

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 mechanistic model – and
33 these equations are motivated by biophysics, neurophysiology, and other conceptual considerations
34 [1–4]. 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 compu-
41 tational properties determine model parameters. Seminal examples of this gold standard include
42 our field’s understanding of memory capacity in associative neural networks [7], chaos and au-
43 tocorrelation timescales in random neural networks [8], the paradoxical effect [9], and decision
44 making [10]. Unfortunately, as circuit models include more biological realism, theory via analytical
45 derivation becomes intractable. Still, we can gain insight into these complex models by identifying
46 the distribution of parameters that produce computations. By solving the inverse problem in this
47 way, scientific analysis of biologically realistic models is made possible [6, 11–14].

48 While theoretical neuroscience is concerned with how model parameters govern computational
49 properties, existing methodology for statistical inference in neuroscience [15–36] (see review, [37])
50 requires that parameters be conditioned on an explicit dataset. The scientific insight for a model

51 of computation is then limited by the quantity and quality of available neural data. Even with a
52 vast amount of high-quality recordings, neural data often reflect uninstructed behaviors [38–40],
53 and thus may only reflect the computation of interest amidst a sea of task-irrelevant factors. A
54 common alternative is to synthesize an explicit dataset that is exemplary of that computation, so
55 that the framework of statistical inference can be applied for parameter identification. In this case,
56 well-defined computational properties are being shoehorned into artificial datasets for the purpose
57 of methodological compatibility.

58 Another key challenge is that as models of computation become more complex, statistical inference
59 becomes intractable. Such mechanistic models in theoretical neuroscience are noisy systems of
60 differential equations that can only be sampled or realized through forward simulation [41, 42];
61 they lack a tractable likelihood function, which is necessary for statistical inference. Therefore, the
62 most popular approaches to parameter inference in mechanistic models have been likelihood-free
63 inference methods [43, 44], in which reasonable parameters are obtained via simulation and rejection.
64 A new class of techniques [45–47] use deep learning to improve upon traditional likelihood-free
65 inference approaches. However, to use these methods in theoretical neuroscience, we must represent
66 computation with an explicit dataset in some way. Theorists are therefore barred from using the
67 probabilistic modeling toolkit for science with circuit models, unless they reformulate their inverse
68 problem into a framework for observational datasets.

69 To address the methodological incongruity between explicit datasets and emergent properties, we
70 present a statistical inference method for conditioning parameters of neural circuit models directly
71 on computation. In this work, we define computation by an emergent property, which is a statistical
72 description of the phenomena to be produced by the neural circuit model. In emergent property
73 inference (EPI), we infer the distribution of model parameters that produce this emergent property.
74 With EPI, parameters are conditioned directly on an implicit dataset defined by the computation
75 of interest. By using recent optimization techniques [48], EPI uses deep learning to make rich,
76 flexible approximations to the parameter distributions [49], the structure of which reveals scientific
77 insight about how parameters govern the emergent property.

78 Equipped with this method, we prove out the potential of EPI by demonstrating its capabilities and
79 presenting novel theoretical findings borne from its analysis. First, we show EPI’s ability to handle
80 mechanistic models using a classic model of parametric degeneracy in biology: the stomatogastric
81 ganglion [50, 51]. Then, we show EPI’s scalability to high dimensional parameter distributions by
82 inferring connectivities of recurrent neural networks (RNNs) that exhibit stable, yet amplified re-

sponses – a hallmark of neural responses throughout the brain [52–54]. In a model of primary visual cortex (V1) [55, 56] with different neuron-types, we show that the equation for excitatory variability become analytically intractable as more populations are added. Strikingly, the way in which noisy inputs across neuron-types governs excitatory variability is salient in the visualized structure of the EPI inferred parameter distribution. Finally, we investigated the possible connectivities of superior colliculus (SC) that allow execution of different tasks on interleaved trials [57]. EPI discovered a rich distribution containing two connectivity regimes with different solution classes. We queried the deep probability distribution learned by EPI to produce a mechanistic understanding of cortical responses in each regime. Intriguingly, all inferred connectivities reproduced results from optogenetic inactivation experiments in this behavioral paradigm – emergent phenomena that EPI was not conditioned upon. These theoretical insights afforded by EPI illustrate the value of deep inference for the interrogation of neural circuit models.

3 Results

3.1 Motivating emergent property inference of theoretical models

Consideration of the typical workflow of theoretical modeling clarifies the need for emergent property inference. First, one designs or chooses an existing model that, it is hypothesized, captures the computation of interest. To ground this process in a well-known example, consider the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic muscle activation patterns for digestion [58]. Despite full knowledge of STG connectivity and a precise characterization of its rhythmic pattern generation, biophysical models of the STG have complicated relationships between circuit parameters and computation [12, 50]. A subcircuit model of the STG [51] is shown schematically in Figure 1A. The jagged connections indicate electrical coupling having electrical conductance g_{el} , smooth connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto the hub neuron, and $g_{synB} = 5nS$ for mutual inhibitory connections. Note that the behavior of this model will be critically dependent on its parameterization – the choices of conductance parameters $\mathbf{z} = [g_{el}, g_{synA}]$. 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 this STG

model, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model 1B. An emergent property studied by Guttierrez et al. of this stochastic model is the hub neuron firing 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. Our notion of intermediate hub frequency is not strictly 0.55Hz, but also moderate deviations of this frequency between the fast (.35Hz) and slow (.68Hz) frequencies, which are quantified in the emergent property with variance 0.025^2Hz^2 .

