

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 experi-
⁴ mentally observed phenomenon – whether behavioral or a pattern of neural activity – and thus
⁵ can offer insights into neural computation. The operation of these mechanistic circuits, like all
⁶ models, critically depends on the choices of model parameters. A key process in circuit modeling
⁷ is then to identify the model parameters consistent with observed phenomena: to solve the inverse
⁸ problem. To solve challenging inverse problems modeling neural datasets, neuroscientists have used
⁹ statistical inference techniques to much success. However, most research in theoretical neuroscience
¹⁰ focuses on how computation emerges in biologically interpretable circuit models, and how the model
¹¹ parameters govern computation; it is not focused on the latent structure of empirical models of
¹² noisy experimental datasets. In this work, we present a novel technique that brings the power
¹³ and versatility of the probabilistic modeling toolkit to theoretical inverse problems. Our method
¹⁴ uses deep neural networks to learn parameter distributions with rich structure that have specific
¹⁵ computational properties in biologically relevant models. This methodology is explained through
¹⁶ a motivational example inferring conductance parameters in an STG subcircuit model. Then, with
¹⁷ RNNs of increasing size, we show that only EPI allows precise control over the behavior of inferred
¹⁸ parameters, and that EPI scales better in parameter dimension than alternative techniques. In the
¹⁹ remainder of this work, we explain novel theoretical insights through the examination of intricate
²⁰ parametric structure in complex circuit models. In a model of primary visual cortex with multiple

21 neuron-types, where analysis becomes untenable with each additional neuron-type, we discovered
22 how noise distributed across neuron-types governs the excitatory population. Finally, in a model
23 of superior colliculus, we identified and characterized two distinct regimes of connectivity that
24 facilitate switching between opposite tasks amidst interleaved trials. We also found that all task-
25 switching connectivities in this model reproduce behaviors from inactivation experiments, further
26 establishing this hypothesized circuit model. Beyond its scientific contribution, this work illustrates
27 the variety of analyses possible once deep learning is harnessed towards solving theoretical inverse
28 problems.

29 2 Introduction

30 The fundamental practice of theoretical neuroscience is to use a mathematical model to understand
31 neural computation, whether that computation enables perception, action, or some intermediate
32 processing. A neural circuit is systematized with a set of equations – the model – and these
33 equations are motivated by biophysics, neurophysiology, and other conceptual considerations [1,
34 2, 3, 4]. The function of this system is governed by the choice of model *parameters*, which when
35 configured in a particular way, give rise to a measurable signature of a computation. The work
36 of analyzing a model then requires solving the inverse problem: given a computation of interest,
37 how can we reason about particular parameter configurations? The inverse problem is crucial for
38 reasoning about likely parameter values, uniquenesses and degeneracies, and predictions made by
39 the model [5, 6].

40 Consider the idealized practice: one carefully designs a model and analytically derives how com-
41 putational properties determine model parameters. Seminal examples of this gold standard (which
42 often adopt approaches from statistical physics) include our field’s understanding of memory ca-
43 pacity in associative neural networks [7], chaos and autocorrelation timescales in random neural
44 networks [8], the paradoxical effect [9], and decision making [10]. Unfortunately, as circuit models
45 include more biological realism, theory via analytical derivation becomes intractable. Still, we can
46 gain insight into these complex models by identifying the distribution of parameters that produce
47 computations. By solving the inverse problem in this way, scientific analysis of complex biologi-
48 cally realistic models is made possible [11, 12, 13, 6, 14]. However, this work clarifies an important
49 incongruity between this methodology and theoretical approaches to neuroscience.

50 A pivotal detail in theoretical neuroscience is the manner in which the computation is specified. One

51 common approach is to use data that is exemplary of that computation, so that the framework of
52 statistical inference can be applied for parameter identification. While a host of suitable algorithms
53 have been developed for inferring parameters in models of neural data [15, 16, 17, 18, 19, 20, 21,
54 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36] (see review, [37]), the level of insight gained
55 strongly depends on the quantity and quality of available data. Ultimately theoretical neuroscience
56 is concerned with the computational properties – the *emergent phenomena* – of our models [7, 8,
57 9, 10], not noisy observed datasets. To use the aforementioned inference paradigm, scientists must
58 shoehorn such mathematical criteria into an artificial dataset compatible with existing statistical
59 approaches.

60 In addition, another crucial challenge in neuroscientific modeling has been the inversion of bio-
61 logically realistic or “mechanistic” models. Most neural circuit models in theoretical neuroscience
62 are noisy systems of differential equations that can only be sampled or realized through forward
63 simulation; they lack the explicit likelihood necessary for statistical inference. Therefore, the most
64 popular approaches to theoretical inverse problems have been likelihood-free inference (LFI) meth-
65 ods [38, 39], in which reasonable parameters are obtained via simulation and rejection. A flourishing
66 new class of techniques [40, 41, 42] use deep learning to improve upon traditional LFI approaches.
67 However, to use these methods in theoretical neuroscience, we still must represent computation with
68 data in some way. Theorists are therefore barred from using the probabilistic modeling toolkit for
69 science with circuit models, unless they reformulate their inverse problem into an empirical frame-
70 work.

71 These challenges motivate the development of a novel inference framework called emergent property
72 inference (EPI). As an adaption of variational inference [43], EPI infers parameter distributions
73 that produce an emergent property: not a singular dataset, but a collection of datasets exhibiting
74 some mathematical criteria. EPI constrains the predictions of the inferred parameter distribution
75 to produce the emergent property, which requires a variant of probabilistic inference methods [44].
76 Importantly, EPI uses deep learning to make rich, flexible approximations to the parameter distri-
77 bution [45] that produces an emergent property. The structure captured by these deep probability
78 distributions are scientifically valuable, revealing the sensitivity and robustness of the emergent
79 property to different parameter combinations. Perhaps most powerfully, EPI facilitates inference
80 in mechanistic models, allowing theorists to capture rich parametric structure in biologically real-
81 istic models that is conditioned upon the emergent phenomena of interest.

82 Equipped with this method, we prove out the potential of EPI by demonstrating its capabilities and

83 presenting novel theoretical findings borne from its analysis. First, we show EPI’s ability to handle
84 mechanistic models using a classic model of parametric degeneracy in biology: the stomatogastric
85 ganglion [46, 47]. Then, we show EPI’s superior scaling properties by inferring connectivities of re-
86 current neural networks (RNNs) that exhibit stable, yet amplified responses – a hallmark of neural
87 responses throughout the brain [48, 49, 50]. In a model of primary visual cortex (V1) with different
88 neuron-types, we show that the equation for excitatory variability become analytically intractable
89 as more populations are added. Strikingly, the way in which noisy inputs across neuron-types
90 governs excitatory variability is salient in the visualized structure of the EPI inferred parameter
91 distribution. Finally, we investigated the possible connectivities of superior colliculus (SC) that al-
92 low execution of different tasks on interleaved trials. EPI discovered a rich distribution containing
93 two connectivity regimes with different solution classes. We quantified parametric sensitivity and
94 robustness in each regime by simply querying the deep probability distribution learned by EPI,
95 which produced a mechanistic understanding of cortical responses in each regime. Intriguingly,
96 all inferred connectivities reproduced results from optogenetic inactivation experiments in this be-
97 havioral paradigm – emergent phenomena that EPI was not conditioned upon. These theoretical
98 insights afforded by EPI illustrate the value of deep inference for the interrogation of neural circuit
99 models.

100 3 Results

101 3.1 Motivating emergent property inference of theoretical models

102 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-
103 erty inference. First, one designs or chooses an existing model that, it is hypothesized, captures
104 the computation of interest. To ground this process in a well-known example, consider the stom-
105 atogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic
106 muscle activation patterns for digestion [51]. Despite full knowledge of STG connectivity and a
107 precise characterization of its rhythmic pattern generation, biophysical models of the STG have
108 complicated relationships between circuit parameters and computation [46, 12]. A subcircuit model
109 of the STG [47] is shown schematically in Figure 1A. The jagged connections indicate electrical cou-
110 pling having electrical conductance g_{el} , smooth connections in the diagram are inhibitory synaptic
111 projections having strength g_{synA} onto the hub neuron, and $g_{synB} = 5nS$ for mutual inhibitory con-
112 nections. Note that the behavior of this model will be critically dependent on its parameterization

– the choices of conductance parameters $\mathbf{z} = [g_{el}, g_{synA}]$. Specifically, the two fast neurons ($f1$ and $f2$) mutually inhibit one another, and oscillate at a faster frequency than the mutually inhibiting slow neurons ($s1$ and $s2$). The hub neuron (hub) couples with either the fast or slow population, or both.

Second, once the model is selected, one must specify what the model should produce. In typical statistical inference, this is a dataset – either collected experimentally or constructed by scientists to fit this empirical paradigm. In EPI, a different approach is taken, in which we define an emergent property: a set of mathematical criteria to be obeyed by the datasets predicted by the inferred distribution. In the STG example, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model 1B. An interesting emergent property of this stochastic model is when the hub neuron fires at an intermediate frequency between the intrinsic spiking rates of the fast and slow populations. This emergent property is shown in Figure 1C at an average frequency of 0.55Hz.

Third, the model parameters producing these outputs are inferred. Most often, brute-force parameter sweeps or rejection sampling techniques [39] are used to identify parameters whose model simulations are close to data or some desired feature. In this last step lies the opportunity for a paradigmatic shift away from empirical data-oriented representations of model output. By precisely quantifying the emergent property of interest as a statistical feature of the model, we can infer a probability distribution over parameter configurations that produce this emergent property. This unlocks the deep probabilistic modeling toolkit for treating theoretical inverse problems.

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



Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based biophysical model of the STG subcircuit. **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 model and emergent property, emergent property inference (EPI) learns a deep probability distribution of parameters \mathbf{z} . **E.** The EPI posterior 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 degeneracy \mathbf{v}_2 . **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.

143 **3.2 A deep generative modeling approach to emergent property inference**

144 Emergent property inference (EPI) formalizes the three-step procedure of the previous section with
 145 deep probability distributions. First, as is typical, we consider the model as a coupled set of differ-
 146 ential equations [47]. In the running STG example, the model activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$ is
 147 the membrane potential for each neuron, which evolves according to the biophysical conductance-
 148 based equation:

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

149 where $C_m = 1\text{nF}$, and \mathbf{h} is a sum of the leak, calcium, potassium, hyperpolarization, electrical, and
 150 synaptic currents, all of which have their own complicated dependence on activity \mathbf{x} and parameters
 151 $\mathbf{z} = [g_{el}, g_{synA}]$, and $d\mathbf{B}$ is white gaussian noise (see Section 5.2.1 for more detail).

152 Second, we define the emergent property, which as above is “intermediate hub frequency” (Figure
 153 1C). Quantifying this phenomenon is straightforward: we stipulate that the hub neuron’s spiking
 154 frequency – denoted $\omega_{hub}(\mathbf{x})$ is close to a frequency of 0.55Hz. Mathematically, we achieve this
 155 with two constraints: by fixing the mean hub frequency over the inferred parameter distribution of
 156 \mathbf{z} and its resulting simulations \mathbf{x} to 0.55Hz,

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

157 and requiring that the variance of the hub frequency over the produced simulations is small

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

158 This level of variance was chosen to be low enough to exclude the fast and slow frequencies of
 159 the two populations, but large enough to allow structural examination of the inferred parameter
 160 distribution. By constraining the means and variances of emergent property statistics over \mathbf{z} as
 161 well as the stochasticity of \mathbf{x} , we can precisely control the behavior that the inferred distribution
 162 that EPI infers. In general, an emergent property

$$\mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \text{Var}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2 \quad (4)$$

163 defines a collection of datasets with a statistic $f(\mathbf{x}; \mathbf{z})$ (which may be comprised of multiple statis-
 164 tics) and the means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ of those statistics over the datasets. The choice of $\boldsymbol{\sigma}^2$
 165 predicates the degree of variability around the mean $\boldsymbol{\mu}$ that is consistent with the emergent prop-
 166 erty.

167 Third, we perform emergent property inference: we find a distribution over parameter configura-
168 tions \mathbf{z} , and insist that samples from this distribution produce the emergent property; in other
169 words, they obey the constraints introduced in Equation 4. This distribution will be chosen from a
170 family of probability distributions $\mathcal{Q} = \{q_{\theta}(\mathbf{z}) : \theta \in \Theta\}$, defined by a deep neural network [45, 52, 53]
171 (Figure 1D, EPI box). Deep probability distributions map a simple random variable \mathbf{z}_0 through a
172 deep neural network with weights and biases θ to parameters $\mathbf{z} = g_{\theta}(\mathbf{z}_0)$ to a suitable complicated
173 distribution (see Section 5.1.2 for more details). Many distributions in \mathcal{Q} will respect the emergent
174 property constraints, so we select the most random (or “entropic”) distribution, which is the same
175 choice made in Bayesian inference (see Section 5.1.6). In EPI optimization, stochastic gradient
176 steps in θ are taken such that entropy is maximized, and the emergent property \mathcal{X} is produced
177 (see Section 5.1) The inferred EPI distribution is denoted $q_{\theta}(\mathbf{z} | \mathcal{X})$, to emphasize that we have
178 conditioned our parameter distribution on emergent property \mathcal{X} .

179 The major scientific value of EPI is in the rich, queryable structure of these deep probability
180 distributions. The probabilities of $q_{\theta}(\mathbf{z} | \mathcal{X})$ are the densities of these parameters in the distribution
181 producing the emergent property. The greatest probabilities (the modes) indicate prototypical
182 parameter configurations, and the manner in which probabilities change away from the modes shows
183 how different parameter combinations preserve or diminish the emergent property. The dimensions
184 of greatest sensitivity (e.g. Fig. 1E solid arrow) or degeneracy (e.g. Fig. 1E dashed arrow) can be
185 measured directly from the second order derivative of $\log q_{\theta}(\mathbf{z} | \mathcal{X})$ called the “Hessian.” Around
186 the mode, eigenvalues of the Hessian are negative; probabilities decrease locally in all directions
187 away from the mode. The eigenvector with most negative eigenvalue is the parameter combination
188 causing probability to decrease the fastest, making it the most sensitive dimension. Likewise, the
189 flattest eigenvector, corresponding to the least negative eigenvalue, points in the most degenerate
190 dimension. Once an EPI distribution has been inferred, this second order derivative requires trivial
191 computation (when correct architecture class is chosen, see Section 5.1.2).

192 In the following sections, we showcase the versatility of EPI for scientific analysis on three neural
193 circuit models across ranges of biological realism, neural system function, and network scale. First,
194 we demonstrate the superior scalability of EPI compared to alternative techniques by inferring high-
195 dimensional distributions of RNN connectivities that exhibit amplified, yet stable responses. Also
196 in this RNN example, we emphasize that EPI is the only technique that controls the predictions
197 made by the inferred parameter distribution. Next, in a model of primary visual cortex [54,
198 55], we show how to gain insight by comparing multiple inferred distributions. Finally, we used

199 EPI to capture subtle parametric structure allowing the mechanistic characterization of multiple
 200 parametric regimes of superior colliculus activity in a model of task switching [56]. This work is
 201 the first to produce this level of theoretical insight via the quantification and examination of the
 202 intricate structure captured by deep probability distributions.

203 **3.3 Scaling inference of RNN connectivity with EPI**

204 Transient amplification is a hallmark of neural activity throughout cortex, and is often thought to
 205 be intrinsically generated by recurrent connectivity in the responding cortical area [48, 49, 50]. It
 206 has been shown that to generate such amplified, yet stabilized responses, the connectivity of RNNs
 207 must be non-normal [57, 48], and satisfy additional constraints [58]. In theoretical neuroscience,
 208 RNNs are optimized and then examined to show how dynamical systems could execute a given
 209 computation [59, 60], but such biologically realistic constraints on connectivity are ignored during
 210 optimization for practical reasons. In general, access to distributions of connectivity adhering to
 211 theoretical criteria like stable amplification, chaotic fluctuations [8], or low tangling [61] would add
 212 great scientific value and contextualization to existing research with RNNs. Here, we use EPI to
 213 learn RNN connectivities producing stable amplification, and demonstrate the superior scalability
 214 and efficiency of EPI to alternative approaches.

