\epsilon}$ are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \quad (66)$$

1212 where the noise is an Ornstein-Uhlenbeck process $\boldsymbol{\epsilon} \sim OU(\tau_{\text{noise}}, \boldsymbol{\sigma})$

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (67)$$

1213 with $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$. The noisy process is parameterized as

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (68)$$

1214 so that $\boldsymbol{\sigma}$ parameterizes the variance of the noisy input in the absence of recurrent connectivity
1215 ($W = \mathbf{0}$). As contrast $c \in [0, 1]$ increases, input to the E- and P-populations increases relative to
1216 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
1217 using the deterministic V1 circuit model [45]

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (69)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (70)$$

1218 and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (71)$$

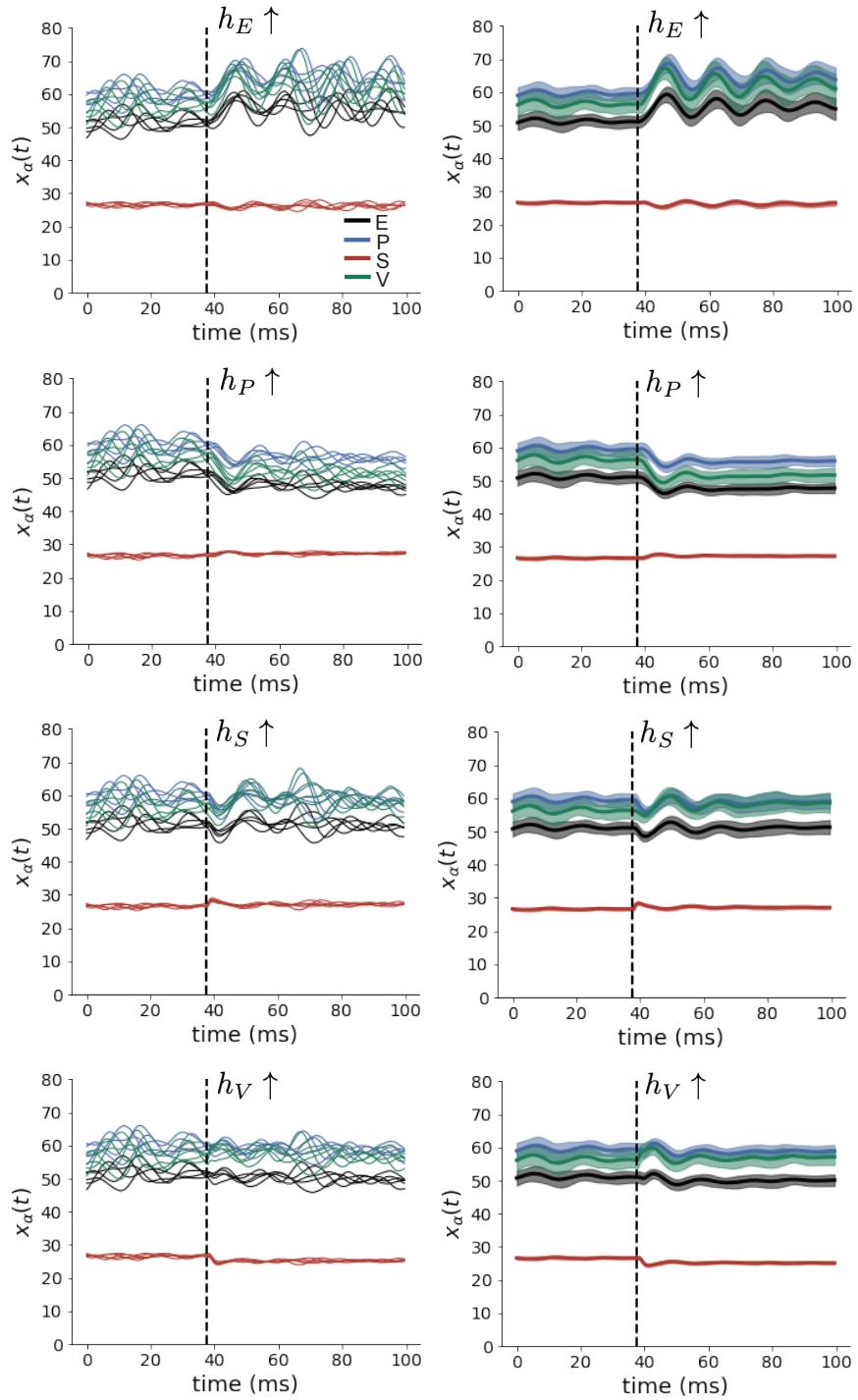


Figure 13: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

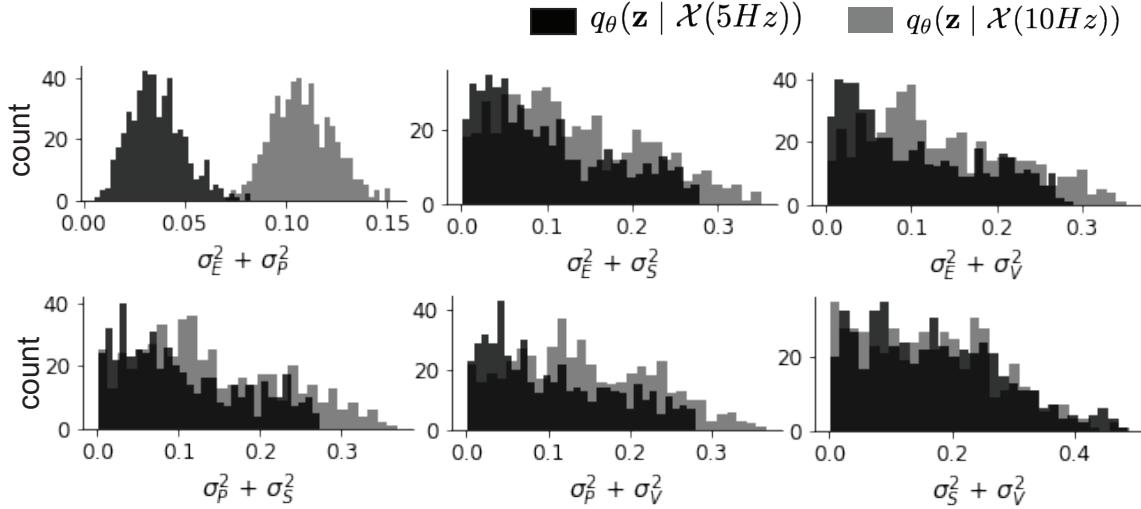


Figure 14: (V1 2) EPI predictive distributions of the sum of squares of each pair of noise parameters.

1219 To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equiv-
1220 alence class

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (72)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (73)$$

1221 and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (74)$$