Third, the model parameters producing these outputs are inferred. To infer the STG parameters of intermediate hub frequency with existing methodology, we need an explicit dataset: experimentally recorded or synthesized. By precisely quantifying the emergent property of interest as a statistical feature of the model, we use EPI to condition directly on this emergent property. EPI learns a probability distribution of model parameters constrained to produce the emergent property. In this last step lies the opportunity for a shift away from a dataset-oriented representation of model output towards that of an implicit dataset, where the only structure is the emergent property of interest.

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 that produces the emergent property.

3.2 A deep generative modeling approach to emergent property inference

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

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

where $C_m=1\text{nF}$, and \mathbf{h} is a sum of the leak, calcium, potassium, hyperpolarization, electrical, and synaptic currents, all of which have their own complicated dependence on activity \mathbf{x} and parameters $\mathbf{z} = [g_{el}, g_{synA}]$, and $d\mathbf{B}$ is white gaussian noise [51, 59] (see Section 5.2.1 for more detail).



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 distribution producing intermediate hub frequency. Samples are colored by \log probability density. Contours of hub neuron frequency error are shown at levels of $.525, .53, \dots, .575$ 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.

142 Second, we stipulate that our model should produce the emergent property of “intermediate hub
 143 frequency” (Figure 1C). We stipulate that the hub neuron’s spiking frequency – denoted $\omega_{\text{hub}}(\mathbf{x})$
 144 is close to a frequency of 0.55Hz, between that of the slow and fast frequencies. Mathematically,
 145 we define this emergent property with two statistical constraints: that the mean hub frequency is
 146 0.55Hz,

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

147 and that the variance of the hub frequency is moderate

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

148 The hub neuron frequency is constrained over the distribution of parameters \mathbf{z} and the distribution
 149 of the data \mathbf{x} that those parameters produce. Formally, the emergent property is the collection of
 150 these two constraints

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

151 In general, an emergent property is a collection of first-, second- and higher moments of statistics
 152 that together define the phenomena.

153 Third, we perform emergent property inference: we find a distribution over parameter configu-
 154 rations \mathbf{z} that produces the emergent property; in other words, they obey the constraints intro-
 155 duced in Equation 4. This distribution will be chosen from a family of probability distributions
 156 $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$, defined by a deep neural network [49, 60, 61] (Figure 1D, EPI box). Deep
 157 probability distributions map a simple random variable \mathbf{z}_0 through a deep neural network with
 158 weights and biases $\boldsymbol{\theta}$ to parameters $\mathbf{z} = g_{\boldsymbol{\theta}}(\mathbf{z}_0)$ to a suitably complicated distribution (see Section
 159 5.1.2 for more details). Many distributions in \mathcal{Q} will respect the emergent property constraints,
 160 so we select the most random (highest entropy) distribution, which is the same choice made in
 161 Bayesian posterior inference (see Section 5.1.6). In EPI optimization, stochastic gradient steps in
 162 $\boldsymbol{\theta}$ are taken such that entropy is maximized, and the emergent property \mathcal{X} is produced (see Section
 163 5.1) The inferred EPI distribution is denoted $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X})$, since it is conditioned upon emergent
 164 property \mathcal{X} . This is meant to share the same notation as a posterior distribution $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathbf{x})$ that is
 165 conditioned upon an explicit dataset.

166 The major scientific value of EPI is in the rich, queryable structure of these deep probability
 167 distributions. The probabilities of $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X})$ are the densities of these parameters in the distribution
 168 producing the emergent property. The greatest probabilities (the modes) indicate prototypical

169 parameter configurations, and the manner in which probabilities change away from the modes shows
170 how different parameter combinations preserve or diminish the emergent property. The dimensions
171 of greatest sensitivity (e.g. Fig. 1E solid arrow) or degeneracy (e.g. Fig. 1E dashed arrow) can be
172 measured directly from the second order derivative of $\log q_{\theta}(\mathbf{z} \mid \mathcal{X})$ called the “Hessian.” Around
173 the mode, eigenvalues of the Hessian are negative; probabilities decrease locally in all directions
174 away from the mode. The eigenvector with most negative eigenvalue is the parameter combination
175 causing probability to decrease the fastest, making it the most sensitive dimension. Likewise, the
176 flattest eigenvector, corresponding to the least negative eigenvalue, points in the most degenerate
177 dimension. Once an EPI distribution has been inferred, this second order derivative requires trivial
178 computation (when correct architecture class is chosen, see Section 5.1.2).

179 In the following sections, we showcase the versatility of EPI for scientific analysis on three neural
180 circuit models across ranges of biological realism, neural system function, and network scale. First,
181 we demonstrate the superior scalability of EPI compared to alternative techniques by inferring high-
182 dimensional distributions of RNN connectivities that exhibit amplified, yet stable responses. Also
183 in this RNN example, we emphasize that EPI is the only technique that controls the predictions
184 made by the inferred parameter distribution. Next, in a model of primary visual cortex [55,
185 56], we show how to gain insight by comparing multiple inferred distributions. Finally, we used
186 EPI to capture subtle parametric structure allowing the mechanistic characterization of multiple
187 parametric regimes of superior colliculus activity in a model of task switching [57]. This work is
188 the first to produce this level of theoretical insight via the quantification and examination of the
189 intricate structure captured by deep probability distributions.