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

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

216 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 $\chi_{i,j}^{(U)}, \chi_{i,j}^{(V)} \sim$
 217 $\mathcal{N}(0, 1)$. 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 am-
 218 plification. Two conditions are necessary and sufficient for RNNs to exhibit stable amplification
 219 [58]: $\text{real}(\lambda_1) < 1$ and $\lambda_1^s > 1$, where λ_1 is the eigenvalue of W with greatest real part and λ^s is
 220 the maximum eigenvalue of $W^s = \frac{W+W^\top}{2}$. RNNs with $\text{real}(\lambda_1) = 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$
 221 will be stable with modest decay rate ($\text{real}(\lambda_1)$ close to its upper bound of 1) and exhibit modest
 222 amplification (λ_1^s close to its lower bound of 1). EPI can naturally condition on this emergent
 223 property

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



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 $\mathbf{x}_0 = \mu$ (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.

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 likelihood-free inference approaches. We ran sequential Monte Carlo approximate Bayesian computation (SMC-ABC) [38] and sequential neural posterior estimation (SNPE) [40] with observation $\mathbf{x}_0 = \boldsymbol{\mu}$. SMC-ABC is a rejection sampling approach that 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 control the mean or variance 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 5.2.2).

As the number of neurons N in the RNN are 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 mightily in dimensions greater than about 30 [62], yet we were careful to assess the scalability of the more comparable approach SNPE. 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). From this analysis, we see that deep inference techniques EPI and SNPE are far more amenable to inference of high dimensional parameter distributions than rejection sampling techniques like SMC-ABC, and that EPI outperforms SNPE in both criteria in high-dimensions.

No matter the number of neurons, EPI always produces connectivity distributions with mean and variance of $\text{real}(\lambda_1)$ and λ_1^s according to \mathcal{X} (Fig. 2C, blue). For the dimensionalities in which SMC-ABC is tractable, the inferred parameters are concentrated and offset from \mathbf{x}_0 (Fig. 2C, green). When using SNPE the predictions of the inferred parameters are highly concentrated at some RNN sizes and widely varied in others (Fig. 2C, orange). We see these properties reflected in simulations from the inferred distributions: EPI produces a consistent variety of stable, amplified activity norms $|r(t)|$, SMC-ABC produces a limited variety of responses, and the changing variety of responses from SNPE emphasizes the control of EPI on parameter predictions.

Through this example, we have shown that EPI can be used for well-controlled insight into RNNs

256 with respect to their theoretical properties. EPI outperforms SNPE in high dimensions by using
 257 gradient information (from $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z}) = \nabla_{\mathbf{z}}[\text{real}(\lambda_1), \lambda_1^s]^\top$) on each optimization iteration. This
 258 agrees with recent speculation that such gradient information could improve the efficiency of LFI
 259 techniques [85]. While scaling to high dimensions is important, we show in the next two sections
 260 how insight can be gained by inspecting structure in lower dimensional parameter distributions.

261 **3.4 EPI reveals how noisy input across neuron-types governs excitatory vari-
 262 ability in a V1 model**

263 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output
 264 function have succeeded in explaining a host of experimentally documented phenomena. In a
 265 regime characterized by inhibitory stabilization of strong recurrent excitation, these models give
 266 rise to paradoxical responses [9], selective amplification [57, 48], surround suppression [63] and
 267 normalization [64]. Despite their strong predictive power, E-I circuit models rely on the assump-
 268 tion that inhibition can be studied as an indivisible unit. However, experimental evidence shows
 269 that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) –
 270 composing 80% of GABAergic interneurons in V1 [65, 66, 67], and that these inhibitory cell types
 271 follow specific connectivity patterns (Fig. 3A) [68]. While research has shown that V1 only shares
 272 specific dimensions of neuronal variability with downstream areas [69], the role played by recurrent
 273 dynamics and the connectivity across neuron-type populations is not understood. Here, in a model
 274 of V1 with biologically realistic connectivity, we use EPI to show how the structure of input across
 275 neuron types affects the variability of the excitatory population – the population largely responsible
 276 for projecting to other brain areas [70].

277 We considered response variability of a nonlinear dynamical V1 circuit model (Fig. 3A) with a
 278 state comprised of each neuron-type population’s rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population
 279 receives recurrent input $W\mathbf{x}$, where W is the effective connectivity estimated from post-synaptic
 280 potential and connectivity rate measurements (see Section 5.2.3). Each population also experiences
 281 an external input \mathbf{h} , which determines population rate via supralinear nonlinearity $\phi(\cdot) = [\cdot]_+^2$. To
 282 build on previous work, we model visual contrast-dependent input to the E- and P-populations
 283 $\mathbf{h} = \mathbf{b} + c\mathbf{h}_c$. There is also an additive noisy input ϵ parameterized by variances for each neuron
 284 type population $\mathbf{z} = \sigma^2 = [\sigma_E^2, \sigma_P^2, \sigma_S^2, \sigma_V^2]$. This noise has a slower dynamical timescale $\tau_{\text{noise}} > \tau$
 285 then the population rate, allowing fluctuations around a stimulus-dependent steady-state

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

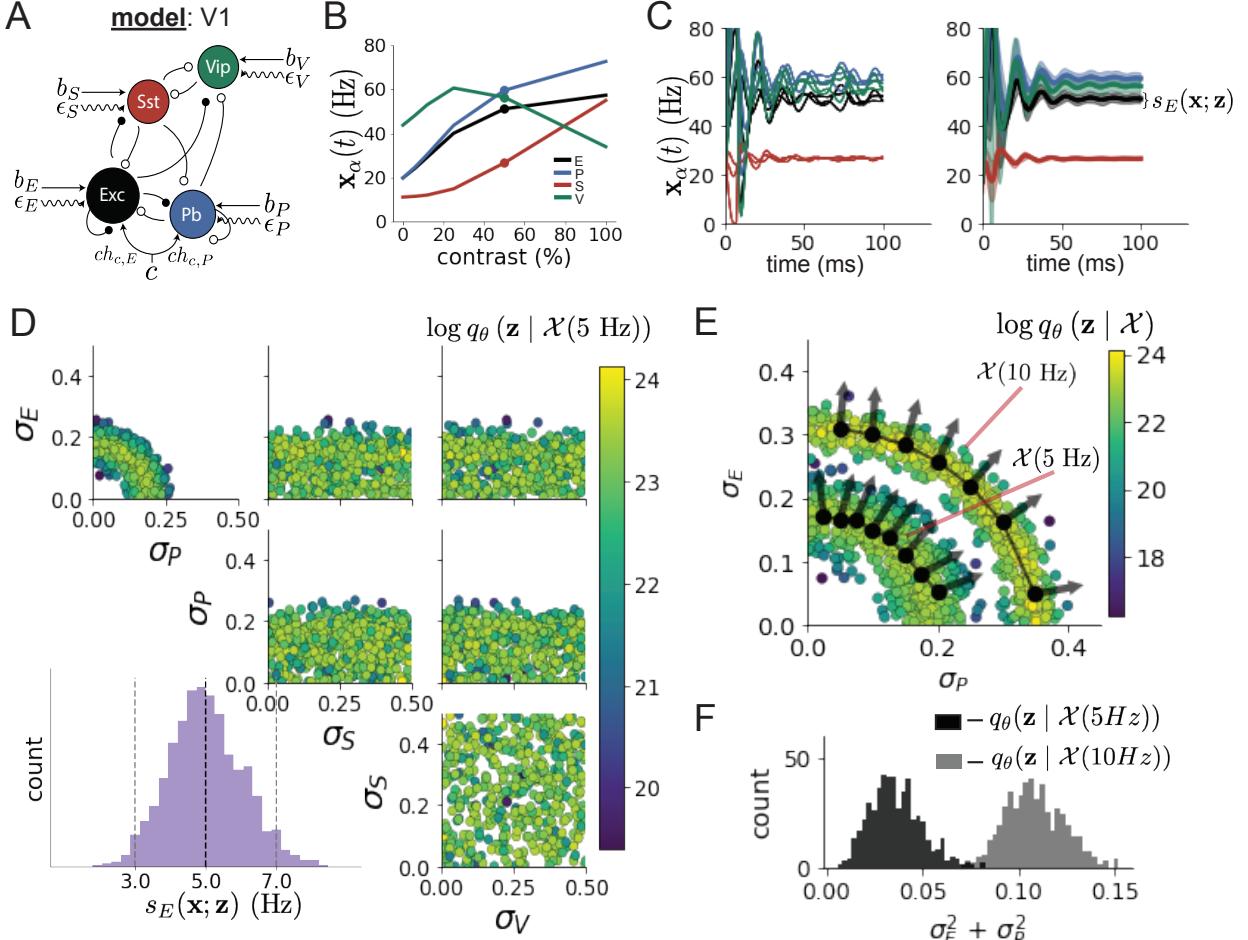


Figure 3: Emergent property inference in the stochastic stabilized supralinear network (SSSN) **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). 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.** Steady-state responses of the SSN model (deterministic, $\sigma = \mathbf{0}$) to varying contrasts. The response at 50% contrast (dots) is the focus of our analysis. **C.** Transient network responses of the SSSN model at 50 % contrast. (Left) Traces are independent trials with varying initialization $\mathbf{x}(0)$ and noise realization. (Right) Mean (solid line) and standard deviation (shading) of responses. **D.** EPI 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 parameter distribution $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ and $q_\theta(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$.

286 This model is the stochastic stabilized supralinear network (SSSN) [71] generalized to have mul-
 287 tiple inhibitory neuron types, and introduces stochasticity to previous four neuron-type models
 288 of V1 [54]. Stochasticity and inhibitory multiplicity introduce substantial complexity to mathe-
 289 matical derivations (see Section 5.2.4) motivating the treatment of this model with EPI. Here, we
 290 consider fixed weights W and input \mathbf{h} [55] (Fig. 3B), and study the effect of input variability
 291 $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on excitatory variability at 50% contrast.

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

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

293 where $s_E(\mathbf{x}; \mathbf{z})$ is the standard deviation of the stochastic E-population response about its steady
 294 state (Fig. 3C).

295 We ran EPI to obtain parameter distribution $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ producing E-population variability
 296 around 5 Hz (Fig. 3D). From the marginal distribution of σ_E and σ_P (Fig. 3D, top-left), we can see
 297 that $s_E(\mathbf{x}; \mathbf{z})$ is sensitive to various combinations of σ_E and σ_P . Alternatively, both σ_S and σ_V are
 298 degenerate with respect to $s_E(\mathbf{x}; \mathbf{z})$ evidenced by the high variability in those dimensions (Fig. 3D,
 299 bottom-right). Together, these observations imply a curved path of parametric degeneracy with
 300 respect to $s_E(\mathbf{x}; \mathbf{z})$ of 5 Hz, which is indicated by the modes along σ_P (Fig. 3E). The dimensions
 301 of sensitivity conferred by EPI and this plain visual structure suggest a quadratic relationship
 302 in the emergent property statistic $s_E(\mathbf{x}; \mathbf{z})$ and parameters \mathbf{z} , which is preserved at a greater
 303 level of variability $\mathcal{X}(10 \text{ Hz})$ (Fig. 3E). Indeed, the sum of squares of σ_E and σ_P is larger in
 304 $q_{\theta}(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$ than $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ (Fig 3F, $p = 0$), while the sum of squares of σ_S and σ_V are
 305 not significantly different in the two EPI distributions (Fig. 15, $p = .402$). The strong compensatory
 306 influence of the E- and P-population input variability on excitatory variability is intriguing, since
 307 this circuit exhibited a paradoxical effect in the P-population (and no other inhibitory types) at
 308 50% contrast (Fig. 15) meaning that the E-population is P-stabilized. Future research may uncover
 309 a link between the populations of stabilizations and compensatory interactions governing excitatory
 310 variability.

311 By defining a clear theoretical question and executing the EPI optimization, we used EPI to reveal
 312 the quadratic relationship between $s_E(\mathbf{x}; \mathbf{z})$ and \mathbf{z} . While this property is ultimately derivable,
 313 we show that with each additional neuron-type population, the formula becomes quite unruly and
 314 likely escapes comprehensible analysis in our case (see Section 5.2.4). This emphasizes the need

315 for streamlined methods for gaining understanding about theoretical models when mathematical
316 analysis becomes prohibitive.

317 **3.5 EPI identifies two regimes of rapid task switching**

318 It has been shown that rats can learn to switch from one behavioral task to the next on randomly
319 interleaved trials [72], and an important question is what types of neural connectivity allow this
320 ability. In this experimental setup, rats were explicitly cued on each trial to either orient towards
321 a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A) task
322 (Fig. 4A). Neural recordings in superior colliculus (SC) exhibited two populations of neurons that
323 represented task context (Pro or Anti). Furthermore, Pro/Anti neurons in each hemisphere were
324 strongly correlated with the animal’s decision [56]. These results motivated a model of SC that is
325 a four-population dynamical system with functionally-defined neuron-types. Here, our goal is to
326 understand how connectivity in this circuit model governs the ability to switch tasks rapidly.

327 In this SC model, there are Pro- and Anti-populations in each hemisphere (left (L) and right
328 (R)) with activity variables $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$. The connectivity of these populations is
329 parameterized by self sW , vertical vW , diagonal dW and horizontal hW connections (Fig. 4B). The
330 input \mathbf{h} is comprised of a positive cue-dependent signal to the Pro or Anti populations, a positive
331 stimulus-dependent input to either the Left or Right populations, and a choice-period input to the
332 entire network (see Section 5.2.5). Model responses are bounded from 0 to 1 as a function ϕ of an
333 internal variable \mathbf{u}

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

334 The model responds to the side with greater Pro neuron activation; e.g. the response is left if
335 $x_{LP} > x_{RP}$ at the end of the trial. Here, we use EPI to determine the network connectivity
336 $\mathbf{z} = [sW, vW, dW, hW]^\top$ that produces rapid task switching.

337 We define the computation of rapid task switching as accurate execution of each task. Inferred
338 models should not exhibit fully random responses (50%), or perfect performance (100%), since
339 perfection is never attained by even the best trained rats. We formulate rapid task switching as an
340 emergent property by stipulating that the average accuracy in the Pro task $p_P(\mathbf{x}; \mathbf{z})$ and Anti task



Figure 4: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. **C.** The EPI inferred distribution of rapid task switching networks. Red and purple stars indicate modes \mathbf{z}^* of each connectivity regime. 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}).

³⁴¹ $p_A(\mathbf{x}; \mathbf{z})$ be 75% with variance 7.5%².

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 75\% \\ 75\% \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 7.5\%^2 \\ 7.5\%^2 \end{bmatrix} \end{aligned} \quad (10)$$

³⁴² The EPI inferred distribution (Fig. 4C) produces task accuracies (Fig. 4C, middle-left) according
³⁴³ to our mathematical definition of rapid task switching (Equation 10). The patterns of connectivity
³⁴⁴ that govern each task accuracy are nonlinear (Fig. 17A-B); they are not captured well by linear
³⁴⁵ correlations (Fig. 17C). For example, there appear to be two regimes of connectivity: the local
³⁴⁶ structure of the EPI distribution changes dramatically after crossing a threshold of sW (Fig. 17A
³⁴⁷ $sW-hW$ marginal distribution). Not only has EPI captured this intricate, nonlinear distribution,
³⁴⁸ we can use the distribution $q_{\theta}(\mathbf{z} | \mathcal{X})$ returned by EPI to understand these two parametric regimes
³⁴⁹ of SC connectivity.