1222 Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5\text{ms}$ from an initial condition
1223 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

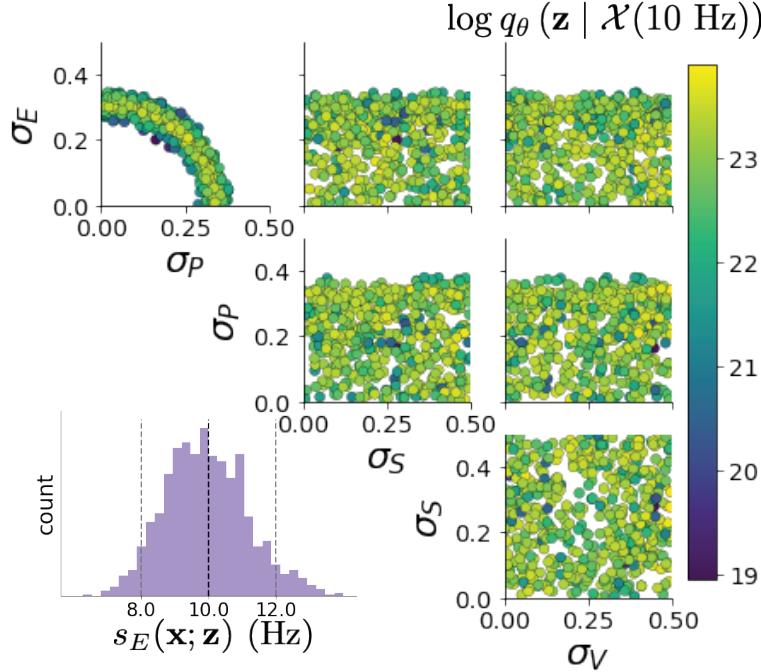


Figure 15: (V1 3) EPI inferred distribution for $\mathcal{X}(10 \text{ Hz})$.

as the square root of the temporal variance from $t_{ss} = 75\text{ms}$ to $Tdt = 100\text{ms}$ averaged over 100 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[\sqrt{\mathbb{E}_{t > t_{ss}} \left[(x_E(t) - \mathbb{E}_{t > t_{ss}} [x_E(t)])^2 \right]} \right] \quad (75)$$

5.4.2 EPI details for the V1 model

For EPI in Fig 3D-E, we used a real NVP architecture with three Real NVP coupling layers and two-layer neural networks of 50 units per layer. The normalizing flow architecture 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 using three different random seeds for architecture initialization $\boldsymbol{\theta}$ with an augmented Lagrangian coefficient of $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the architectures converging with criteria $N_{\text{test}} = 100$ at greatest entropy across three random seeds.

1233 **5.4.3 Sensitivity analyses**

1234 In Fig. 3E, we visualize the modes of $q_{\theta}(\mathbf{z} \mid \mathcal{X})$ throughout the σ_E - σ_P marginal. Specifically, we
 1235 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_{\theta}(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (76)$$

1236 At each mode \mathbf{z}^* , we calculated the Hessian and visualized the sensitivity dimension in the direction
 1237 of positive σ_E .