190 3.3 Scaling inference of RNN connectivity with EPI

191 Transient amplification is a hallmark of neural activity throughout cortex, and is often thought to be
192 intrinsically generated by recurrent connectivity in the responding cortical area [52–54]. It has been
193 shown that to generate such amplified, yet stabilized responses, the connectivity of RNNs must be
194 non-normal [52, 62], and satisfy additional constraints [63]. In theoretical neuroscience, RNNs are
195 optimized and then examined to show how dynamical systems could execute a given computation
196 [64, 65], but such biologically realistic constraints on connectivity are ignored during optimization
197 for practical reasons. In general, access to distributions of connectivity adhering to theoretical
198 criteria like stable amplification, chaotic fluctuations [8], or low tangling [66] would add scientific
199 value and context to existing research with RNNs. Here, we use EPI to learn RNN connectivities

200 producing stable amplification, and demonstrate the superior scalability and efficiency of EPI to
 201 alternative approaches.

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

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

203 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 $\chi_{i,j}^{(U)}, \chi_{i,j}^{(V)} \sim$
 204 $\mathcal{N}(0, 1)$. 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 ampli-
 205 fication. Two conditions are necessary and sufficient for RNNs to exhibit stable amplification [63]:
 206 $\text{real}(\lambda_1) < 1$ and $\lambda_1^s > 1$, where λ_1 is the eigenvalue of W with greatest real part and λ^s is the max-
 207 imum eigenvalue of $W^s = \frac{W+W^\top}{2}$. RNNs with $\text{real}(\lambda_1) = 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$ will be stable
 208 with modest decay rate ($\text{real}(\lambda_1)$ close to its upper bound of 1) and exhibit modest amplification
 209 (λ_1^s close to its lower bound of 1). EPI can naturally condition on this emergent property

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

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

212 For comparison, we infer the parameters \mathbf{z} likely to produce stable amplification using two alter-
 213 native likelihood-free inference approaches. We ran sequential Monte Carlo approximate Bayesian
 214 computation (SMC-ABC) [43] and sequential neural posterior estimation (SNPE) [45] with ob-
 215 servation $\mathbf{x}_0 = \boldsymbol{\mu}$. SMC-ABC is a rejection sampling approach that SMC techniques to improve
 216 efficiency, and SNPE approximates posteriors with deep probability distributions using a two-
 217 network architecture (see Section 5.1.1). Unlike EPI, these statistical inference techniques do not
 218 control the mean or variance of the predictive distribution, and these predictions of the inferred
 219 posteriors are typically affected by model characteristics (e.g. N and g , Fig. 11). To compare the
 220 efficiency of these different techniques, we measured the time and number of simulations necessary
 221 for the distance of the predictive mean to be less than 0.5 from $\boldsymbol{\mu} = \mathbf{x}_0$ (see Section 5.2.2).

222 As the number of neurons N in the RNN is scaled, and thus the dimension of the parameter
 223 space $\mathbf{z} \in [-1, 1]^{4N}$, we see that EPI converges at greater speed and at greater dimension than
 224 SMC-ABC and SNPE (Fig. 2A). It also becomes most efficient to use EPI in terms of simulation
 225 count at $N = 50$ (Fig. 2B). It is well known that ABC techniques struggle in dimensions greater



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.

than about 30 [67], 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 with respect to their theoretical properties. EPI outperforms SNPE in high dimensions by using gradient information (from $\nabla_{\mathbf{z}}f(\mathbf{x}; \mathbf{z}) = \nabla_{\mathbf{z}}[\text{real}(\lambda_1), \lambda_1^s]^\top$) on each optimization iteration. This agrees with recent speculation that such gradient information could improve the efficiency of LFI techniques [68]. While scaling to high dimensions is important, we show in the next two sections how insight can be gained by inspecting structure in lower dimensional parameter distributions.

3.4 EPI reveals how noisy input across neuron-types governs excitatory variability in a V1 model

Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output function have succeeded in explaining a host of experimentally documented phenomena. In a regime characterized by inhibitory stabilization of strong recurrent excitation, these models give rise to paradoxical responses [9], selective amplification [52, 62], surround suppression [69] and normalization [70]. Despite their strong predictive power, E-I circuit models rely on the assumption that inhibition can be studied as an indivisible unit. However, experimental evidence shows that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) – composing 80% of GABAergic interneurons in V1 [71–73], and that these inhibitory cell types follow specific