³⁵⁰ To distinguish these two parts of the distribution, we point out that for many fixed values of
³⁵¹ parameter hW , there are two modes in the EPI distribution. Thus, by fixing hW to different
³⁵² values and doing gradient ascent on $\log q_{\theta}(\mathbf{z} | \mathcal{X})$ in the parameter spaces proximal to each mode,
³⁵³ we can identify a set of modes $\mathbf{z}^*(hW_{\text{fixed}}, r)$ for each putative regime $r \in \{1, 2\}$ (see Section 5.2.5).
³⁵⁴ As hW_{fixed} increases, the modes coalesce to intermediate parameters reflecting a transition between
³⁵⁵ the two sets of modes (Fig. 20 top). However, the sensitivity dimensions of these modes \mathbf{v}_1 (refer
³⁵⁶ to Section 3.2), which reflect the structure of the EPI distribution around each mode, are different
³⁵⁷ across putative regime, yet consistent across hW_{fixed} . This categorical difference in sensitivity
³⁵⁸ dimension across the two sets of modes shows that they indeed represent two different regimes of
³⁵⁹ computation in which connectivity governs computation in different ways.

³⁶⁰ To understand how SC connectivity governs computation in each regime, we can examine how
³⁶¹ perturbations along $\mathbf{v}_1(\mathbf{z}^*)$ affect task performance in each regime; we measure task accuracy for
³⁶² connectivity changes in the dimension that rapid task switching is most sensitive. While the
³⁶³ monotonic increase in Pro accuracy with \mathbf{v}_1 perturbation is largely unaffected by regime (Fig. 4D,
³⁶⁴ top-left), we see a stark difference in Anti accuracy: Anti accuracy falls in either direction of \mathbf{v}_1
³⁶⁵ in regime 1, yet monotonically increases along with Pro accuracy in regime 2 (Fig. 4D, top-right).
³⁶⁶ These two rapid task switching pathologies are caused by distinct connectivity changes ($\mathbf{v}_1(\mathbf{z}^*(\cdot, 1))$
³⁶⁷ vs $\mathbf{v}_1(\mathbf{z}^*(\cdot, 2))$) and explain the sharp change in local structure of the EPI distribution.

368 To understand the connectivity mechanisms that distinguish these two regimes, we perturb connec-
369 tivity at each mode in dimensions that have well defined roles in processing for the Pro and Anti
370 tasks. A convenient property of this connectivity parameterization is that there are \mathbf{z} -invariant
371 eigenmodes of connectivity, whose eigenvalues (or degree of amplification) change with \mathbf{z} . These
372 eigenmodes have intuitive roles in processing in each task, and are accordingly named the *all*,
373 *side*, *task*, and *diag* eigenmodes (see Section 5.2.5). Furthermore, the parameter dimension \mathbf{v}_a
374 ($a \in \{\text{all}, \text{side}, \text{task}, \text{and diag}\}$) that increases the eigenvalue of connectivity λ_a is \mathbf{z} -invariant (un-
375 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
376 of each processing mode by perturbing \mathbf{z} along \mathbf{v}_a , we can elicit the differentiating properties of
377 the two regimes.

378 Through these connectivity perturbation analyses, we found that increasing λ_{task} strongly reduced
379 Pro accuracy in regime 1, yet strongly reduced Anti accuracy in regime 2. This suggests that
380 stronger task representations can inhibit both Pro and Anti task performance in different contexts.
381 Furthermore, changing λ_{task} in either direction decreases Anti performance in regime 1, showing
382 that Anti task performance in regime 1 is dependent on a specific level of task representation.
383 We also found that with increasing λ_{diag} , Pro accuracy increased in both regimes, but there were
384 opposite effects on Anti accuracy. In regime 1, stronger amplification of diagonal population pat-
385 terns decreased Anti accuracy, while in regime 2 accuracy increased. These findings give us an
386 understanding of the mechanistic differences in computation enabling rapid task switching in each
387 regime.

388 **3.6 EPI inferred SC connectivities reproduce results from optogenetic inacti-
389 vation experiments**

390 During the delay period of this task, the circuit must prepare to execute the correct task based on
391 the cue input. Experimental results from Duan et al. found that optogenetic inactivation of SC
392 during the delay period consistently decreased performance in the Anti task, but had no effect on
393 the Pro task (Fig. 5A). All network connectivities inferred by EPI exhibited this same effect, when
394 network activities were silenced during the delay period (see Section 5.2.5). Notably, EPI inferred
395 connectivities were only conditioned upon the emergent property of rapid task switching, not on
396 Anti task failure during delay period silencing.

397 Similarities across Pro and Anti trials in choice period responses following delay period inactivation
398 (Fig. 21A) suggested that connectivity patterns inducing greater Pro task accuracy increase error

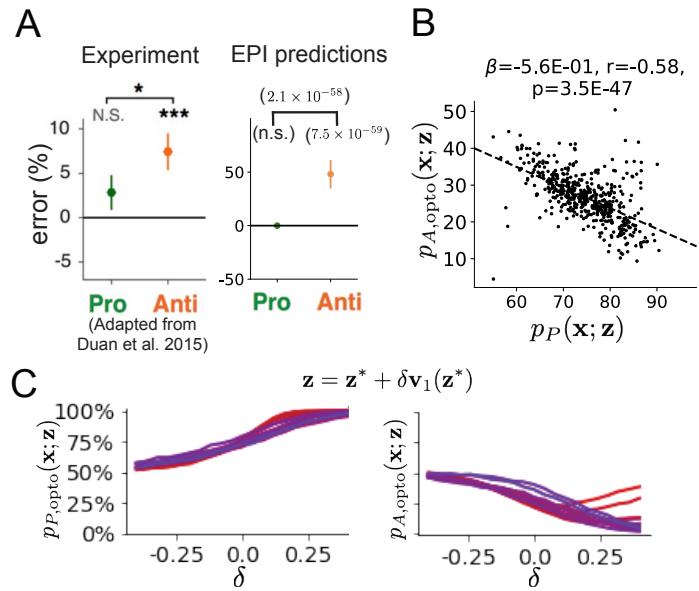


Figure 5: **A.** The EPI distribution predicts experimental results (left) showing no change in the Pro task, but larger error in the Anti task (right). **B.** Accuracy in the Anti task during delay period optogenetic inactivation $p_{A,\text{opto}}$ is strongly anticorrelated with accuracy in the Pro task. **C.** Mean and standard error ($N_{\text{test}} = 25$) of accuracy with respect to perturbation along the sensitivity dimension of each mode \mathbf{z}^* .

399 in delay period inactivated Anti trials (Fig. 5B). The strong anticorrelation between p_P and $p_{A,\text{opto}}$
 400 across EPI inferred connectivities led to the following hypothesis about each connectivity regime:
 401 the sensitivity dimension of each regime decreases $p_{A,\text{opto}}$ irrespective of its effect on p_A , since
 402 both \mathbf{v}_1 and \mathbf{v}_2 increase p_P . Indeed, in regimes 1 and 2 where sensitivity dimensions elicit different
 403 responses in p_A , $p_{A,\text{opto}}$ decreases since the connectivity changes enhancing p_P exacerbate Anti trial
 404 error (Fig. 5C). Thus, the altered state caused by delay period silencing makes the connectivity
 405 governing p_P more influential than the connectivity governing p_A .

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

412 **4 Discussion**

413 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets
414 [37]. Such careful inference procedures are developed for these statistical models allowing precise,
415 quantitative reasoning, which clarifies the way data informs beliefs about the model parameters.
416 However, these statistical models lack resemblance to the underlying biology, making it unclear
417 how to go from the structure revealed by these methods, to the neural mechanisms giving rise
418 to it. In contrast, theoretical neuroscience has focused on careful mechanistic modeling and the
419 production of emergent properties of computation. The careful steps of *i.*) model design and
420 *ii.*) emergent property definition, are followed by *iii.*) practical inference methods resulting in an
421 opaque characterization of the way model parameters govern computation. In this work, we improve
422 upon parameter inference techniques in theoretical neuroscience with emergent property inference,
423 harnessing deep learning towards careful inference in careful models of neural computation (see
424 Section 5.1.1).

425 Specifically, approximate Bayesian computation [73, 74, 38] has been the standard approach to
426 parameter inference in neural circuit models lacking tractable likelihoods. ABC methods do not
427 confer probabilities on accepted parameters, require an acceptance threshold chosen to trade-off
428 inference quality with tractability, do not scale efficiently to high-dimensional parameter spaces, and
429 require independent techniques to analyze sensitivity for local parameter choices [62]. In contrast,
430 EPI allows probability evaluations at any point in parameter space, conditions posteriors on the
431 natural quantification of emergent properties, scales to high dimensional parameter spaces, and
432 naturally admits sensitivity quantification via fast evaluations of the posterior Hessian.

433 Technically, EPI is a maximum entropy method, which learns parameter distributions that are
434 as random as possible given that they produce the emergent property. Conceptually, maximally
435 random distributions given some constraints are useful for understanding parametric sensitivity.
436 This is well understood in Bayesian inference, where maximum entropy is the chosen normative
437 principle. This is emphasized by an innovative formalism unifying top-down maximum entropy
438 normative models with bottom-up statistical models [75]. Indeed, EPI is an adaptive variational
439 inference program, and may be considered to have a Bayesian uniform prior (see Section 5.1.6).

440 Biologically realistic models of neural circuits often prove formidable to analyze for two main rea-
441 sons. A primary challenge is that the number of parameters scales dramatically with the number of
442 neurons, limiting analysis of its parameter space. We see in Section 3.3 that EPI scales seamlessly

443 to high dimensional parameter spaces of RNN connectivities, while maintaining the production
444 of the specified emergent property. EPI strongly outperforms the standard likelihood-free infer-
445 ence technique (SMC-ABC [38]), and a recently developed deep likelihood-free inference technique
446 (SNPE [40]), most likely because of it's ability to leverage the gradient information of the emer-
447 gent property statistics and to adapt it's paramter sampling distribution at every step of gradient
448 descent.

449 A secondary challenge is that the structure of the parametric regimes governing emergent properties
450 is intricate. For example, even in low dimensional circuits, models can support more than one steady
451 state [76] and non-trivial dynamics on strange attractors [77]. With EPI, we use deep probabillity
452 distributions to capture the complex nonlinear parameter distributions governing model behavior.
453 In Section 3.4, we used EPI to reveal a curved parametric manifolds governing curcuit variability
454 in the stochastic stabilized supralinear network, and used hypothesis testing techniques to validate
455 our findings. In Section 3.6, we identified two regimes of SC connectivity resulting in rapid task
456 switching, and found that the full distribution of rapid task switching networks reproduced an
457 experimental result.

458 EPI leverages deep learning technology for neuroscientific inquiry in a categorically different way
459 than approaches focused on training neural networks to execute behavioral tasks [78]. These works
460 focus on examining optimized deep neural networks while considering the objective function, learn-
461 ing rule, and architecture used. This endeavor efficiently obtains sets of parameters that can be
462 reasoned about with respect to such considerations, but lacks the careful probabilistic treatment of
463 parameter inference in EPI. All of these approaches can be used complementarily to enhance the
464 practice of theoretical neuroscience.

465 **Acknowledgements:**

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

472 **Data availability statement:**

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

475 **Code availability statement:**

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

477 **References**

- 478 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
479 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 480 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
481 21(1):25–45, 1998.
- 482 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 483 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in
484 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 485 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
486 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
487 models. *PLoS Comput Biol*, 3(10):e189, 2007.
- 488 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
489 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
490 sion model. *Neuron*, 82(4):809–821, 2014.
- 491 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-
492 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 493 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
494 networks. *Physical review letters*, 61(3):259, 1988.
- 495 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
496 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
497 17(11):4382–4388, 1997.
- 498 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration
499 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 500 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
501 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.

- 502 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
503 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 504 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
505 model. *PLoS computational biology*, 2(7):e94, 2006.
- 506 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
507 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 508 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
509 13(8):1713–1720, 2001.
- 510 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
511 A statistical paradigm for neural spike train decoding applied to position prediction from
512 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
513 7425, 1998.
- 514 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
515 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 516 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.
517 A point process framework for relating neural spiking activity to spiking history, neural
518 ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 519 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations
520 imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–
521 1012, 2006.
- 522 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan
523 Segev. A novel multiple objective optimization framework for constraining conductance-based
524 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 525 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature
526 analysis. *Neural computation*, 19(4):1022–1038, 2007.
- 527 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
528 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of
529 neural population activity. In *Advances in neural information processing systems*, pages
530 1881–1888, 2009.

- 531 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and
532 Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural*
533 *information processing systems*, 24:1350–1358, 2011.
- 534 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In
535 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 536 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-
537 dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,
538 9(3):e1002922, 2013.
- 539 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
540 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
541 *Science*, 349(6244):184–187, 2015.
- 542 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq
543 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in
544 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 545 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
546 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th In-*
547 *ternational Conference on Machine Learning*, 2019.
- 548 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.
549 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.
550 *Nature Communications*, 10(4933), 2019.
- 551 [30] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
552 neural population models through nonlinear embeddings. In *Advances in neural information*
553 *processing systems*, pages 163–171, 2016.
- 554 [31] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-
555 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 556 [32] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-
557 encoders. *bioRxiv*, page 399246, 2018.
- 558 [33] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
559 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R

- 560 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-
561 encoders. *Nature methods*, page 1, 2018.
- 562 [34] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
563 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
564 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 565 [35] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
566 Datta. Composing graphical models with neural networks for structured representations and
567 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 568 [36] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
569 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
570 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural
571 Information Processing Systems*, 2019.
- 572 [37] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
573 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
574 2018.
- 575 [38] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
576 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 577 [39] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
578 A framework for parameter estimation and model selection from experimental data in systems
579 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 580 [40] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
581 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P
582 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
583 dynamics. *bioRxiv*, page 838383, 2019.
- 584 [41] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
585 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on
586 Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 587 [42] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
588 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
589 4248. PMLR, 2020.

- 590 [43] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
591 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 592 [44] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
593 networks. *International Conference on Learning Representations*, 2017.
- 594 [45] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
595 *International Conference on Machine Learning*, 2015.
- 596 [46] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-
597 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,
598 2001.
- 599 [47] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
600 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
601 *Neuron*, 77(5):845–858, 2013.
- 602 [48] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
603 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 604 [49] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dy-
605 namics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–
606 1406, 2014.
- 607 [50] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding
608 and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.
- 609 [51] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
610 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 611 [52] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
612 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 613 [53] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for
614 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,
615 2017.
- 616 [54] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and
617 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,
618 115(3):1399–1409, 2016.

- 619 [55] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
620 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
621 operating regime of cortex. *bioRxiv*, 2020.
- 622 [56] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
623 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
624 routing. *bioRxiv*, page 245613, 2018.
- 625 [57] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
626 2009.
- 627 [58] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural
628 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 629 [59] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in*
630 *neurobiology*, 25:156–163, 2014.
- 631 [60] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*
632 *opinion in neurobiology*, 46:1–6, 2017.
- 633 [61] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H
634 Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex
635 embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966,
636 2018.
- 637 [62] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian com-*
638 *putation*. CRC Press, 2018.
- 639 [63] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
640 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
641 592, 2009.
- 642 [64] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralin-
643 ear network: a unifying circuit motif underlying multi-input integration in sensory cortex.
644 *Neuron*, 85(2):402–417, 2015.
- 645 [65] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
646 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
647 5(10):793, 2004.

- 648 [66] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
649 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-*
650 *robiology*, 71(1):45–61, 2011.
- 651 [67] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-
652 tex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 653 [68] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
654 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
655 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 656 [69] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn.
657 Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.
- 658 [70] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate
659 cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
- 660 [71] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
661 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
662 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 663 [72] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and
664 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,
665 2015.
- 666 [73] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-
667 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 668 [74] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
669 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
670 2003.
- 671 [75] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkacik. Statistical
672 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.
- 673 [76] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give
674 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of
675 Sciences*, 115(13):3464–3469, 2018.