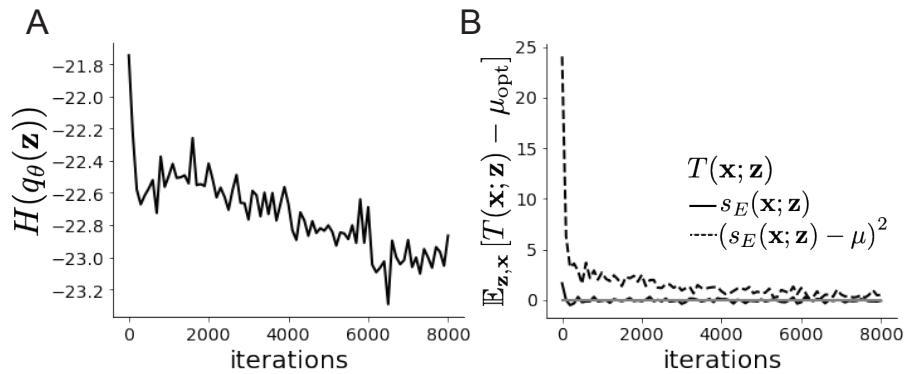


Figure 16: (V1 4) Optimization for V1

1238 **5.4.4 Primary visual cortex: Mathematical intuition and challenges**

1239 The dynamical system that we are working with can be written as

$$\begin{aligned} dx &= \frac{1}{\tau}(-x + f(Wx + h + \epsilon))dt \\ d\epsilon &= -\frac{dt}{\tau_{\text{noise}}} \epsilon + \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}} \Sigma_{\epsilon} dW \end{aligned} \quad (77)$$

1240 Where in this paper we chose

$$\Sigma_{\epsilon} = \tau_{\text{noise}} \begin{bmatrix} \tilde{\sigma}_E & 0 & 0 & 0 \\ 0 & \tilde{\sigma}_P & 0 & 0 \\ 0 & 0 & \tilde{\sigma}_S & 0 \\ 0 & 0 & 0 & \tilde{\sigma}_V \end{bmatrix} \quad (78)$$

1241 where $\tilde{\sigma}_\alpha$ is the reparameterized standard deviation of the noise for population α from Equation
 1242 68.

1243 In order to compute this covariance, we define $v = \omega x + h + \epsilon$ and $S = I - \omega f'(v)$, to re-write Eq.
 1244 (77) as an 8-dimensional system:

$$d \begin{pmatrix} \delta v \\ \epsilon \end{pmatrix} = - \begin{pmatrix} S & -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I \\ 0 & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} \begin{pmatrix} \delta v \\ \epsilon \end{pmatrix} dt + \begin{pmatrix} 0 & \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}} \Sigma_\epsilon \\ 0 & \frac{\sqrt{2}}{\sqrt{\tau_{\text{noise}}}} \Sigma_\epsilon \end{pmatrix} d\mathbf{W} \quad (79)$$

1245 Where $d\mathbf{W}$ is a vector with the private noise of each variable. The $d\mathbf{W}$ term is multiplied by a
 1246 non-diagonal matrix is because the noise that the voltage receives is the exact same than the one
 1247 that comes from the OU process and not another process. The solution of this problem is given by
 1248 the Lyapunov Equation [60, 68]:

$$\begin{pmatrix} S & -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I \\ 0 & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} \begin{pmatrix} \Lambda_v & \Lambda_c \\ \Lambda_c^T & \Lambda_\epsilon \end{pmatrix} + \begin{pmatrix} \Lambda_v & \Lambda_c \\ \Lambda_c^T & \Lambda_\epsilon \end{pmatrix} \begin{pmatrix} S^T & 0 \\ -\frac{\tau_{\text{noise}} - \tau}{\tau \tau_{\text{noise}}} I & \frac{1}{\tau_{\text{noise}}} I \end{pmatrix} = \begin{pmatrix} \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon & \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon \\ \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon & \frac{2}{\tau_{\text{noise}}} \Lambda_\epsilon \end{pmatrix} \quad (80)$$

1249 To obtain an equation for Λ_v , we solve this block matrix multiplication:

$$S\Lambda_v + \Lambda_v S^T = \frac{2\Lambda_\epsilon}{\tau_{\text{noise}}} + \frac{\tau_{\text{noise}}^2 - \tau^2}{(\tau \tau_{\text{noise}})^2} \left(\left(\frac{1}{\tau_{\text{noise}}} I + S \right)^{-1} \Lambda_\epsilon + \Lambda_\epsilon \left(\frac{1}{\tau_{\text{noise}}} I + S^T \right)^{-1} \right) \quad (81)$$

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 given S . Due to symmetry arguments [68], in this case the diagonal elements of the covariance matrix of the voltage will have the form:

$$\Lambda_{v_{ii}} = \sum_{i=\{E,P,S,V\}} g_i(S) \sigma_{ii}^2 \quad (82)$$

1250 These coefficients $g_i(S)$ are complicated functions of the Jacobian of the system. Although expres-
 1251 sions for these coefficients can be found explicitly, only numerical evaluation of those expressions
 1252 determine which components of the noisy input are going to strongly influence the variability of ex-
 1253 citatory population. Showing the generality of this dependence in more complicated noise scenarios
 1254 (e.g. $\tau_{\text{noise}} > \tau$ as in Section 3.4), is the focus of current research.

1255 **5.5 Superior colliculus**

1256 **5.5.1 SC model**

1257 The ability to switch between two separate tasks throughout randomly interleaved trials, or “rapid
 1258 task switching,” has been studied in rats, and midbrain superior colliculus (SC) has been shown to
 1259 play an important role in this computation [69]. Neural recordings in SC exhibited two populations of
 1260 neurons that simultaneously represented both task context (Pro or Anti) and motor response (con-
 1261 tralateral or ipsilateral to the recorded side), which led to the distinction of two functional classes:
 1262 the Pro/Contra and Anti/Ipsi neurons [46]. Given this evidence, Duan et al. proposed a model
 1263 with four functionally-defined neuron-type populations: two in each hemisphere corresponding to
 1264 the Pro/Contra and Anti/Ipsi populations. We study how the connectivity of this neural circuit
 1265 governs rapid task switching ability.