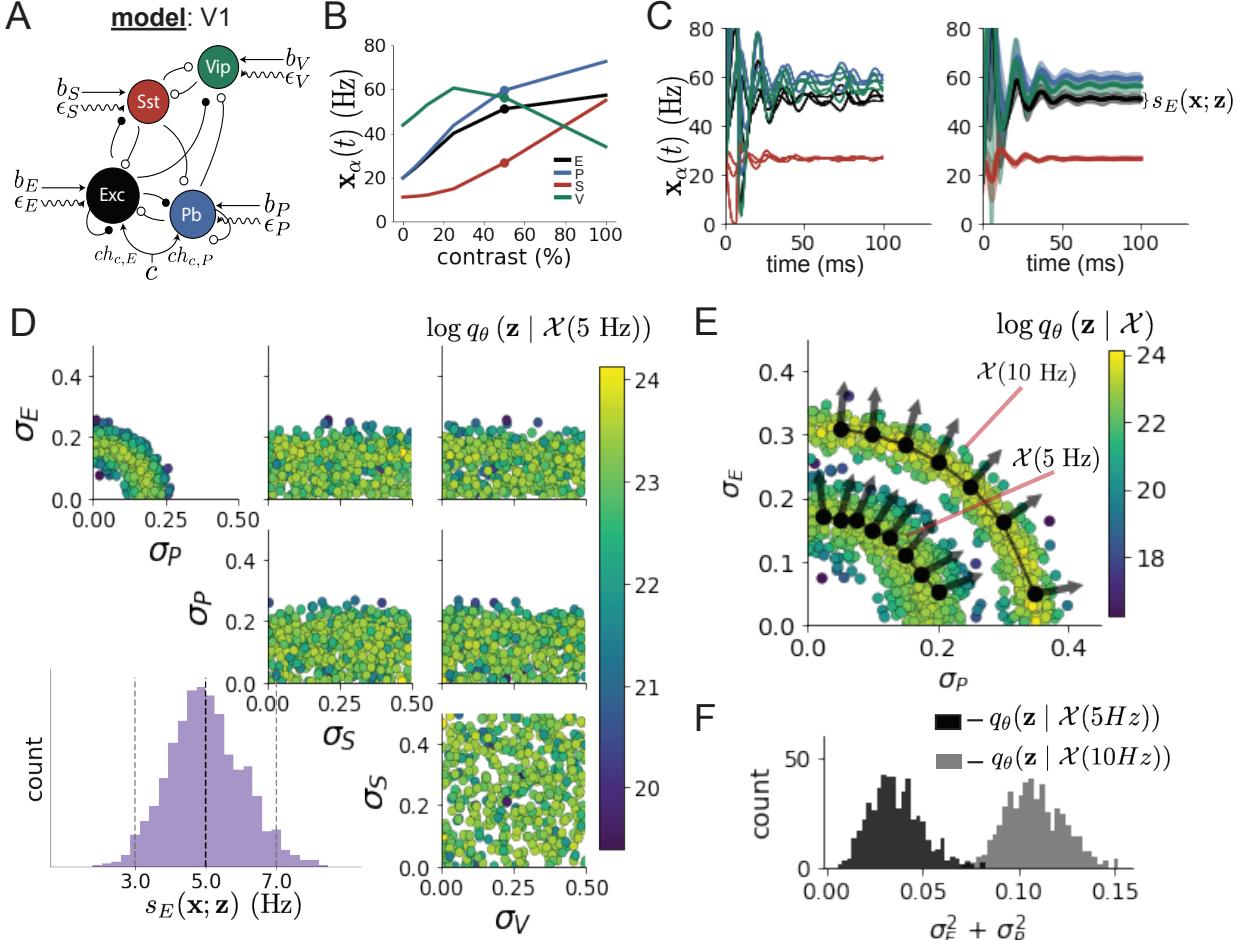


Figure 3: Emergent property inference in the stochastic stabilized supralinear network (SSSN) **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). 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 $q_\theta(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$.

257 connectivity patterns (Fig. 3A) [74]. While research has shown that V1 only shares specific dimensions
 258 of neuronal variability with downstream areas [75], the role played by recurrent dynamics and
 259 the connectivity across neuron-type populations is not understood. Here, in a model of V1 with
 260 biologically realistic connectivity, we use EPI to show how the structure of input across neuron
 261 types affects the variability of the excitatory population – the population largely responsible for
 262 projecting to other brain areas [76].

263 We considered response variability of a nonlinear dynamical V1 circuit model (Fig. 3A) with a
 264 state comprised of each neuron-type population’s rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population
 265 receives recurrent input $W\mathbf{x}$, where W is the effective connectivity estimated from post-synaptic
 266 potential and connectivity rate measurements (see Section 5.2.3). Each population also experiences
 267 an external input \mathbf{h} , which determines population rate via supralinear nonlinearity $\phi(\cdot) = [\cdot]_+^2$. To
 268 build on previous work, we model visual contrast-dependent input to the E- and P-populations
 269 $\mathbf{h} = \mathbf{b} + c\mathbf{h}_c$. There is also an additive noisy input ϵ parameterized by variances for each neuron
 270 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 $\tau_{\text{noise}} > \tau$
 271 then the population rate, allowing fluctuations around a stimulus-dependent steady-state

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

272 This model is the stochastic stabilized supralinear network (SSSN) [77] generalized to have mul-
 273 tiple inhibitory neuron types, and introduces stochasticity to previous four neuron-type models
 274 of V1 [55]. Stochasticity and inhibitory multiplicity introduce substantial complexity to mathe-
 275 matical derivations (see Section 5.2.4) motivating the treatment of this model with EPI. Here, we
 276 consider fixed weights W and input \mathbf{h} [56] (Fig. 3B), and study the effect of input variability
 277 $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on excitatory variability at 50% contrast.