- 676 [77] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emer-
677 gent dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint*
678 *arXiv:1605.04463*, 2016.
- 679 [78] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,
680 2019.
- 681 [79] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
682 1957.
- 683 [80] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
684 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 685 [81] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
686 Edward Teller. Equation of state calculations by fast computing machines. *The journal of*
687 *chemical physics*, 21(6):1087–1092, 1953.
- 688 [82] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
689 1970.
- 690 [83] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
691 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
692 73(2):123–214, 2011.
- 693 [84] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-
694 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
695 820, 2011.
- 696 [85] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based infer-
697 ence. *Proceedings of the National Academy of Sciences*, 2020.
- 698 [86] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
699 free variational inference. In *Advances in Neural Information Processing Systems*, pages
700 5523–5533, 2017.
- 701 [87] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
702 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience*
703 *Meeting (COSYNE), Lisbon, Portugal*, 2019.

- 704 [88] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
705 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience
706 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 707 [89] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
708 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
709 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 710 [90] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-
711 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,*
712 *Germany*, 2019.
- 713 [91] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
714 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
715 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 716 [92] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
717 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 718 [93] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
719 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 720 [94] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
721 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
722 2012.
- 723 [95] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
724 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
725 2018.
- 726 [96] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
727 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 728 [97] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
729 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
730 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
731 1929, 2009.

- 732 [98] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
733 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 734 [99] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
735 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 736 [100] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-
737 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv*
738 preprint arXiv:1912.02762, 2019.
- 739 [101] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
740 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 741 [102] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
742 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-*
743 *tion processing systems*, 29:4743–4751, 2016.
- 744 [103] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Internation-*
745 *al Conference on Learning Representations*, 2015.
- 746 [104] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for
747 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- 748 [105] Emmanuel Klinger, Dennis Rickert, and Jan Hasenauer. pyabc: distributed, likelihood-free
749 inference. *Bioinformatics*, 34(20):3591–3593, 2018.
- 750 [106] David S Greenberg, Marcel Nonnenmacher, and Jakob H Macke. Automatic posterior trans-
751 formation for likelihood-free inference. *International Conference on Machine Learning*, 2019.

752 **5 Methods**

753 **5.1 Emergent property inference (EPI)**

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

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

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

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

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

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

780 Such deep probability distributions embed the posterior distribution in a deep network. Once
 781 optimized, this deep network representation has remarkably useful properties: immediate posterior
 782 sampling, and immediate probability, gradient, and Hessian evaluation at any parameter choice.
 783 Given a choice of model $p(\mathbf{x} \mid \mathbf{z})$ and emergent property of interest \mathcal{X} , $q_{\theta}(\mathbf{z})$ is optimized via
 784 the neural network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep
 785 variational family \mathcal{Q} producing the emergent property \mathcal{X} :

$$q_{\theta}(\mathbf{z} \mid \mathcal{X}) = q_{\theta}^*(\mathbf{z}) = \operatorname{argmax}_{q_{\theta} \in \mathcal{Q}} H(q_{\theta}(\mathbf{z})) \quad (13)$$

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

786 Entropy is chosen as the normative selection principle, since we want the posterior to only contain
 787 structure predicated by the emergent property [79, 80]. This choice of selection principle is also
 788 that of standard Bayesian inference, and we derive an exact relation between EPI and variational
 789 inference (see Section 5.1.5). However, a key difference is that variational inference and other
 790 Bayesian methods do not constrain the predictions of their inferred posteriors. This optimization
 791 is executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [44].

792 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.
 793 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-
 794 ond, we describe the special class of deep probability distributions used in EPI called normalizing
 795 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-
 796 tion 13 (Section 5.1.3). Then, we demonstrate the details of this optimization in a toy example
 797 (Section 5.1.4). Finally, we establish the known relationship between maximum entropy distribu-
 798 tions and exponential families (Section 5.1.5), which is used to explain the relation between EPI
 799 and variational inference (Section 5.1.6).

800 5.1.1 Related approaches

801 When Bayesian inference problems lack conjugacy, scientists use approximate inference methods
 802 like variational inference (VI) [43] and Markov chain Monte Carlo (MCMC) [81, 82]. After opti-
 803 mization, variational methods return a parameterized posterior distribution, which we can analyze.
 804 Also, the variational approximating distribution class is often chosen such that it permits fast
 805 sampling. In contrast MCMC methods only produce samples from the approximated posterior dis-
 806 tribution. No parameterized distribution is estimated, and additional samples are always generated
 807 with the same sampling complexity. Inference in models defined by systems of differential has been

808 demonstrated with MCMC [83], although this approach requires tractable likelihoods. Advances
809 have leveraged structure in stochastic differential equation models to improve likelihood
810 approximations, thus expanding the domain of applicable models [84].

811 Likelihood-free (or “simulation-based”) inference (LFI) [85] is model parameter inference in the
812 absence of a tractable likelihood function. The most prevalent approach to LFI is approximate
813 Bayesian computation [73], in which satisfactory parameter samples are kept from random prior
814 sampling according to a rejection heuristic. The obtained set of parameters do not have a prob-
815 abilities, and further insight about the model must be gained from examination of the parameter
816 set and their generated activity. Methodological advances to ABC methods have come through
817 the use of Markov chain Monte Carlo (MCMC-ABC) [74] and sequential Monte Carlo (SMC-ABC)
818 [38] sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still
819 struggles to scale in dimensionality (cf. Fig. 2). Furthermore, once a parameter set has been
820 obtained by SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again
821 with a new population of initialized particles to obtain additional samples.

822 For scientific model analysis, we seek a posterior distribution exhibiting the properties of a well-
823 chosen variational approximation: a parametric form conferring analytic calculations, and trivial
824 sampling time. For this reason, ABC and MCMC techniques are unattractive, since they only
825 produce a set of parameter samples and have unchanging sampling rate. EPI executes likelihood-
826 free inference using the MEFN [44] algorithm using a deep variational posterior approximation.
827 The deep neural network of EPI defines the parametric form of the posterior approximation. Fur-
828 thermore, the EPI distribution is constrained to produce an emergent property. In other words,
829 the summary statistics of the posterior predictive distribution are fixed to have certain first and
830 second moments. EPI optimization is enabled using stochastic gradient techniques in the spirit
831 of likelihood-free variational inference [86]. The analytic relationship between EPI and variational
832 inference is explained in Secton 5.1.6.

833 We note that, during our preparation and early presentation of this work [87, 88], another work
834 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
835 circuits ([89, 90, 40]). We are encouraged by this general problem being recognized by others in the
836 community, and we emphasize that these works offer complementary neuroscientific contributions
837 (different theoretical models of focus) and use different technical methodologies (ours is built on
838 our prior work [44], theirs similarly [91]).

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

840 recently developed LFI methods in which two neural networks are used for posterior inference.
841 This first neural network is a normalizing flow used to estimate the posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE)
842 or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural likelihood (SNL [41])). A recent advance uses an
843 unconstrained neural network to estimate the likelihood ratio (sequential neural ratio estimation
844 (SNRE [42])). In SNL and SNRE, MCMC sampling techniques are used to obtain samples from
845 the approximated posterior. This contrasts with EPI and SNPE, which afford a normalizing flow
846 approximation to the posterior, which facilitates immediate measurements of sample probability,
847 gradient, or Hessian for system analysis. The second neural network in this sequential class of
848 methods is the amortizer. This network maps data \mathbf{x} (or statistics $f(\mathbf{x}; \mathbf{z})$ or model parameters \mathbf{z})
849 to the weights and biases of the first neural network. These methods are optimized on a conditional
850 density (or ratio) estimation objective on a sequentially adapting finite sample-based approximation
851 to the posterior.

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

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

870 **5.1.2 Normalizing flows**

871 Deep probability distributions are comprised of multiple layers of fully connected neural networks
 872 (Equation). When each neural network layer is restricted to be a bijective function, the sample
 873 density can be calculated using the change of variables formula at each layer of the network. For
 874 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

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

875 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 876 restricting our layers to normalizing flows [45, 100] – bijective functions with fast log determinant
 877 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log
 878 probability calculation confers efficient optimization of the maximum entropy objective (see Section
 879 5.1.3). We use the Real NVP [52] normalizing flow class, because its coupling architecture confers
 880 both fast sampling (forward) and fast log probability evaluation (backward). Fast probability
 881 evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout
 882 parameter space. Glow permutations were used in between coupling stages [101]. This is in contrast
 883 to autoregressive architectures [53, 102], in which only forward or backward passes are efficient. In
 884 this work, normalizing flows are used as flexible posterior approximations $q_{\boldsymbol{\theta}}(\mathbf{z})$ having weights and
 885 biases $\boldsymbol{\theta}$. We specify the architecture used in each application by the number of Real-NVP affine
 886 coupling stages, and the number of neural network layers and units per layer of the conditioning
 887 functions.

888 **5.1.3 Augmented Lagrangian optimization**

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

$$L(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{opt}}, c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top R(\boldsymbol{\theta}) + \frac{c}{2} \|R(\boldsymbol{\theta})\|^2 \quad (15)$$

891 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
 892 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.
 893 The sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter $\boldsymbol{\mu}_{\text{opt}}$ are determined by the means $\boldsymbol{\mu}$ and
 894 variances $\boldsymbol{\sigma}^2$ of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 13. Specifically, $T(\mathbf{x}; \mathbf{z})$ is
 895 a concatenation of the first and second moments, $\boldsymbol{\mu}_{\text{opt}}$ is a concatenation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ (see section
 896 5.1.5), and the Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of exponential

897 families (see Section 5.1.6). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution are optimized
898 according to Equation 15 using the Adam optimizer with learning rate 10^{-3} [103].

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

$$H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \int -q_{\boldsymbol{\theta}}(\mathbf{z}) \log(q_{\boldsymbol{\theta}}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [-\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (16)$$

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

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

904 The lagrangian parameters $\boldsymbol{\eta}_{\text{opt}}$ are initialized to zero and adapted following each augmented
905 Lagrangian epoch, which is a period of optimization with fixed $(\boldsymbol{\eta}_{\text{opt}}, c)$ for a given number of
906 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
907 after each epoch based on constraint error reduction. The penalty coefficient is updated based
908 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
909 $\mathbb{E}[|R(\boldsymbol{\theta}_{k+1})|] > \gamma \mathbb{E}[|R(\boldsymbol{\theta}_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
910 other update rule is $\boldsymbol{\eta}_{\text{opt},k+1} = \boldsymbol{\eta}_{\text{opt},k} + c_k \frac{1}{n} \sum_{i=1}^n (T(\mathbf{x}^{(i)}) - \boldsymbol{\mu}_{\text{opt}})$ given a batch size n . Throughout
911 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
912 according to application.

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

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

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

934 When assessing the suitability of EPI for a particular modeling question, there are some important
 935 technical considerations. First and foremost, as in any optimization problem, the defined emergent
 936 property should always be appropriately conditioned (constraints should not have wildly different
 937 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution
 938 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-
 939 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,
 940 one should consider the computational cost of the gradient calculations. In the best circumstance,
 941 there is a simple, closed form expression (e.g. Section 5.2.2) for the emergent property statistic
 942 given the model parameters. On the other end of the spectrum, many forward simulation iterations
 943 may be required before a high quality measurement of the emergent property statistic is available
 944 (e.g. Section 5.2.1). In such cases, backpropagating gradients through the SDE evolution will be
 945 expensive.

946 5.1.4 Example: 2D LDS

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

$$\tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (18)$$

949 with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (19)$$

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

954 alternatively of positive imaginary component when the eigenvalues are complex conjugate pairs.
 955 To learn the distribution of real entries of A that produce a band of oscillating systems around
 956 1Hz, we formalized this emergent property as $\text{real}(\lambda_1)$ having mean zero with variance 0.25^2 , and
 957 the oscillation frequency $2\pi\text{imag}(\lambda_1)$ having mean $\omega = 1$ Hz with variance $(0.1\text{Hz})^2$:

$$\mathbb{E}[T(\mathbf{x})] \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}. \quad (20)$$

958

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

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (21)$$

962 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
 963 does not require simulation, we cannot derive the distribution q_{θ}^* directly. This fact is due to the
 964 formally hard problem of the backward mapping: finding the natural parameters η from the mean
 965 parameters $\boldsymbol{\mu}$ of an exponential family distribution [92]. Instead, we used EPI to approximate this
 966 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two
 967 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped
 968 onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

969 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the
 970 parameter distribution. To validate our method, we analytically derived the contours of the prob-
 971 ability density from the emergent property statistics and values. In the a_1 - a_4 plane, the black
 972 line at $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$,
 973 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
 974 of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired
 975 statistical constraints and model degeneracy in the sum of a_1 and a_4 . Intuitively, the parameters
 976 equivalent with respect to emergent property statistic $\text{real}(\lambda_1)$ have similar log densities.

977 To explain the bimodality of the EPI distribution, we examined the imaginary component of λ_1 .

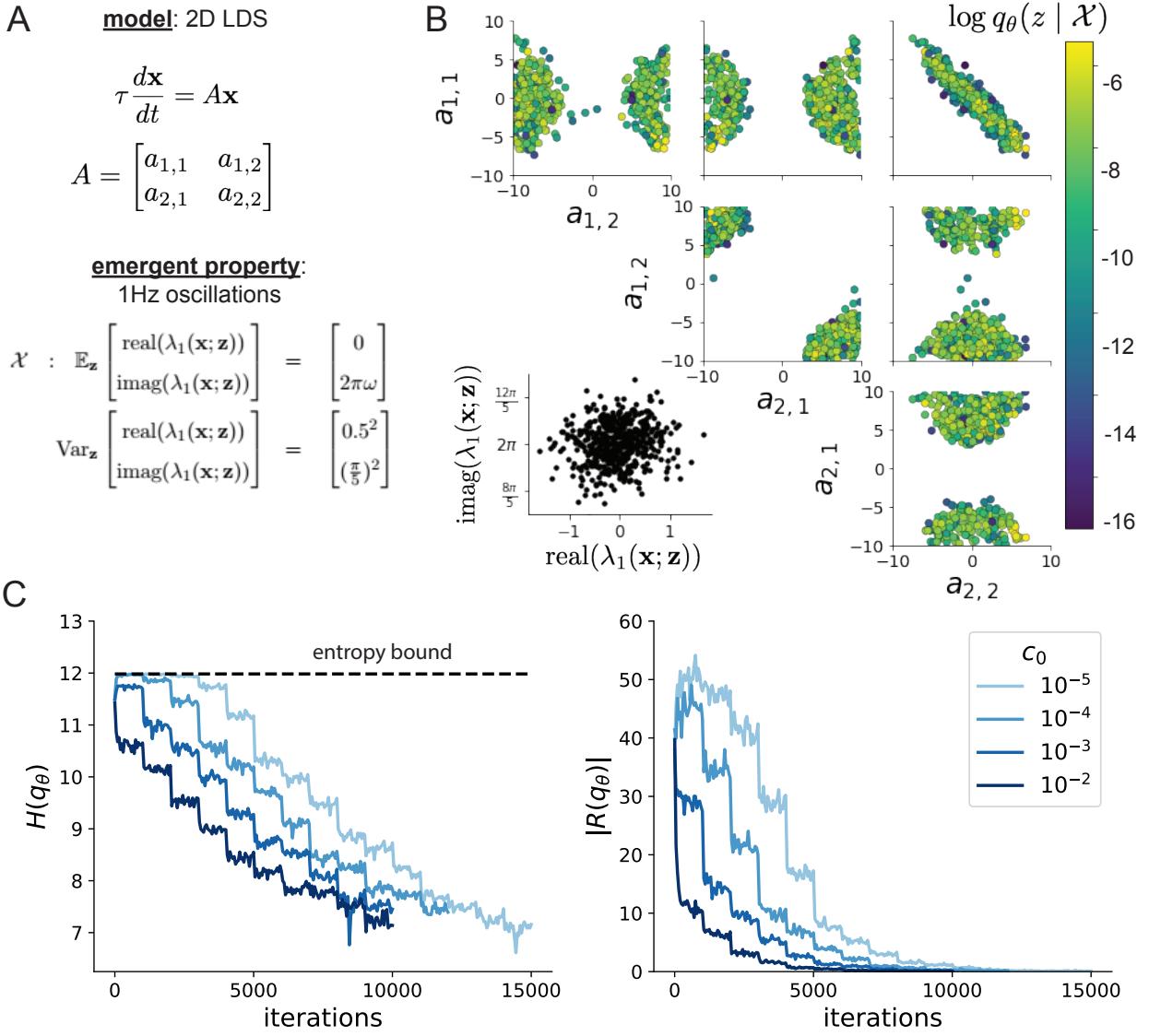


Figure 6: (LDS1): **A.** Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. **B.** The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. 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.