1266 The four populations of this model are denoted as left Pro (LP), left Anti (LA), right Pro (RP)
 1267 and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh\left(\frac{u_\alpha - a}{b}\right) + \frac{1}{2} \right), \quad (83)$$

1268 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-
 1269 earity. We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (84)$$

1270 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (85)$$

1271 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2 . These
 1272 hyperparameter values are motivated by modeling choices and results from [46].

1273 The weight matrix has 4 parameters for self sW , vertical vW , horizontal hW , and diagonal dW
 1274 connections:

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (86)$$

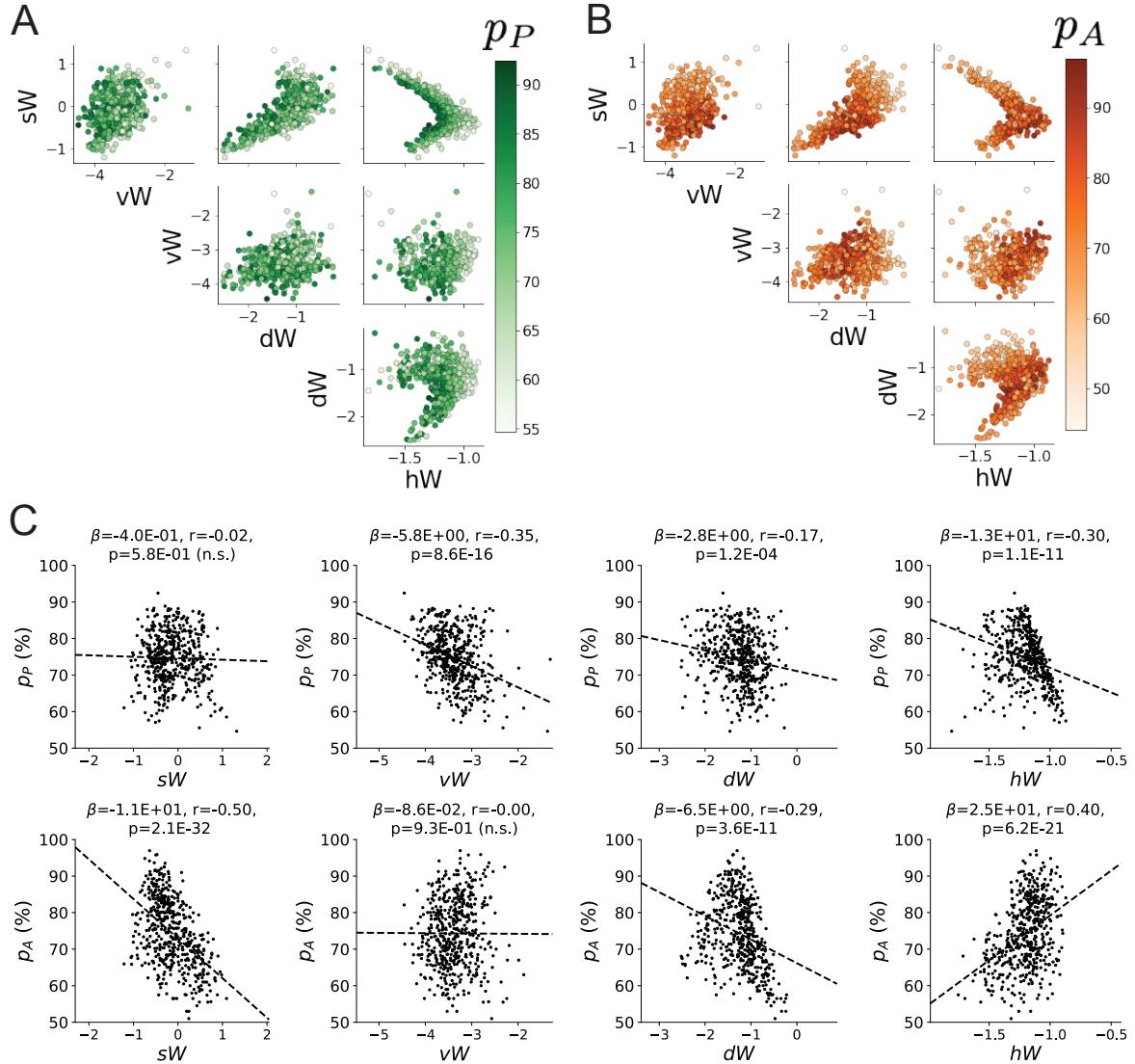


Figure 17: (SC1): **A.** Same pairplot as Fig. 4C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