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

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

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

281 We ran EPI to obtain parameter distribution $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ producing E-population variability
 282 around 5 Hz (Fig. 3D). From the marginal distribution of σ_E and σ_P (Fig. 3D, top-left), we can see
 283 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
 284 degenerate with respect to $s_E(\mathbf{x}; \mathbf{z})$ evidenced by the high variability in those dimensions (Fig. 3D,

bottom-right). Together, these observations imply a curved path of parametric degeneracy with respect to $s_E(\mathbf{x}; \mathbf{z})$ of 5 Hz, which is indicated by the modes along σ_P (Fig. 3E). The dimensions of sensitivity conferred by EPI and this plain visual structure suggest a quadratic relationship in the emergent property statistic $s_E(\mathbf{x}; \mathbf{z})$ and parameters \mathbf{z} , which is preserved at a greater level of variability $\mathcal{X}(10 \text{ Hz})$ (Fig. 3E). Indeed, the sum of squares of σ_E and σ_P is larger in $q_{\theta}(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$ than $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ (Fig 3F, $p < 1 \times 10^{-10}$), while the sum of squares of σ_S and σ_V are not significantly different in the two EPI distributions (Fig. 15, $p = .40$). The strong compensatory influence of the E- and P-population input variability on excitatory variability is intriguing, since this circuit exhibited a paradoxical effect in the P-population (and no other inhibitory types) at 50% contrast (Fig. 15) meaning that the E-population is P-stabilized. Future research may uncover a link between the populations of stabilizations and compensatory interactions governing excitatory variability.

EPI revealed the quadratic relationship between $s_E(\mathbf{x}; \mathbf{z})$ and \mathbf{z} . While this property is ultimately derivable, we show that with each additional neuron-type population, the formula becomes quite unruly and likely escapes comprehensible analysis in our case (see Section 5.2.4). This emphasizes the need for streamlined methods for gaining understanding about theoretical models when mathematical analysis becomes prohibitive.

3.5 EPI identifies two regimes of rapid task switching

It has been shown that rats can learn to switch from one behavioral task to the next on randomly interleaved trials [78], and an important question is what types of neural connectivity allow this ability. In this experimental setup, rats were explicitly cued on each trial to either orient towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti (A) task (Fig. 4A). Neural recordings in superior colliculus (SC) exhibited two populations of neurons that represented task context (Pro or Anti). Furthermore, Pro/Anti neurons in each hemisphere were strongly correlated with the animal's decision [57]. These results motivated a model of SC that is a four-population dynamical system with functionally-defined neuron-types. Here, our goal is to understand how connectivity in this circuit model governs the ability to switch tasks rapidly.

In this SC model, there are Pro- and Anti-populations in each hemisphere (left (L) and right (R)) with activity variables $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^{\top}$. The connectivity of these populations is parameterized by self sW , vertical vW , diagonal dW and horizontal hW connections (Fig. 4B). The input \mathbf{h} is comprised of a positive cue-dependent signal to the Pro or Anti populations, a positive

316 stimulus-dependent input to either the Left or Right populations, and a choice-period input to the
 317 entire network (see Section 5.2.5). Model responses are bounded from 0 to 1 as a function ϕ of an
 318 internal variable \mathbf{u}

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

$$\mathbf{x} = \phi(\mathbf{u}).$$

319 The model responds to the side with greater Pro neuron activation; e.g. the response is left if
 320 $x_{LP} > x_{RP}$ at the end of the trial. Here, we use EPI to determine the network connectivity
 321 $\mathbf{z} = [sW, vW, dW, hW]^\top$ that produces rapid task switching.

322 We define the computation of rapid task switching as accurate execution of each task. Inferred
 323 models should not exhibit fully random responses (50%), or perfect performance (100%), since
 324 perfection is never attained by even the best trained rats. We formulate rapid task switching as an
 325 emergent property by stipulating that the average accuracy in the Pro task $p_P(\mathbf{x}; \mathbf{z})$ and Anti task
 326 $p_A(\mathbf{x}; \mathbf{z})$ be 75% with variance $7.5\%^2$.

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

327 The EPI inferred distribution (Fig. 4C) produces task accuracies (Fig. 4C, middle-left) according
 328 to our mathematical definition of rapid task switching (Equation 10). The patterns of connectivity
 329 that govern each task accuracy are nonlinear (Fig. 17A-B); they are not captured well by linear
 330 correlations (Fig. 17C). For example, there appear to be two regimes of connectivity: the local
 331 structure of the EPI distribution changes dramatically after crossing a threshold of sW (Fig. 17A
 332 $sW-hW$ marginal distribution). Not only has EPI captured this intricate, nonlinear distribution,
 333 we can use the distribution $q_\theta(\mathbf{z} | \mathcal{X})$ returned by EPI to understand these two parametric regimes
 334 of SC connectivity.