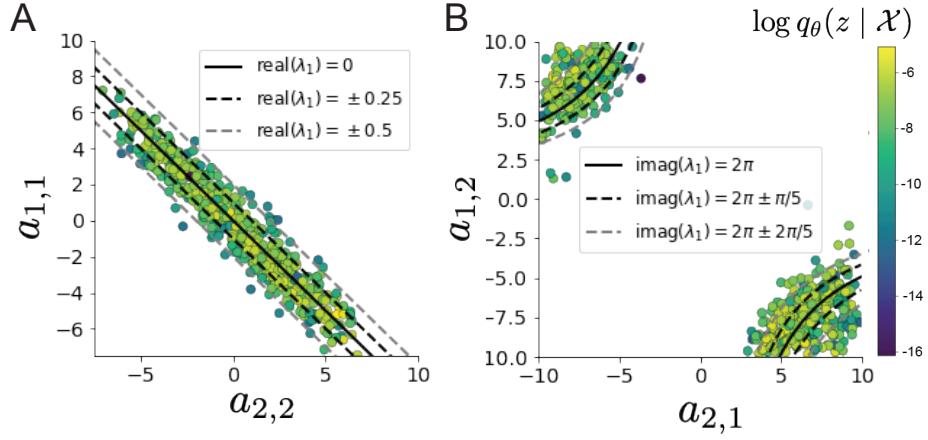


Figure 7: (LDS2): **A.** Probability contours in the a_1 - a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. **B.** Probability contours in the a_2 - a_3 plane were derived from the emergent property statistic of oscillation frequency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

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4 - a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (22)$$

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

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \quad (23)$$

981 Since we constrained $\mathbb{E}_{\mathbf{z} \sim q_\theta} [\text{imag}(\lambda)] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation
982 $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_\theta} [|a_1a_4 - E_{q_\theta}[a_1a_4]|]$,
983 we show the contours as $a_1a_4 = 0$ (black), $a_1a_4 = -\sigma_{1,4}$ (black dotted), and $a_1a_4 = -2\sigma_{1,4}$ (grey
984 dotted). This validates the curved structure of the inferred distribution learned through EPI. We
985 took steps in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive
986 values a_1a_4 in the learned distribution. Subtler combinations of model and emergent property will
987 have more complexity, further motivating the use of EPI for understanding these systems. As we
988 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz
989 (Fig. S3).

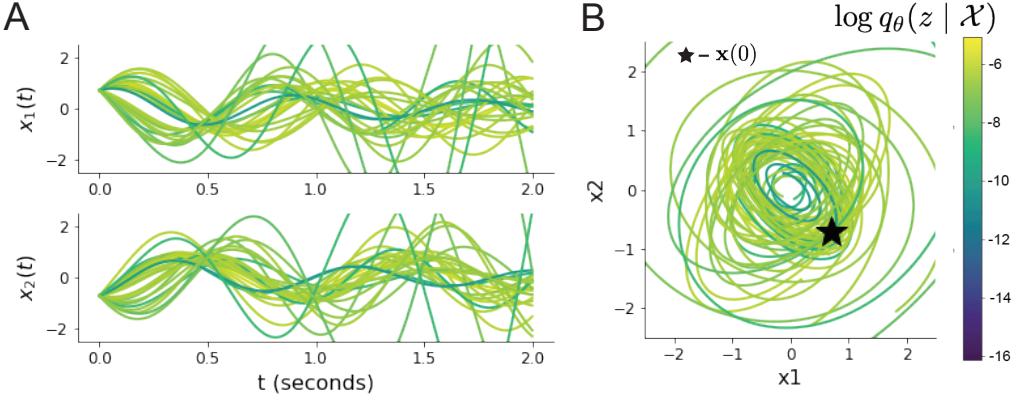


Figure 8: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_\theta(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. **A.** Each dimension of the simulated trajectories throughout time. **B.** The simulated trajectories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

990 5.1.5 Maximum entropy distributions and exponential families

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

$$p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \quad (24)$$

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

993 will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \quad (25)$$

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

996 In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equation
 997 4

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (26)$$

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

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

999 As a general maximum entropy distribution (Equation 24), the sufficient statistics vector contains

1000 both first and second order moments of $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (28)$$

1001 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (29)$$

1002 **5.1.6 EPI as variational inference**

1003 In Bayesian inference a prior belief about model parameters \mathbf{z} is stated in a prior distribution $p(\mathbf{z})$,
 1004 and the statistical model capturing the effect of \mathbf{z} on observed data points \mathbf{x} is formalized in the
 1005 likelihood distribution $p(\mathbf{x} | \mathbf{z})$. In Bayesian inference, we obtain a posterior distribution $p(z | \mathbf{x})$,
 1006 which captures how the data inform our knowledge of model parameters using Bayes' rule:

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \quad (30)$$

1007 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
 1008 However, conjugacy is rare in practice, and alternative methods, such as variational inference [104],
 1009 are utilized.

1010 In variational inference, a posterior approximation $q_{\boldsymbol{\theta}}^*$ is chosen from within some variational family
 1011 \mathcal{Q}

$$q_{\boldsymbol{\theta}}^*(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})). \quad (31)$$

1012 The KL divergence can be written in terms of entropy of the variational approximation:

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

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

1014 Since the marginal distribution of the data $p(\mathbf{x})$ (or ‘evidence’) is independent of $\boldsymbol{\theta}$, variational
 1015 inference is executed by optimizing the remaining expression. This is usually framed as maximizing
 1016 the evidence lower bound (ELBO)

$$\underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}} || p(\mathbf{z} | \mathbf{x})) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmax}} H(q_{\boldsymbol{\theta}}) + \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z}))]. \quad (34)$$

1017 Now, consider the setting where we have chosen a uniform prior, and stipulate a mean-field gaussian
 1018 likelihood on a chosen statistic of the data $f(\mathbf{x}; \mathbf{z})$

$$p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) | \boldsymbol{\mu}_f, \Sigma_f), \quad (35)$$

1019 where $\Sigma_f = \text{diag}(\boldsymbol{\sigma}_f^2)$. The log likelihood is then proportional to a dot product of the natural
 1020 parameter of this mean-field gaussian distribution and the first and second moment statistics.

$$\log p(\mathbf{x} | \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \quad (36)$$

1021 where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\boldsymbol{\mu}_f}{\sigma_f^2} \\ \frac{-1}{2\sigma_f^2} \end{bmatrix}, \text{ and} \quad (37)$$

$$1022 \quad T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \quad (38)$$

1023 The variational objective is then

$$\underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta) + \boldsymbol{\eta}_f^\top \mathbb{E}_{\mathbf{z} \sim q_\theta} [T(\mathbf{x}; \mathbf{z})] \quad (39)$$

1024 Comparing this to the Lagrangian objective (without augmentation) of EPI, we see they are the
 1025 same

$$\begin{aligned} q_\theta^*(\mathbf{z}) &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top (\mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}_{\text{opt}}) \\ &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})]. \end{aligned} \quad (40)$$

1026 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
 1027 (Equation 28). Thus, EPI is implicitly executing variational inference with a uniform prior and a
 1028 mean-field gaussian likelihood on the emergent property statistics. The data \mathbf{x} used by this implicit
 1029 variational inference program would be that generated by the adapting variational approximation
 1030 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{z})q_\theta(\mathbf{z})$, and the likelihood parameters $\boldsymbol{\eta}_f$ of EPI optimization epoch k are predicated
 1031 by $\boldsymbol{\eta}_{\text{opt}, k}$. However, in EPI we have not specified a prior distribution, or collected data, which can
 1032 inform us about model parameters. Instead we have a mathematical specification of an emergent
 1033 property, which the model must produce, and a maximum entropy selection principle. Accordingly,
 1034 we replace the notation of $p(\mathbf{z} | \mathbf{x})$ with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys
 1035 emergent property \mathcal{X} (see Section 5.1).

1036 5.2 Theoretical models

1037 In this study, we used emergent property inference to examine several models relevant to theoretical
 1038 neuroscience. Here, we provide the details of each model and the related analyses.

1039 **5.2.1 Stomatogastric ganglion**

1040 We analyze how the parameters $\mathbf{z} = [g_{el}, g_{synA}]$ govern the emergent phenomena of intermediate
 1041 hub frequency in a model of the stomatogastric ganglion (STG) [47] shown in Figure 1A with
 1042 activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$, using the same hyperparameter choices as Gutierrez et al.
 1043 Each neuron's membrane potential $x_\alpha(t)$ for $\alpha \in \{f1, f2, hub, s1, s2\}$ is the solution of the following
 1044 stochastic differential equation:

$$C_m \frac{dx_\alpha}{dt} = -[h_{leak}(\mathbf{x}; \mathbf{z}) + h_{Ca}(\mathbf{x}; \mathbf{z}) + h_K(\mathbf{x}; \mathbf{z}) + h_{hyp}(\mathbf{x}; \mathbf{z}) + h_{elec}(\mathbf{x}; \mathbf{z}) + h_{syn}(\mathbf{x}; \mathbf{z})] + dB. \quad (41)$$

1045 The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization,
 1046 electrical and synaptic currents as well as gaussian noise dB . Each current component is a function
 1047 of all membrane potentials and the conductance parameters \mathbf{z} .

1048 The capacitance of the cell membrane was set to $C_m = 1nF$. Specifically, the currents are the
 1049 difference in the neuron's membrane potential and that current type's reversal potential multiplied
 1050 by a conductance:

$$h_{leak}(\mathbf{x}; \mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (42)$$

$$h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{post} - x_\alpha^{pre}) \quad (43)$$

$$h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (44)$$

$$h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (45)$$

$$h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (46)$$

$$h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (47)$$

1051 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 1052 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$, g_{Ca} ,
 1053 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 1054 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
 1055 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
 1056 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

1057 Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating
 1058 dynamics dependent on steady-state gating variables M_∞ , N_∞ and H_∞ , respectively:

$$M_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_1}{v_2} \right) \right) \quad (48)$$

$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (49)$$

$$N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_3}{v_4} \right) \right) \quad (50)$$

$$\lambda_N = \phi_N \cosh \left(\frac{x_\alpha - v_3}{2v_4} \right) \quad (51)$$

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (52)$$

$$H_\infty = \frac{1}{1 + \exp \left(\frac{x_\alpha + v_5}{v_6} \right)} \quad (53)$$

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp \left(\frac{-x_\alpha + v_7}{v_8} \right)} \right). \quad (54)$$

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

Finally, there is a synaptic gating variable as well:

$$S_\infty = \frac{1}{1 + \exp \left(\frac{v_{th} - x_\alpha}{v_9} \right)}. \quad (55)$$

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

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

$$\Phi = [0.0, 0.01, \dots, 1.0]^\top .. \quad (56)$$

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

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

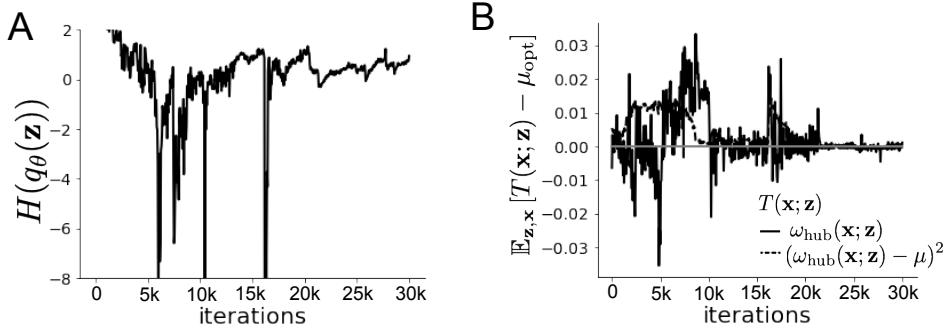


Figure 9: (STG1): EPI optimization of the STG model producing network syncing. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

1086 $\{f_1, f_2, \text{hub}, s_1, s_2\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta|X_\alpha| \odot i), \quad (57)$$

1087 where $i = [0, 1, \dots, 100]$. The frequency is then calculated as

$$\omega_\alpha = 0.01\psi_\alpha \text{Hz}. \quad (58)$$

1088 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
1089 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
1090 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be $(0.025\text{Hz})^2$ to
1091 capture variation in frequency between 0.5Hz and 0.6Hz (Equation 4). As a maximum entropy dis-
1092 tribution, $T(\mathbf{x}; \mathbf{z})$ is comprised of both these first and second moments of the hub neuron frequency
1093 (as in Equations 28 and 29)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \quad (59)$$

1094

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \quad (60)$$

1095 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
1096 epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

1097 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-
1098 layer neural networks of 25 units per layer. The normalizing flow architecture mapped $z_0 \sim \mathcal{N}(\mathbf{0}, I)$

1099 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
 1100 returned by a preliminary ABC search. We did not include $g_{\text{synA}} < 0.01$, for numerical stability.
 1101 EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an
 1102 augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The distribution
 1103 shown is that of the architecture converging with criteria $N_{\text{test}} = 100$ at greatest entropy across
 1104 random seeds.

1105 We calculated the Hessian at the mode of the inferred EPI distribution. The Hessian of a probability
 1106 model is the second order gradient of the log probability density $\log q_{\boldsymbol{\theta}}(\mathbf{z})$ with respect to the
 1107 parameters \mathbf{z} : $\frac{\partial^2 \log q_{\boldsymbol{\theta}}(\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^\top}$. With EPI, we can examine the Hessian, which is analytically available
 1108 throughout distribution, to indicate the dimensions of parameter space that are sensitive (strongly
 1109 negative eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the
 1110 emergent property produced. In Figure 1D, the eigenvectors of the Hessian v_1 (solid) and v_2
 1111 (dashed) are shown evaluated at the mode of the distribution. The length of the arrows is inversely
 1112 proportional to the square root of absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$.
 1113 Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D parameter
 1114 space are chosen arbitrarily.

1115 5.2.2 Scaling EPI for stable amplification in RNNs

1116 We examined the scaling properties of EPI by learning connectivities of RNNs of increasing size
 1117 that exhibit stable amplification. Rank-2 RNN connectivity was modeled as $W = UV^\top$, where
 1118 $U = [\mathbf{u}_1 \ \mathbf{u}_2] + g\chi^{(W)}$, $V = [\mathbf{v}_1 \ \mathbf{v}_2] + g\chi^{(V)}$, and $\chi_{i,j}^{(W)}, \chi_{i,j}^{(V)} \sim \mathcal{N}(0, 1)$. This RNN model has
 1119 dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}. \quad (61)$$

1120 In this analysis, we inferred connectivity parameterizations $\mathbf{z} = [\mathbf{u}_1^\top, \mathbf{u}_2^\top, \mathbf{v}_1^\top, \mathbf{v}_2^\top]^\top \in [-1, 1]^{(4N)}$
 1121 that produced stable amplification using EPI, SMC-ABC [38], and SNPE [40] (see Section Related
 1122 Methods).