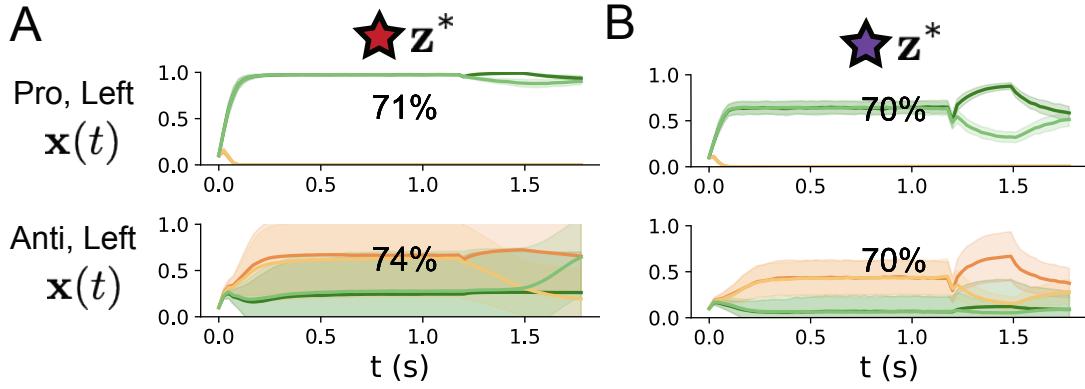


Figure 18: (SC2): **A.** Simulations in network regime 1 ($hW_{\text{fixed}} = -1.5$). **B.** Simulations in network regime 2 ($hW_{\text{fixed}} = -1.5$) .

₁₂₇₅ We study the role of parameters $\mathbf{z} = [sW, vW, hW, dW]^\top$ in rapid task switching.

₁₂₇₆ The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (87)$$

₁₂₇₇ There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (88)$$

₁₂₇₈ a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (89)$$

₁₂₇₉ rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (90)$$

₁₂₈₀

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (91)$$

₁₂₈₁ a choice-period input

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

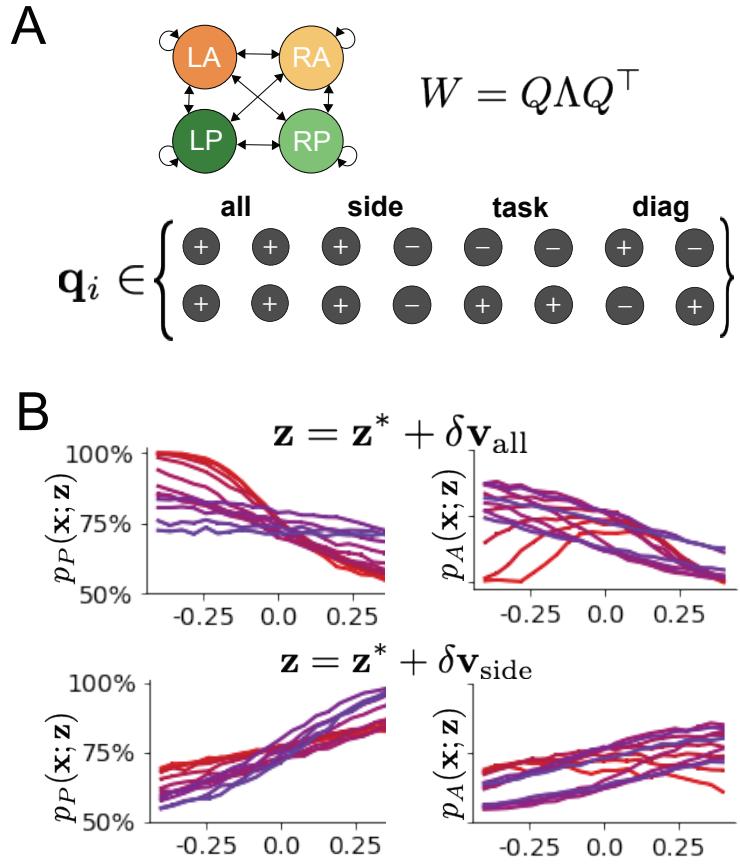


Figure 19: (SC3): **A.** Invariant eigenvectors of connectivity matrix W . **B.** Accuracies for connectivity perturbations for increasing λ_{all} and λ_{side} (rest shown in Fig. 4D).

1282 and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases} \quad (93)$$

1283 The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{P,\text{bias}} = 0.5$, $I_{P,\text{rule}} = 0.6$, $I_{A,\text{rule}} = 0.6$,