335 To distinguish these two parts of the distribution, we point out that for many fixed values of
 336 parameter hW , there are two modes in the EPI distribution. Thus, by fixing hW to different
 337 values and doing gradient ascent on $\log q_\theta(\mathbf{z} | \mathcal{X})$ in the parameter spaces proximal to each mode,
 338 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).
 339 As hW_{fixed} increases, the modes coalesce to intermediate parameters reflecting a transition between
 340 the two sets of modes (Fig. 20 top). However, the sensitivity dimensions of these modes \mathbf{v}_1 (refer



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 to Section 3.2), which reflect the structure of the EPI distribution around each mode, are different
342 across putative regime, yet consistent across hW_{fixed} . This categorical difference in sensitivity
343 dimension across the two sets of modes shows that they indeed represent two different regimes of
344 computation in which connectivity governs computation in different ways.

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

353 To understand the connectivity mechanisms that distinguish these two regimes, we perturb connec-
354 tivity at each mode in dimensions that have well defined roles in processing for the Pro and Anti
355 tasks. A convenient property of this connectivity parameterization is that there are \mathbf{z} -invariant
356 eigenmodes of connectivity, whose eigenvalues (or degree of amplification) change with \mathbf{z} . These
357 eigenmodes have intuitive roles in processing in each task, and are accordingly named the *all*,
358 *side*, *task*, and *diag* eigenmodes (see Section 5.2.5). Furthermore, the parameter dimension \mathbf{v}_a
359 ($a \in \{\text{all}, \text{side}, \text{task}, \text{and diag}\}$) that increases the eigenvalue of connectivity λ_a is \mathbf{z} -invariant (un-
360 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
361 of each processing mode by perturbing \mathbf{z} along \mathbf{v}_a , we can elicit the differentiating properties of
362 the two regimes.

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

373 **3.6 EPI inferred SC connectivities reproduce results from optogenetic inacti-**
374 **vation experiments**

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

382 Similarities across Pro and Anti trials in choice period responses following delay period inactivation
383 (Fig. 21A) suggested that connectivity patterns inducing greater Pro task accuracy increase error
384 in delay period inactivated Anti trials (Fig. 5B). The strong anticorrelation between p_P and $p_{A,opto}$
385 across EPI inferred connectivities led to the following hypothesis about each connectivity regime:
386 the sensitivity dimension of each regime decreases $p_{A,opto}$ irrespective of its effect on p_A , since
387 both \mathbf{v}_1 and \mathbf{v}_2 increase p_P . Indeed, in regimes 1 and 2 where sensitivity dimensions elicit different
388 responses in p_A , $p_{A,opto}$ decreases since the connectivity changes enhancing p_P exacerbate Anti trial
389 error (Fig. 5C). Thus, the altered state caused by delay period silencing makes the connectivity
390 governing p_P more influential on Anti accuracy than the connectivity governing p_A .

391 In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task
392 switching. This EPI distribution revealed two regimes of rapid task switching, which we char-
393 acterized using the probabilistic toolkit EPI seemlessly affords. We found that both of these
394 parametric regimes identified by EPI reproduce results from optogenetic inactivation experiments:
395 when activity is silenced during the delay period, only Anti accuracy suffers. We then identified
396 the connectivity mechanisms governing Anti accuracy during delay period silencing, and showed
397 that they are regime invariant.

398 **4 Discussion**

399 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets [37].
400 Careful inference procedures are developed for these statistical models allowing precise, quantitative
401 reasoning, which clarifies the way data informs beliefs about the model parameters. However,
402 these statistical models often lack resemblance to the underlying biology, making it unclear how

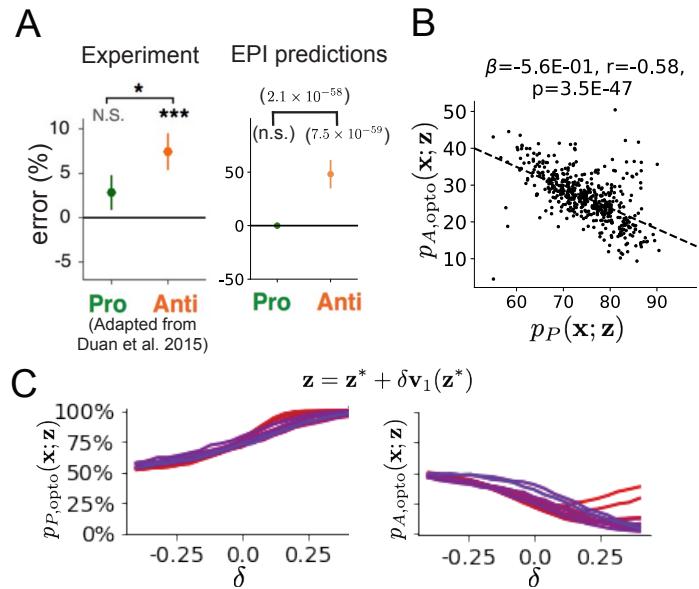


Figure 5: **A.** The EPI distribution predicts experimental results (left) showing no change in the Pro task, but larger error in the Anti task (right). **B.** Accuracy in the Anti task during delay period optogenetic inactivation $p_{A,\text{opto}}$ is strongly anticorrelated with accuracy in the Pro 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}^* .

403 to go from the structure revealed by these methods, to the neural mechanisms giving rise to
 404 it. In contrast, theoretical neuroscience has focused on careful mechanistic modeling and the
 405 production of emergent properties of computation, rather than measuring structure in some noisy
 406 observed dataset. The careful steps of *i.)* model design and *ii.)* emergent property definition,
 407 are followed by *iii.)* practical inference methods resulting in an opaque characterization of the
 408 way model parameters govern computation. In this work, we improve upon parameter inference
 409 techniques in theoretical neuroscience with emergent property inference, harnessing deep learning
 410 towards parameter inference with respect to emergent phenomena in interpretable models of neural
 411 computation (see Section 5.1.1).