1123 For this RNN model to be stable, all real eigenvalues of W must be less than 1: $\text{real}(\lambda_1) < 1$,
 1124 where λ_1 denotes the greatest real eigenvalue of W . For a stable RNN to amplify at least one input
 1125 pattern, the symmetric connectivity $W^s = \frac{W+W^\top}{2}$ must have an eigenvalue greater than 1: $\lambda_1^s > 1$,
 1126 where λ^s is the maximum eigenvalue of W^s . These two conditions are necessary and sufficient for
 1127 stable amplification in RNNs [58]. We defined the emergent property of stable amplification with
 1128 means of these eigenvalues (0.5 and 1.5, respectively) that satisfy these conditions. To complete

1129 the emergent property definition, we chose variances (0.25^2) about those means such that samples
 1130 rarely violate the eigenvalue constraints. In terms of the EPI optimization variables, this is written
 1131 as

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \lambda_1^s(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0.5)^2 \\ (\lambda_1^s(\mathbf{x}; \mathbf{z}) - 1.5)^2 \end{bmatrix}, \quad (62)$$

1132

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.25^2 \\ 0.25^2 \end{bmatrix}. \quad (63)$$

1133 Gradients of maximum eigenvalues of Hermitian matrices like W^s are available with modern auto-
 1134 matic differentiation tools. To differentiate through the $\text{real}(\lambda_1)$, we solved the following equation
 1135 for eigenvalues of rank-2 matrices using the rank reduced matrix $W^r = V^\top U$

$$\lambda_{\pm} = \frac{\text{Tr}(W^r) \pm \sqrt{\text{Tr}(W^r)^2 - 4\text{Det}(W^r)}}{2}. \quad (64)$$

1136 For EPI in Fig. 2, we used a real NVP architecture with three coupling layers of affine transfor-
 1137 mations parameterized by two-layer neural networks of 100 units per layer. The initial distribution
 1138 was a standard isotropic gaussian $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to the support of $\mathbf{z}_i \in [-1, 1]$. We used
 1139 an augmented Lagrangian coefficient of $c_0 = 10^3$, a batch size $n = 200$, $\beta = 4$, and chose to use
 1140 500 iterations per augmented Lagrangian epoch and emergent property constraint convergence was
 1141 evaluated at $N_{\text{test}} = 200$ (Fig. 2B blue line, and Fig. 2C-D blue).

1142 We compared EPI to two alternative likelihood-free inference (LFI) techniques, since the likelihood
 1143 of these eigenvalues given \mathbf{z} is not available. Approximate Bayesian computation (ABC) [73] is a
 1144 rejection sampling technique for obtaining sets of parameters \mathbf{z} that produce activity \mathbf{x} close to some
 1145 observed data \mathbf{x}_0 . Sequential Monte Carlo approximate Bayesian computation (SMC-ABC) is the
 1146 state-of-the-art ABC method, which leverages SMC techniques to improve sampling speed. We ran
 1147 SMC-ABC with the pyABC package [105] to infer RNNs with stable amplification: connectivities
 1148 having eigenvalues within an ϵ -defined l_2 distance of

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}. \quad (65)$$

1149 SMC-ABC was run with a uniform prior over $\mathbf{z} \in [-1, 1]^{(4N)}$, a population size of 1,000 particles
 1150 with simulations parallelized over 32 cores, and a multivariate normal transition model.

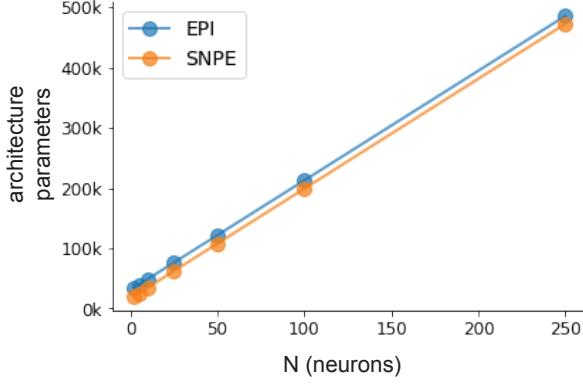


Figure 10: (RNN1): Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size (N).

1151 SNPE, the next LFI approach in our comparison, is far more similar to EPI. Like EPI, SNPE
 1152 treats parameters in mechanistic models with deep probability distributions, yet the two learning
 1153 algorithms are categorically different. SNPE uses a two-network architecture to approximate the
 1154 posterior distribution of the model conditioned on observed data \mathbf{x}_0 . The amortizing network maps
 1155 observations \mathbf{x}_i to the parameters of the deep probability distribution. The weights and biases of the
 1156 parameter network are optimized by sequentially augmenting the training data with additional pairs
 1157 $(\mathbf{z}_i, \mathbf{x}_i)$ based on the most recent posterior approximation. This sequential procedure is important
 1158 to get training data \mathbf{z}_i to be closer to the true posterior, and \mathbf{x}_i to be closer to the observed data.
 1159 For the deep probability distribution architecture, we chose a masked autoregressive flow with affine
 1160 couplings (the default choice), three transforms, 50 hidden units, and a normalizing flow mapping
 1161 to the support as in EPI. This architectural choice closely tracked the size of the architecture used
 1162 by EPI (Fig. 10). As in SMC-ABC, we ran SNPE with $\mathbf{x}_0 = \mu$. All SNPE optimizations were
 1163 run for a limit of 1.5 days on a Tesla V100 GPU, or until two consecutive rounds resulted in a
 1164 validation log probability lower than the maximum observed for that random seed.

1165 To clarify the difference in objectives of EPI and SNPE, we show their results on RNN models
 1166 with different numbers of neurons N and random strength g . The parameters inferred by EPI
 1167 consistently produces the same mean and variance of $\text{real}(\lambda_1)$ and λ_1^s , while those inferred by
 1168 SNPE change according to the model definition (Fig. 11A). For $N = 2$ and $g = 0.01$, the SNPE
 1169 posterior has greater concentration in eigenvalues around \mathbf{x}_0 than at $g = 0.1$, where the model has
 1170 greater randomness (Fig. 11B top, orange). At both levels of g when $N = 2$, the posterior of SNPE
 1171 has lower entropy than EPI at convergence (Fig. 11B top). However at $N = 10$, SNPE results in

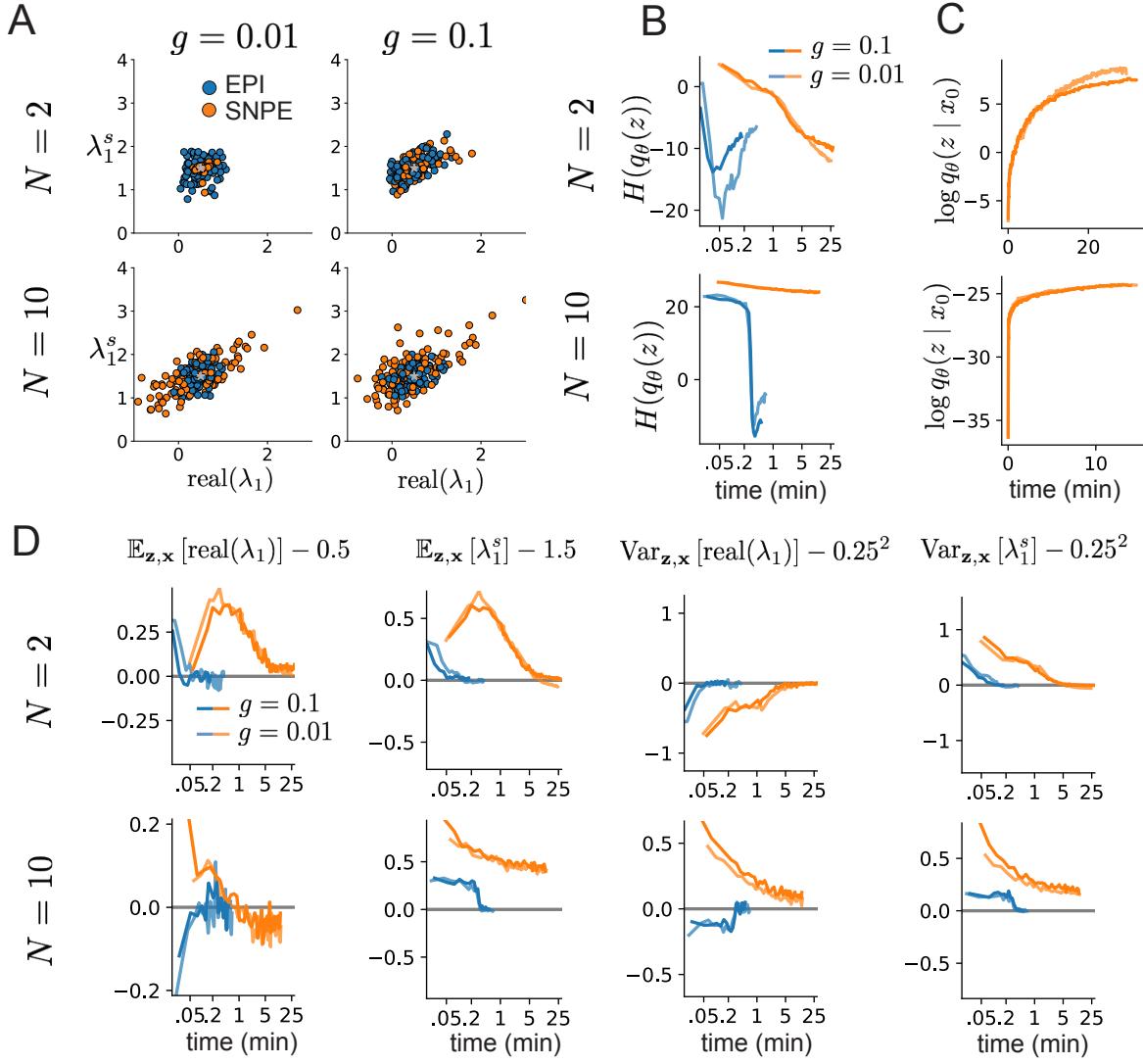


Figure 11: (RNN2): Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A.** Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with $N = 2$ (top), $N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

1172 a predictive distribution of more widely dispersed eigenvalues (Fig. 11A bottom), and an inferred
1173 posterior with greater entropy than EPI (Fig. 11B bottom). We highlight these differences not
1174 to focus on an insightful trend, but to emphasize that these methods optimize different objectives
1175 with different implications.

1176 Note that SNPE converges when it's validation log probability has saturated after several rounds
1177 of optimization (Fig. 11C), and that EPI converges after several epochs of its own optimization
1178 to enforce the emergent property constraints (Fig. 11D blue). Importantly, as SNPE optimizes
1179 its posterior approximation, the predictive means change, and at convergence may be different
1180 than \mathbf{x}_0 (Fig. 11D orange, left). It is sensible to assume that predictions of a well-approximated
1181 SNPE posterior should closely reflect the data on average (especially given a uniform prior and
1182 a low degree of stochasticity), however this is not a given. Furthermore, no aspect of the SNPE
1183 optimization controls the variance of the predictions (Fig. 11D orange, right).

1184 To compare the efficiency of these algorithms for inferring RNN connectivity distributions producing
1185 stable amplification, we develop a convergence criteria that can be used across methods. While EPI
1186 has its own hypothesis testing convergence criteria for the emergent property, it would not make
1187 sense to use this criteria on SNPE and SMC-ABC which do not constrain the means and variances
1188 of their predictions. Instead, we consider EPI and SNPE to have converged after completing its
1189 most recent optimization epoch (EPI) or round (SNPE) in which the distance

$$d(q_\theta(z)) = |\mathbb{E}_{\mathbf{z}, \mathbf{x}} [f(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}|_2 \quad (66)$$

1190 is less than 0.5. We consider SMC-ABC to have converged once the population produces samples
1191 within the $\epsilon = 0.5$ ball ensuring stable amplification.

1192 When assessing the scalability of SNPE, it is important to check that alternative hyperparameter-
1193 izations could not yield better performance. Key hyperparameters of the SNPE optimization are
1194 the number of simulations per round n_{round} , the number of atoms used in the atomic proposals of
1195 the SNPE-C algorithm [106], and the batch size n . To match EPI, we used a batch size of $n = 200$
1196 for $N \leq 25$, however we found $n = 1,000$ to be helpful for SNPE in higher dimensions. While
1197 $n_{\text{round}} = 1,000$ yielded SNPE convergence for $N \leq 25$, we found that a substantial increase to
1198 $n_{\text{round}} = 25,000$ yielded more consistent convergence at $N = 50$ (Fig. 12A). By increasing n_{round} ,
1199 we also necessarily increase the duration of each round. At $N = 100$, we tried two hyperparameter
1200 modifications. As suggested in [106], we increased n_{atom} by an order of magnitude to improve
1201 gradient quality, but this had little effect on the optimization (much overlap between same random

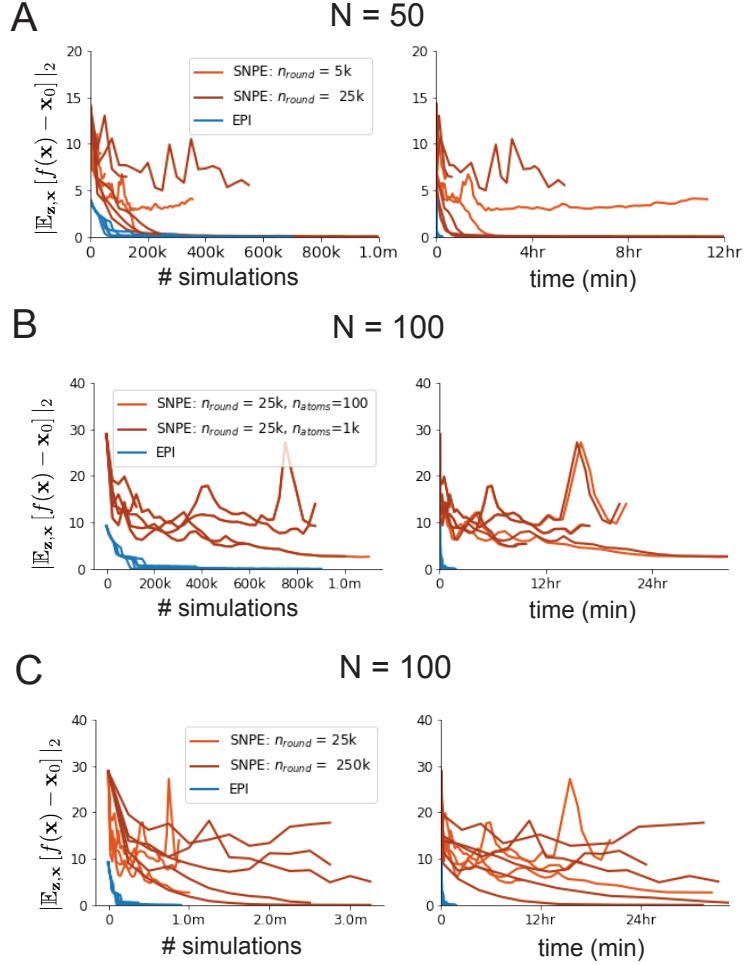


Figure 12: (RNN3): SNPE convergence was enabled by increasing n_{round} , not n_{atom} . **A.** Difference of mean predictions \mathbf{x}_0 throughout optimization at $N = 50$ with by simulation count (left) and wall time (right) of SNPE with $n_{\text{round}} = 5,000$ (light orange), SNPE with $n_{\text{round}}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

seeds) (Fig. 12B). Finally, we increased n_{round} by an order of magnitude, which yielded convergence in one case, but no others. We found no way to improve the convergence rate of SNPE without making more aggressive hyperparameter choices requiring high numbers of simulations.