1284 $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

1285 5.5.2 Task accuracy calculation

1286 The accuracies of each task p_P and p_A are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (94)$$

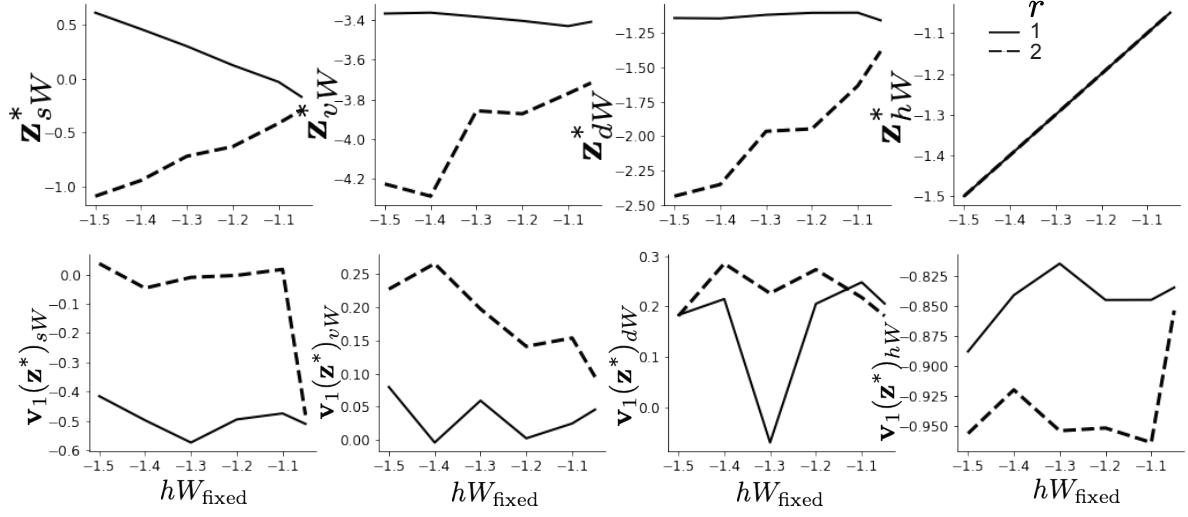


Figure 20: (SC4): **A.** The individual parameters of each mode throughout the two regimes. **B.** The individual sensitivities of parameters of each mode throughout the two regimes.

1287 and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (95)$$

1288 given that the stimulus is on the left side, where Θ is the Heaviside step function, and the accuracy
1289 is averaged over 200 independent trials. The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (96)$$

1290 where $\beta = 100$.

1291 5.5.3 EPI details for the SC model

1292 Writing the EPI distribution as a maximum entropy distribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these
1293 first and second moments of the accuracy in each task (as in Equations 18 and 19)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \\ (p_P(\mathbf{x}; \mathbf{z}) - .75)^2 \\ (p_A(\mathbf{x}; \mathbf{z}) - .75)^2 \end{bmatrix}, \quad (97)$$

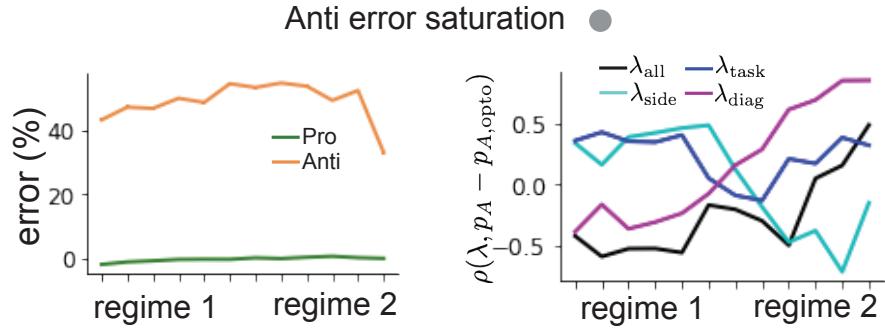


Figure 21: (SC5): (Left) Mean and standard error of Pro and Anti error from regime 1 to regime 2 at $\gamma = 0.85$. (Right) Correlations of connectivity eigenvalues with Anti error from regime 1 to regime 2 at $\gamma = 0.85$.

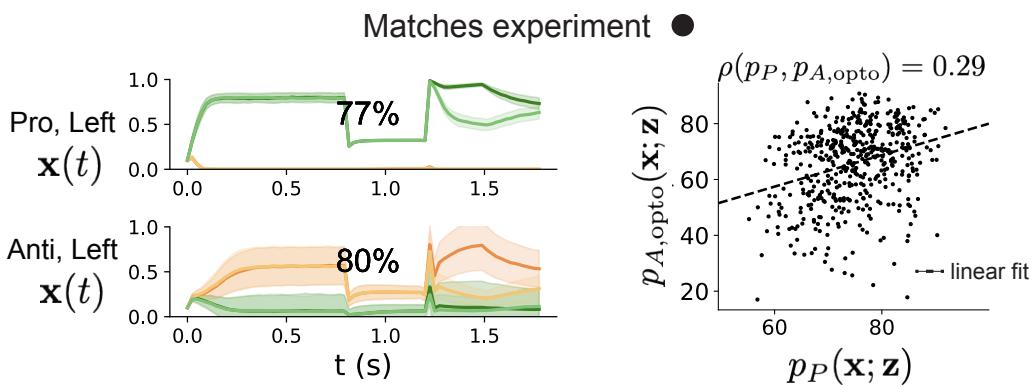


Figure 22: (SC6): (Left) Mean and standard deviation (shading) of responses of the SC model at the mode of the EPI distribution to delay period inactivation at $\gamma = 0.675$. (Right) Anti accuracy following delay period inactivation at $\gamma = 0.675$ versus accuracy in the Pro task across connectivities in the EPI distribution.

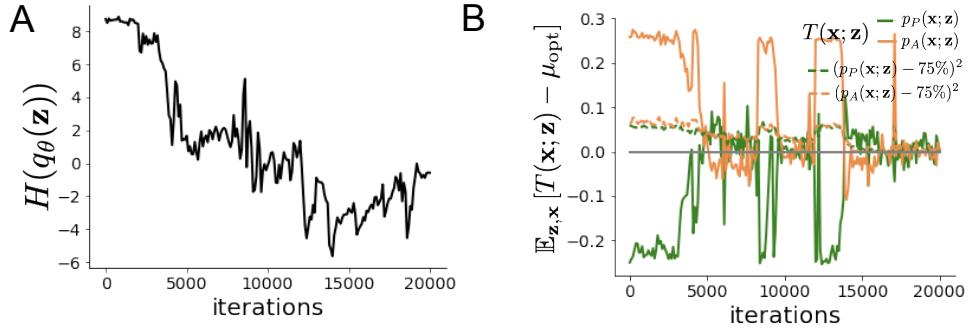


Figure 23: (SC7): **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

1294

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} .75 \\ .75 \\ .075^2 \\ .075^2 \end{bmatrix}. \quad (98)$$

1295 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 1296 epoch of 2,000 iterations (see Section 5.1.4). The optimization converged after ten epochs (Fig.
 1297 22).