412 Methodology for statistical inference in mechanistic models of neural circuits has evolved consider-
 413 ably in recent years. Early work used rejection sampling techniques [43, 79, 80], but more recently
 414 developed methodology employs deep learning to improve efficiency or provide flexible distribution
 415 approximations. SNPE [45] and other sequential techniques for inference in mechanistic models
 416 developed along with EPI (see Section 5.1.1) have been used for posterior inference with noisy
 417 experimental datasets. On the other hand, EPI is a deep inference technique designed to condition
 418 upon mathematical criteria, such that the parameter distribution only produces the specified *emer-*
 419 *gent properties* of computation. EPI is thus ideally suited for questions in theoretical neuroscience,

420 and we show that it has superior scaling properties to these other inference techniques (see Section
421 3.3).

422 In this work, we prove out the utility of deep probability distributions for theoretical neuroscience.
423 While previous work has used SNPE to obtain flexible posterior approximations in mechanistic
424 models conditioned on experimental datasets, we use the rich structure captured by deep probability
425 distributions in EPI to gain new theoretical insights. This is first done in a complex model of V1,
426 where we combine the modeling advancements [55, 77], which make analytic characterization of
427 excitatory variability very complicated. There, EPI clearly and simply revealed the parametric
428 structure of input variability across neuron-type populations that governed excitatory variability,
429 which has implications on the dimensionality and nature information transmission in visual cortex.

430 Finally, we used EPI to identify two distinct regimes of SC connectivity that enabled rapid task
431 switching. By systematically characterizing the local structure of the inferred distribution using the
432 analytic capabilities deep probability distributions, we discerned a mechanistic understanding of
433 each computational regime. Each of these regimes reproduced effects from optogenetic experiments
434 [78], suggesting that both are biologically plausible. These analyses of the V1 and SC models serve
435 as examples of how to leverage the probabilistic toolkit for theoretical insight into models of neural
436 computation.

437 **Acknowledgements:**

438 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
439 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
440 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
441 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
442 Fellowship. Helpful conversations were had with Francesca Mastrogiovanni, Srdjan Ostojic, James
443 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

444 **Data availability statement:**

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

447 **Code availability statement:**

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

449 **References**

- 450 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
451 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 452 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
453 21(1):25–45, 1998.
- 454 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 455 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in
456 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 457 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
458 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
459 models. *PLoS Comput Biol*, 3(10):e189, 2007.
- 460 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
461 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
462 sion model. *Neuron*, 82(4):809–821, 2014.
- 463 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-
464 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 465 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
466 networks. *Physical review letters*, 61(3):259, 1988.
- 467 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
468 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
469 17(11):4382–4388, 1997.
- 470 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration
471 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 472 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
473 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 474 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
475 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.

- 476 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
477 model. *PLoS computational biology*, 2(7):e94, 2006.
- 478 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
479 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 480 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
481 13(8):1713–1720, 2001.
- 482 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
483 A statistical paradigm for neural spike train decoding applied to position prediction from
484 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
485 7425, 1998.
- 486 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
487 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 488 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.
489 A point process framework for relating neural spiking activity to spiking history, neural
490 ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 491 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correlations
492 imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–
493 1012, 2006.
- 494 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan
495 Segev. A novel multiple objective optimization framework for constraining conductance-based
496 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 497 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature
498 analysis. *Neural computation*, 19(4):1022–1038, 2007.
- 499 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
500 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of
501 neural population activity. In *Advances in neural information processing systems*, pages
502 1881–1888, 2009.

- 503 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and
504 Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural*
505 *information processing systems*, 24:1350–1358, 2011.
- 506 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In
507 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 508 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-
509 dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,
510 9(3):e1002922, 2013.
- 511 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
512 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
513 *Science*, 349(6244):184–187, 2015.
- 514 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq
515 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in
516 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 517 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
518 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th In-*
519 *ternational Conference on Machine Learning*, 2019.
- 520 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.
521 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.
522 *Nature Communications*, 10(4933), 2019.
- 523 [30] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
524 neural population models through nonlinear embeddings. In *Advances in neural information*
525 *processing systems*, pages 163–171, 2016.
- 526 [31] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-
527 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 528 [32] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-
529 encoders. *bioRxiv*, page 399246, 2018.
- 530 [33] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
531 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R

- 532 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-
533 encoders. *Nature methods*, page 1, 2018.
- 534 [34] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
535 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
536 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 537 [35] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
538 Datta. Composing graphical models with neural networks for structured representations and
539 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 540 [36] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
541 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
542 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural
543 Information Processing Systems*, 2019.
- 544 [37] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
545 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
546 2018.
- 547 [38] Christopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral
548 state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.
- 549 [39] Aman B Saleem, Aslı Ayaz, Kathryn J Jeffery, Kenneth D Harris, and Matteo Carandini.
550 Integration of visual motion and locomotion in mouse visual cortex. *Nature neuroscience*,
551 16(12):1864–1869, 2013.
- 552 [40] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Church-
553 land. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuro-
554 science*, 22(10):1677–1686, 2019.
- 555 [41] Peter Dayan, Laurence F Abbott, et al. Theoretical neuroscience: computational and mathe-
556 matical modeling of neural systems. *Journal of Cognitive Neuroscience*, 15(1):154–155, 2003.
- 557 [42] Eugene M Izhikevich. *Dynamical systems in neuroscience*. MIT press, 2007.
- 558 [43] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
559 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