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

5.2.3 Primary visual cortex

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

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \quad (67)$$

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

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (68)$$

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

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (69)$$

so that $\boldsymbol{\sigma}$ parameterizes the variance of the noisy input in the absence of recurrent connectivity ($W = \mathbf{0}$). As contrast increases, input to the E- and P-populations increases relative to 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 using the deterministic V1 circuit model [55]

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (70)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (71)$$

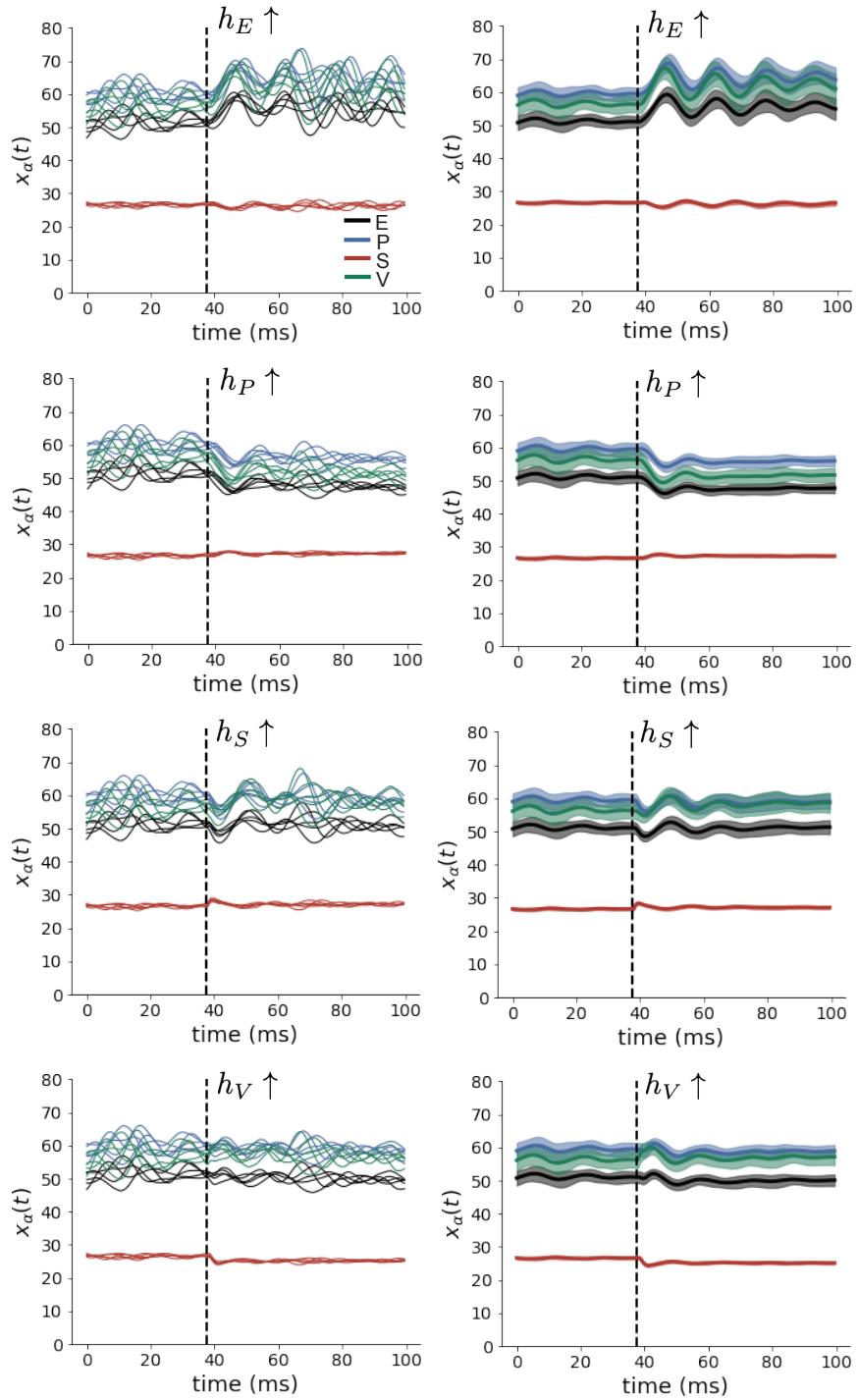


Figure 13: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

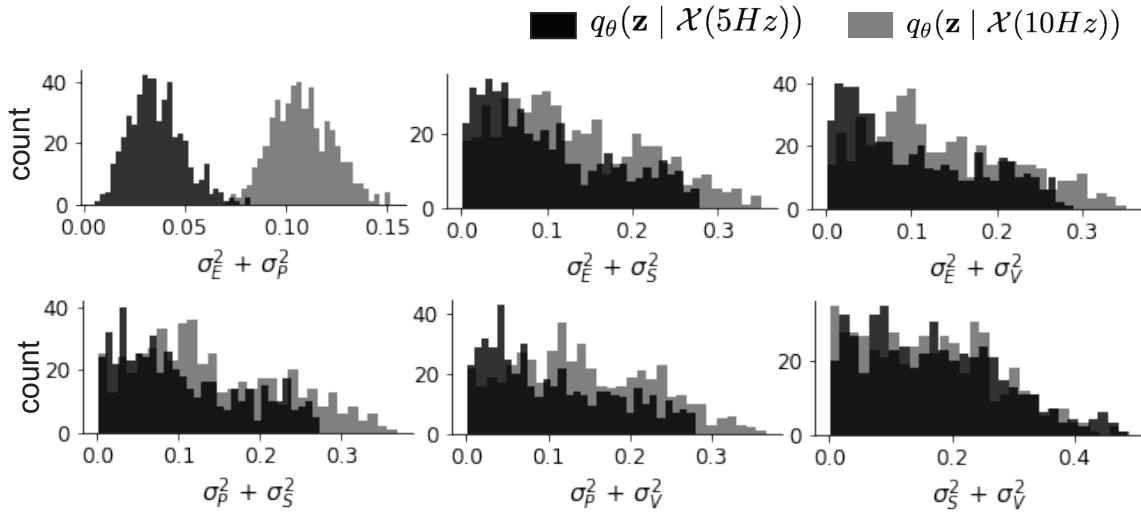


Figure 14: (V1 2) Posterior predictive distributions of the sum of squares of each pair of noise parameters.

1217 and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (72)$$

1218 To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equiv-
 1219 alence class

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (73)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (74)$$

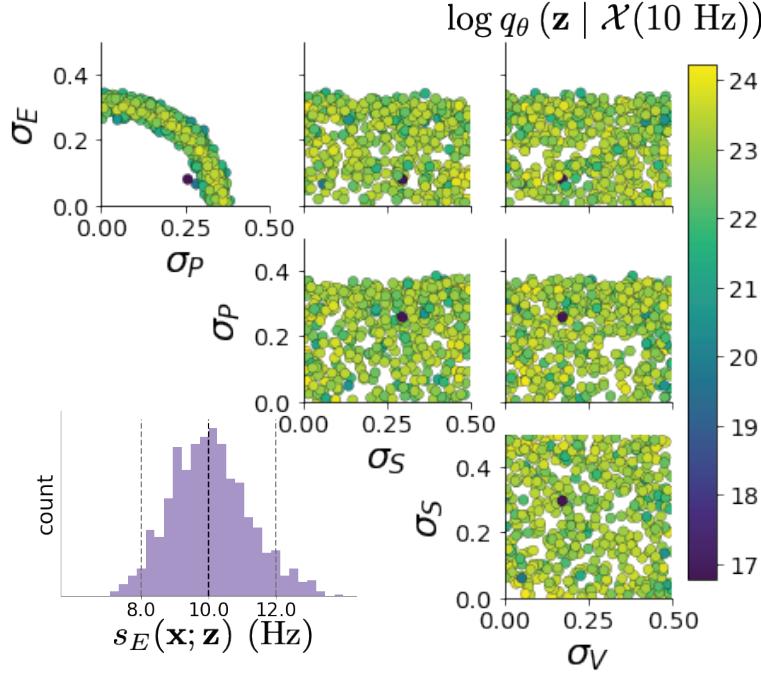


Figure 15: (V1 3) EPI posterior for $\mathcal{X}(10 \text{ Hz})$.

1220 and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (75)$$

1221 Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5\text{ms}$ from an initial condition
 1222 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
 1223 as the square root of the temporal variance from $t_{ss} = 75\text{ms}$ to $Tdt = 100\text{ms}$ averaged over 100
 1224 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[\sqrt{\mathbb{E}_{t > t_{ss}} [(x_E(t) - \mathbb{E}_{t > t_{ss}} [x_E(t)])^2]} \right] \quad (76)$$

1225 For EPI in Fig 3D-E, we used a real NVP architecture with three Real NVP coupling layers
 1226 and two-layer neural networks of 50 units per layer. The normalizing flow architecture mapped
 1227 $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ 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
 1228 different random seeds for architecture initialization θ with an augmented Lagrangian coefficient of
 1229 $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the architectures
 1230 converging with criteria $N_{\text{test}} = 100$ at greatest entropy across random seeds.

1231 In Fig. 3E, we visualize the modes of $q_{\theta}(\mathbf{z} \mid \mathcal{X})$ throughout the σ_E - σ_P marginal. Specifically, we
 1232 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_{\theta}(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (77)$$

1233 At each mode \mathbf{z}^* , we calculated the Hessian and visualized the sensitivity dimension in the direction
 1234 of positive σ_E .

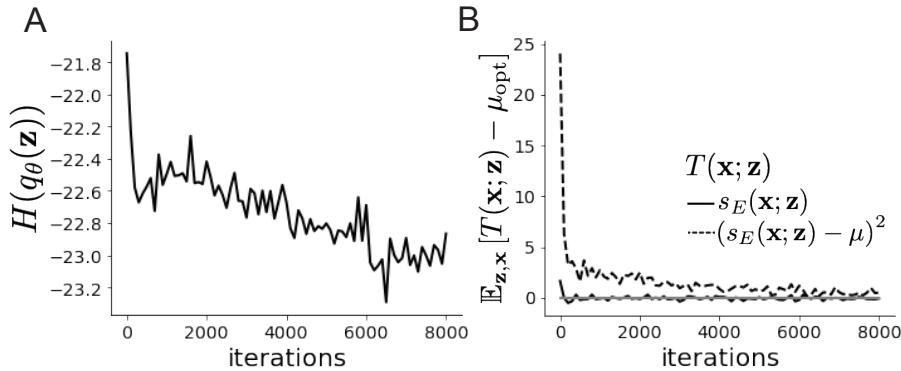


Figure 16: (V1 4) Optimization for V1

1235 5.2.4 Primary visual cortex: challenges to analysis

1236 TODO Agostina and I are putting this together now.

1237 5.2.5 Superior colliculus

1238 The ability to switch between two separate tasks throughout randomly interleaved trials, or “rapid
 1239 task switching,” has been studied in rats, and midbrain superior colliculus (SC) has been shown to
 1240 play an important role in this computation [72]. Neural recordings in SC exhibited two populations of
 1241 neurons that simultaneously represented both task context (Pro or Anti) and motor response (con-
 1242 tralateral or ipsilateral to the recorded side), which led to the distinction of two functional classes:
 1243 the Pro/Contra and Anti/Ipsi neurons [56]. Given this evidence, Duan et al. proposed a model
 1244 with four functionally-defined neuron-type populations: two in each hemisphere corresponding to
 1245 the Pro/Contra and Anti/Ipsi populations. We study how the connectivity of this neural circuit
 1246 governs rapid task switching ability.

1247 The four populations of this model are denoted as left Pro (LP), left Anti (LA), right Pro (RP)
 1248 and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh\left(\frac{u_\alpha - a}{b}\right) + \frac{1}{2} \right), \quad (78)$$

1249 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-
 1250 earity. We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (79)$$

1251 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (80)$$

1252 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2 . These
 1253 hyperparameter values are motivated by modeling choices and results from [56].

1254 The weight matrix has 4 parameters for self sW , vertical vW , horizontal hW , and diagonal dW
 1255 connections:

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (81)$$

1256 We study the role of parameters $\mathbf{z} = [sW, vW, hW, dW]^\top$ in rapid task switching.

1257 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (82)$$

1258 There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (83)$$

1259 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (84)$$

1260 rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (85)$$

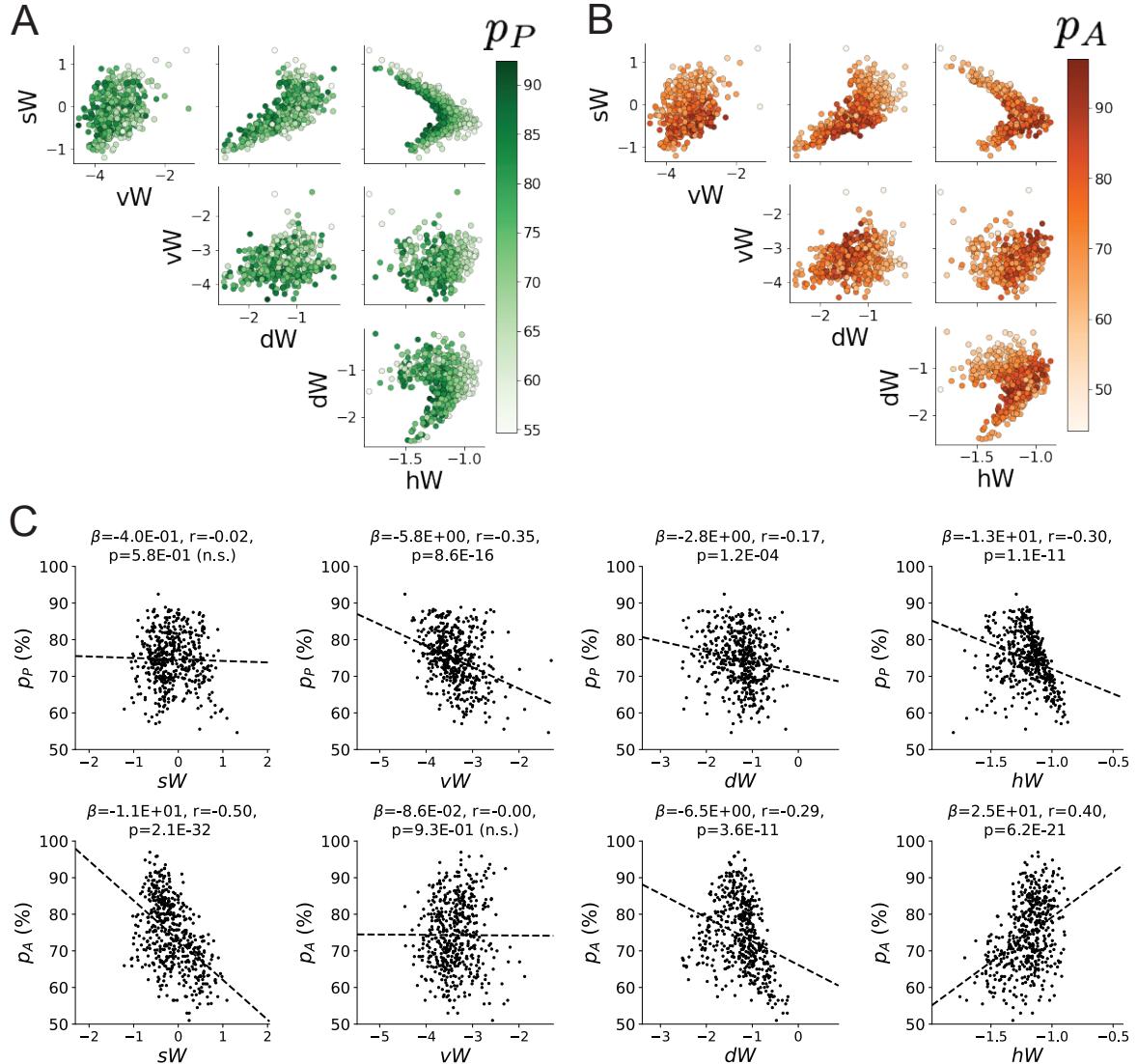


Figure 17: (SC1): **A.** Same pairplot as Fig. 4C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

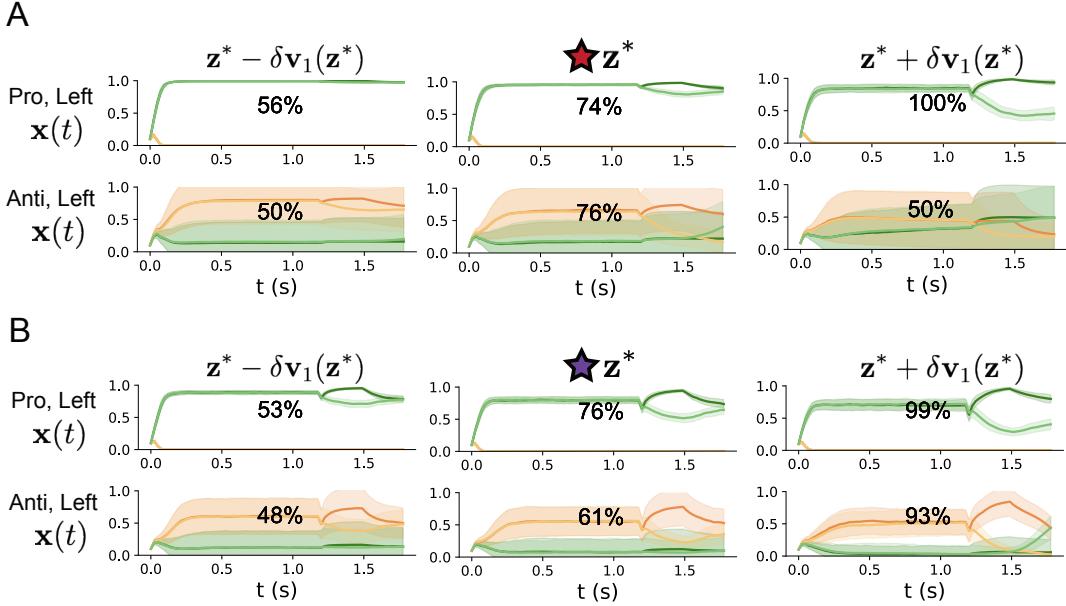


Figure 18: (SC2): **A.** Simulations in network regime 1 ($hW_{\text{fixed}} = -1.2$) (center) with simulations given connectivity perturbations in the negative direction of the sensitivity vector \mathbf{v}_1 (left) and positive direction (right). **B.** Same as A for network regime 2.