1298 For EPI in Fig. 4C, we used a real NVP architecture with three coupling layers of affine transfor-
 1299 mations parameterized by two-layer neural networks of 50 units per layer. The initial distribution
 1300 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
 1301 augmented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution
 1302 was the greatest EPI distribution to converge across 5 random seeds with criteria $N_{\text{test}} = 25$.

1303 The bend in the EPI distribution is not a spurious result of the EPI optimization. The structure
 1304 discovered by EPI matches the shape of the set of points returned from brute-force random sampling
 1305 (Fig. 24A) These connectivities were sampled from a uniform distribution over the range of each
 1306 connectivity parameter, and all parameters producing accuracy in each task within the range of
 1307 60% to 90% were kept. This set of connectivities will not match the distribution of EPI exactly,
 1308 since it is not conditioned on the emergent property. For example the parameter set returned by
 1309 the brute-force search is biased towards lower accuracies (Fig. 24B).

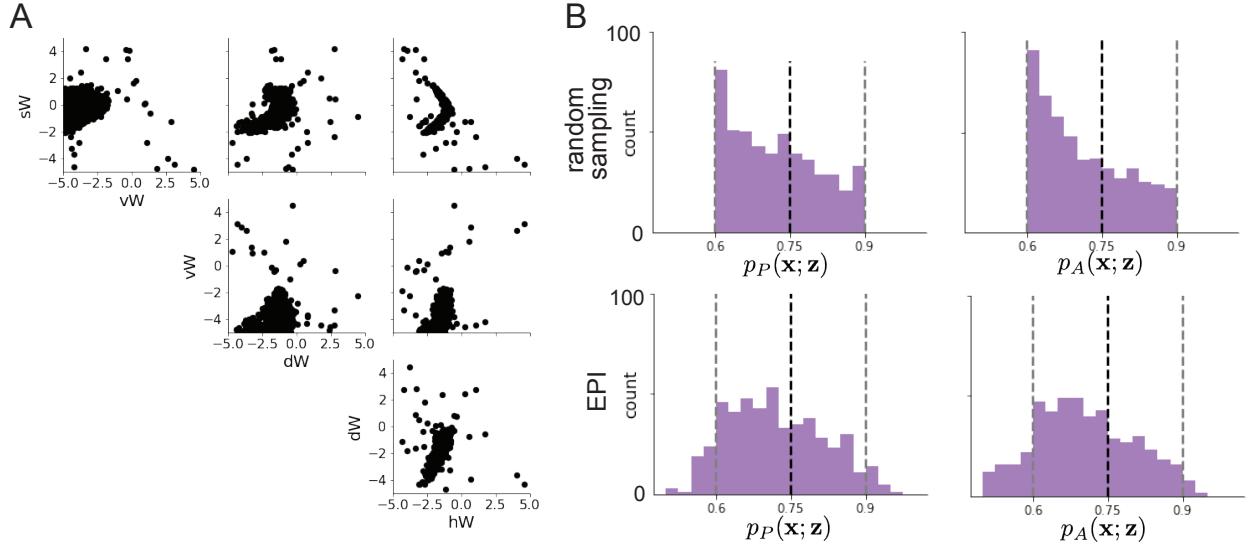


Figure 24: (SC8): **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

1310 5.5.4 Regime identification with EPI

1311 We sought two sets of parameters from $q_{\theta}(\mathbf{z} | \mathcal{X})$ that were representative of each regime, so that we
 1312 could assess their implications on computation. For fixed values of hW , we hypothesized that there
 1313 are two modes: one in each regime of greater and lesser sW . To begin, we found one mode for each
 1314 regime at $hW_{\text{fixed}} = -1.5$ using 200 steps of gradient ascent of the deep probability distribution
 1315 $q_{\theta}(\mathbf{z} | \mathcal{X})$. In regime 1, the initialization had positive sW , and the initialization had negative sW
 1316 in regime 2, which led to disparate modes (Fig. 20 top). These modes were then used as the
 1317 initialization to find the next mode at $hW_{\text{fixed}} = -1.4$ and so on. 200 steps of gradient ascent
 1318 were always taken, and learning rates of 2.5×10^{-4} and 5×10^{-4} were used for regimes 1 and 2,
 1319 respectively. Each of these modes is denoted $\mathbf{z}^*(hW_{\text{fixed}}, r)$ for regime $r \in \{1, 2\}$.

1320 For the analyses in Figure 5C and Figure 21, we obtained parameters for each step along the
 1321 continuum between regimes 1 and 2 by sampling from the EPI distribution. Each sample was
 1322 assigned to the closest mode $\mathbf{z}^*(hW_{\text{fixed}}, r)$. Sampling continued until 500 samples were assigned to
 1323 each mode, which took 7.36 seconds. To obtain this many samples for each mode with brute force
 1324 sampling over the chosen prior, this would take 4.20 days.