1261

$$\mathbf{h}_{A,\text{rule}}(t) = \begin{cases} I_{A,\text{rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (86)$$

1262 a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (87)$$

1263 and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases}. \quad (88)$$

1264 The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{P,\text{bias}} = 0.5$, $I_{P,\text{rule}} = 0.6$, $I_{A,\text{rule}} = 0.6$,
1265 $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

1266 The accuracies of each task p_P and p_A are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (89)$$

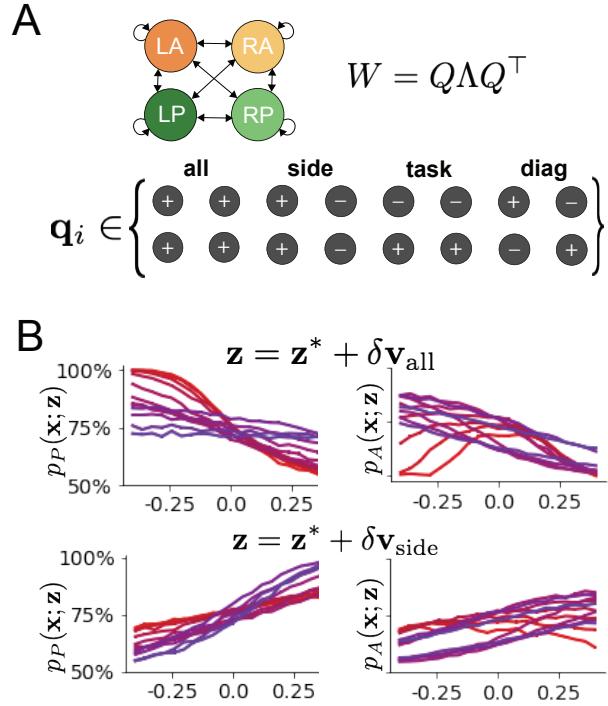


Figure 19: (SC3): **A.** Invariant eigenvectors of connectivity matrix W . **B.** Accuracies for connectivity perturbations for increasing λ_{all} and λ_{side} (rest shown in Fig. 4D).

1267 and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (90)$$

1268 given that the stimulus is on the left side, where Θ is the Heaviside step function, and the accuracy
1269 is averaged over 200 independent trials. The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (91)$$

1270 where $\beta = 100$.

1271 Writing the EPI distribution as a maximum entropy distribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these
1272 first and second moments of the accuracy in each task (as in Equations 28 and 29)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \\ (p_P(\mathbf{x}; \mathbf{z}) - 75\%)^2 \\ (p_A(\mathbf{x}; \mathbf{z}) - 75\%)^2 \end{bmatrix}, \quad (92)$$

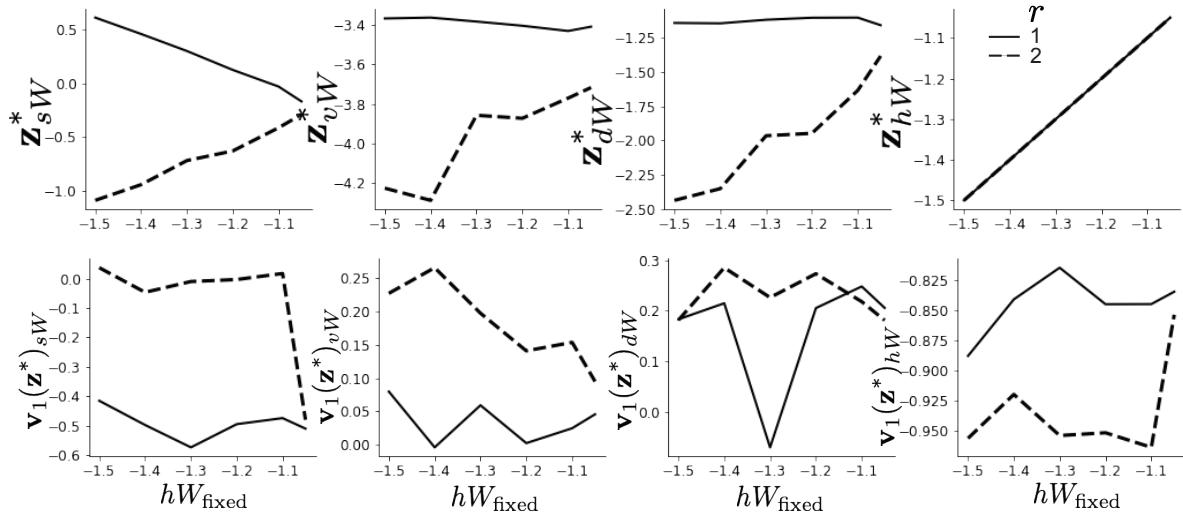


Figure 20: (SC4): **A.** The individual parameters of each mode throughout the two regimes. **B.** The individual sensitivities of parameters of each mode throughout the two regimes.

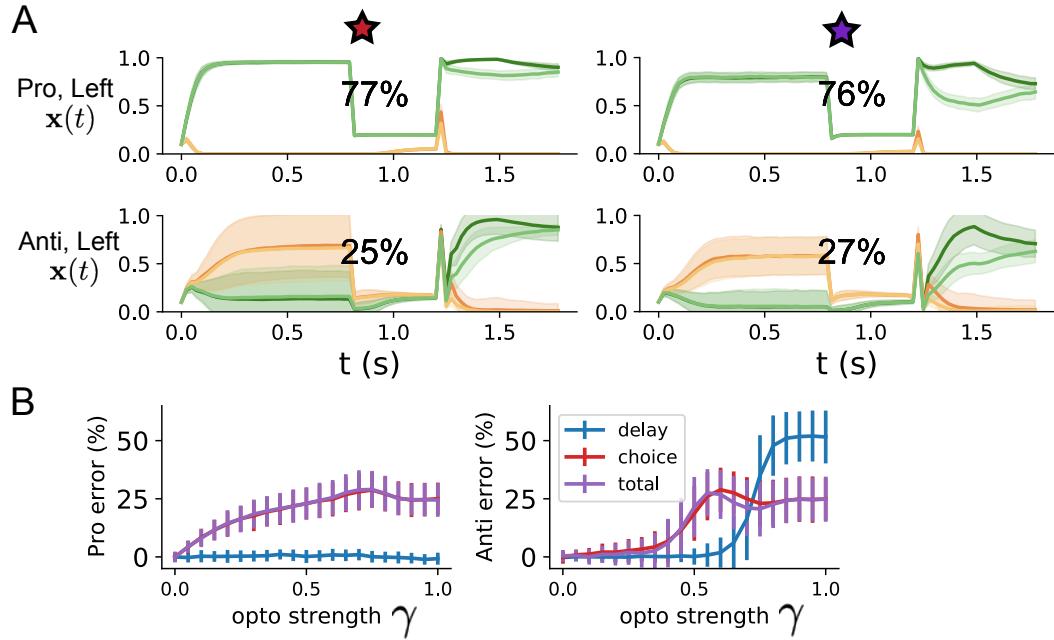


Figure 21: (SC5): **A.** Response of each parameter regime to optogenetic silencing during the delay period. **B.** Error induced by delay period inactivation with increasing optogenetic strength. Means and standard deviations are calculated across the entire EPI posterior.

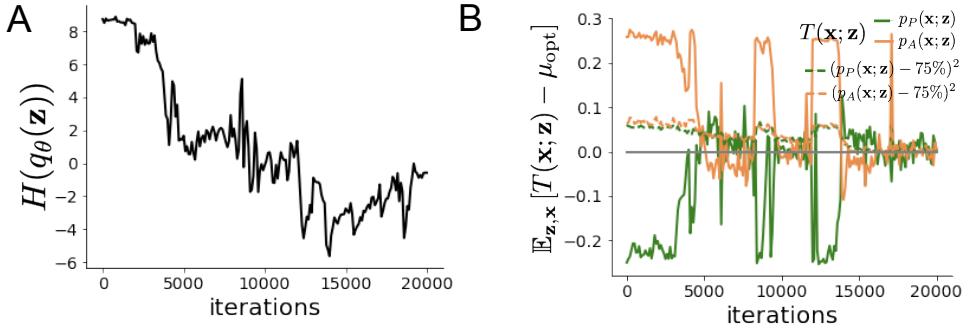


Figure 22: (SC6): **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

1273

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 7.5\%^2 \\ 7.5\%^2 \end{bmatrix}. \quad (93)$$

1274 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 1275 epoch of 2,000 iterations (see Section 5.1.3). The optimization converged after ten epochs (Fig.
 1276 22).

1277 For EPI in Fig. 4C, we used a real NVP architecture with three coupling layers of affine transfor-
 1278 mations parameterized by two-layer neural networks of 50 units per layer. The initial distribution
 1279 was a standard isotropic gaussian $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ mapped to a support of $\mathbf{z}_i \in [-5, 5]$. We used an
 1280 augmented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution
 1281 converged with criteria $N_{\text{test}} = 25$.

1282 The EPI distribution of SC model connectivities producing rapid task switching has interesting
 1283 structure. Throughout $q_{\theta}(\mathbf{z} \mid \mathcal{X})$, we see that the probability distribution is narrow in hW
 1284 (Fig. 4C). This suggests that rapid task switching is sensitive to changes in hW , but this is only a
 1285 single parameter. The local structure of the distribution varies across parameter space, and thus
 1286 the nature in which parameter combinations affect rapid task switching. From visual inspection, we
 1287 may hypothesize that there are two distinct regimes, most easily visualized in the $sW-hW$ marginal
 1288 distribution: one where sW and hW are correlated for greater sW and one where sW and hW are
 1289 anticorrelated for lesser sW .

1290 We sought two sets of parameters in this distribution representative of each regime, so that we
 1291 could assess their implications on computation. For fixed values of hW , we hypothesized that there

1292 are two modes: one in each regime of greater and lesser sW . To begin, we found one mode for each
 1293 regime at $hW_{\text{fixed}} = -1.5$ using 200 steps of gradient ascent of the deep probability distribution
 1294 $q_{\theta}(\mathbf{z} \mid \mathcal{X})$. In regime 1, the initialization had positive sW , and the initialization had negative sW
 1295 in regime 2, which led to disparate modes (Fig. 20 top). These modes were then used as the
 1296 initialization to find the next mode at $hW_{\text{fixed}} = -1.4$ and so on. 200 steps of gradient ascent
 1297 were always taken, and learning rates of 2.5×10^{-4} and 5×10^{-4} were used for regimes 1 and 2,
 1298 respectively. Each of these modes is denoted $\mathbf{z}^*(hW_{\text{fixed}}, r)$ for regime $r \in \{1, 2\}$.

1299 At each mode, we measure the sensitivity dimension (that of most negative eigenvalue in the Hessian
 1300 of the EPI distribution) $\mathbf{v}_1(\mathbf{z}^*)$. To resolve sign degeneracy in eigenvectors, we chose $\mathbf{v}_1(\mathbf{z}^*)$ to have
 1301 negative element in hW . This tells us what parameter combination rapid task switching is most
 1302 sensitive to at this parameter choice in the regime. We see that while the modes of each regime
 1303 gradually converge to similar connectivities at $hW_{\text{fixed}} = -1.05$ (Fig. 20 top), the sensitivity
 1304 dimensions remain categorically different throughout the two regimes (Fig. 20 bottom). Only at
 1305 $hW_{\text{fixed}} = -1.05$ is there a flip in sensitivity from regime 2 to regime 1 (in $\mathbf{v}_1(\mathbf{z}^*)_{sW}$ and $\mathbf{v}_1(\mathbf{z}^*)_{hW}$).
 1306 There is thus some ambiguity regarding the “regime” of $\mathbf{z}^*(-1.05, 2)$, since the mode is derived
 1307 from an initialization in regime 2, but has sensitivity like regime 1. We can consider this as an
 1308 intermediate transitional region of parameter space between the two regimes. To emphasize this,
 1309 $\mathbf{z}^*(-1.05, 1)$ and $\mathbf{z}^*(-1.05, 2)$ have the same color.

1310 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-
 1311 tion of the symmetric connectivity matrices $W = Q\Lambda Q^{-1}$, which results in the same basis vectors
 1312 \mathbf{q}_i for all W parameterized by \mathbf{z} (Fig. 19A). These basis vectors have intuitive roles in processing for
 1313 this task, and are accordingly named the *all* eigenmode - all neurons co-fluctuate, *side* eigenmode
 1314 - one side dominates the other, *task* eigenmode - the Pro or Anti populations dominate the other,
 1315 and *diag* mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair.
 1316 Due to the parametric structure of the connectivity matrix, the parameters \mathbf{z} are a linear function
 1317 of the eigenvalues $\boldsymbol{\lambda} = [\lambda_{\text{all}}, \lambda_{\text{side}}, \lambda_{\text{task}}, \lambda_{\text{diag}}]^T$ associated with these eigenmodes.

$$\mathbf{z} = A\boldsymbol{\lambda} \tag{94}$$

1318

$$A = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}. \tag{95}$$

1319 We are interested in the effect of raising or lowering the amplification of each eigenmode in the

1320 connectivity matrix. To test this, we calculate the unit vector of changes in the connectivity \mathbf{z} that
 1321 result from a change in the associated eigenvalues

$$\mathbf{v}_a = \frac{\frac{\partial \mathbf{z}}{\partial \lambda_a}}{\left\| \frac{\partial \mathbf{z}}{\partial \lambda_a} \right\|_2}, \quad (96)$$

1322 where

$$\frac{\partial \mathbf{z}}{\partial \lambda_a} = A \mathbf{e}_a, \quad (97)$$

1323 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 .

1324 While perturbations in the sensitivity dimension $\mathbf{v}_1(\mathbf{z}^*)$ adapt with the mode \mathbf{z}^* chosen, perturba-
 1325 tions in \mathbf{v}_a for $a \in \{\text{all, side, text, diag}\}$ are invariant to \mathbf{z} (Equation 97).

1326 We tested whether the inferred SC model connectivities could reproduce experimental effects of
 1327 optogenetic inactivation in rats [72]. During periods of simulated optogenetic inactivation, activity
 1328 was decreased proportional to the optogenetic strength γ

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (98)$$

1329 Delay period inactivation was from $0.8 < t < 1.2$, choice period inactivation was for $t > 1.2$ and
 1330 total inactivation was for the entire trial.