1325 **5.5.5 Sensitivity analysis**

1326 At each mode, we measure the sensitivity dimension (that of most negative eigenvalue in the Hessian
 1327 of the EPI distribution) $\mathbf{v}_1(\mathbf{z}^*)$. To resolve sign degeneracy in eigenvectors, we chose $\mathbf{v}_1(\mathbf{z}^*)$ to have
 1328 negative element in hW . This tells us what parameter combination rapid task switching is most
 1329 sensitive to at this parameter choice in the regime. We see that while the modes of each regime
 1330 gradually converge to similar connectivities at $hW_{\text{fixed}} = -1.05$ (Fig. 20 top), the sensitivity
 1331 dimensions remain categorically different throughout the two regimes (Fig. 20 bottom). Only at
 1332 $hW_{\text{fixed}} = -1.05$ is there a flip in sensitivity from regime 2 to regime 1 (in $\mathbf{v}_1(\mathbf{z}^*)_{sW}$ and $\mathbf{v}_1(\mathbf{z}^*)_{hW}$).
 1333 There is thus some ambiguity regarding the “regime” of $\mathbf{z}^*(-1.05, 2)$, since the mode is derived
 1334 from an initialization in regime 2, but has sensitivity like regime 1. We can consider this as an
 1335 intermediate transitional region of parameter space between the two regimes. To emphasize this,
 1336 $\mathbf{z}^*(-1.05, 1)$ and $\mathbf{z}^*(-1.05, 2)$ have the same color.

1337 **5.5.6 Connectivity eigendecomposition and processing modes**

1338 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-
 1339 tion of the symmetric connectivity matrices $W = Q\Lambda Q^{-1}$, which results in the same basis vectors
 1340 \mathbf{q}_i for all W parameterized by \mathbf{z} (Fig. 19A). These basis vectors have intuitive roles in processing for
 1341 this task, and are accordingly named the *all* eigenmode - all neurons co-fluctuate, *side* eigenmode
 1342 - one side dominates the other, *task* eigenmode - the Pro or Anti populations dominate the other,
 1343 and *diag* mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair.
 1344 Due to the parametric structure of the connectivity matrix, the parameters \mathbf{z} are a linear function
 1345 of the eigenvalues $\boldsymbol{\lambda} = [\lambda_{\text{all}}, \lambda_{\text{side}}, \lambda_{\text{task}}, \lambda_{\text{diag}}]^\top$ associated with these eigenmodes.

$$\mathbf{z} = A\boldsymbol{\lambda} \tag{99}$$

1346

$$A = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}. \tag{100}$$

1347 We are interested in the effect of raising or lowering the amplification of each eigenmode in the
 1348 connectivity matrix. To test this, we calculate the unit vector of changes in the connectivity \mathbf{z} that

1349 result from a change in the associated eigenvalues

$$\mathbf{v}_a = \frac{\frac{\partial \mathbf{z}}{\partial \lambda_a}}{\left\| \frac{\partial \mathbf{z}}{\partial \lambda_a} \right\|_2}, \quad (101)$$

1350 where

$$\frac{\partial \mathbf{z}}{\partial \lambda_a} = A \mathbf{e}_a, \quad (102)$$

1351 and e.g. $\mathbf{e}_{\text{all}} = [1, 0, 0, 0]^\top$. So \mathbf{v}_a is the normalized column of A corresponding to eigenmode a .

1352 While perturbations in the sensitivity dimension $\mathbf{v}_1(\mathbf{z}^*)$ adapt with the mode \mathbf{z}^* chosen, perturba-
1353 tions in \mathbf{v}_a for $a \in \{\text{all, side, text, diag}\}$ are invariant to \mathbf{z} (Equation 102).

1354 To understand the connectivity mechanisms that distinguish these two regimes, we perturb connec-
1355 tivity at each mode in dimensions that have well defined roles in processing for the Pro and Anti
1356 tasks. A convenient property of this connectivity parameterization is that there are \mathbf{z} -invariant
1357 eigenmodes of connectivity, whose eigenvalues (or degree of amplification) change with \mathbf{z} . These
1358 eigenmodes have intuitive roles in processing in each task, and are accordingly named the *all*,
1359 *side*, *task*, and *diag* eigenmodes (see Section 5.5). Furthermore, the parameter dimension \mathbf{v}_a
1360 ($a \in \{\text{all, side, task, and diag}\}$) that increases the eigenvalue of connectivity λ_a is \mathbf{z} -invariant (un-
1361 like the sensitivity dimension $\mathbf{v}_1(\mathbf{z})$) and $\mathbf{v}_a \perp \mathbf{v}_{b \neq a}$. Thus, by changing the degree of amplification
1362 of each processing mode by perturbing \mathbf{z} along \mathbf{v}_a , we can elicit the differentiating properties of
1363 the two regimes.

1364 5.5.7 Modeling optogenetic silencing.

1365 We tested whether the inferred SC model connectivities could reproduce experimental effects of
1366 optogenetic inactivation in rats [69]. During periods of simulated optogenetic inactivation, activity
1367 was decreased proportional to the optogenetic strength $\gamma \in [0, 1]$

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (103)$$

1368 Delay period inactivation was from $0.8 < t < 1.2$.