- 560 [44] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
561 A framework for parameter estimation and model selection from experimental data in systems
562 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 563 [45] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
564 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P
565 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
566 dynamics. *bioRxiv*, page 838383, 2019.
- 567 [46] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
568 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on*
569 *Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 570 [47] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
571 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
572 4248. PMLR, 2020.
- 573 [48] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
574 networks. *International Conference on Learning Representations*, 2017.
- 575 [49] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
576 *International Conference on Machine Learning*, 2015.
- 577 [50] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-
578 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,
579 2001.
- 580 [51] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
581 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
582 *Neuron*, 77(5):845–858, 2013.
- 583 [52] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
584 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 585 [53] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dy-
586 namics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–
587 1406, 2014.

- 588 [54] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding
589 and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.
- 590 [55] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and
591 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*,
592 115(3):1399–1409, 2016.
- 593 [56] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
594 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
595 operating regime of cortex. *bioRxiv*, 2020.
- 596 [57] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
597 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
598 routing. *bioRxiv*, page 245613, 2018.
- 599 [58] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
600 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 601 [59] Catherine Morris and Harold Lecar. Voltage oscillations in the barnacle giant muscle fiber.
602 *Biophysical journal*, 35(1):193–213, 1981.
- 603 [60] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
604 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 605 [61] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for
606 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,
607 2017.
- 608 [62] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
609 2009.
- 610 [63] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural
611 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 612 [64] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in
613 neurobiology*, 25:156–163, 2014.
- 614 [65] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current
615 opinion in neurobiology*, 46:1–6, 2017.

- 616 [66] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H
617 Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex
618 embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966,
619 2018.
- 620 [67] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian com-*
621 *putation*. CRC Press, 2018.
- 622 [68] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based infer-
623 ence. *Proceedings of the National Academy of Sciences*, 2020.
- 624 [69] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
625 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
626 592, 2009.
- 627 [70] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralin-
628 ear network: a unifying circuit motif underlying multi-input integration in sensory cortex.
629 *Neuron*, 85(2):402–417, 2015.
- 630 [71] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
631 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
632 5(10):793, 2004.
- 633 [72] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
634 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-*
635 *robiology*, 71(1):45–61, 2011.
- 636 [73] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-
637 *tex*: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 638 [74] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
639 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
640 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 641 [75] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn.
642 Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.
- 643 [76] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate
644 cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

- 645 [77] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
646 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
647 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 648 [78] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and
649 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,
650 2015.
- 651 [79] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-
652 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 653 [80] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
654 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
655 2003.
- 656 [81] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
657 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 658 [82] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
659 Edward Teller. Equation of state calculations by fast computing machines. *The journal of
660 chemical physics*, 21(6):1087–1092, 1953.
- 661 [83] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
662 1970.
- 663 [84] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
664 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
665 73(2):123–214, 2011.
- 666 [85] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-
667 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
668 820, 2011.
- 669 [86] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
670 free variational inference. In *Advances in Neural Information Processing Systems*, pages
671 5523–5533, 2017.

- 672 [87] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
673 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience*
674 *Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 675 [88] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
676 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience
677 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 678 [89] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
679 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
680 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 681 [90] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-
682 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,*
683 *Germany*, 2019.
- 684 [91] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
685 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
686 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 687 [92] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
688 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 689 [93] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
690 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 691 [94] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
692 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
693 2012.
- 694 [95] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
695 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
696 2018.
- 697 [96] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
698 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 699 [97] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
700 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially

- 701 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
702 1929, 2009.
- 703 [98] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
704 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 705 [99] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
706 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 707 [100] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-
708 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv*
709 preprint [arXiv:1912.02762](https://arxiv.org/abs/1912.02762), 2019.
- 710 [101] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
711 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 712 [102] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
713 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-*
714 *tion processing systems*, 29:4743–4751, 2016.
- 715 [103] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Internation-
716 al Conference on Learning Representations*, 2015.
- 717 [104] Emmanuel Klinger, Dennis Rickert, and Jan Hasenauer. pyabc: distributed, likelihood-free
718 inference. *Bioinformatics*, 34(20):3591–3593, 2018.
- 719 [105] David S Greenberg, Marcel Nonnenmacher, and Jakob H Macke. Automatic posterior trans-
720 formation for likelihood-free inference. *International Conference on Machine Learning*, 2019.

721 **5 Methods**

722 **5.1 Emergent property inference (EPI)**

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

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

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

743 Precisely, the emergent property statistics $f(\mathbf{x}; \mathbf{z})$ must have means $\boldsymbol{\mu}$ and variances σ^2 over the
744 EPI distribution of parameters and stochasticity of the data given the parameters. By defining
745 these means and variances over both levels of stochasticity – the inferred distribution and that of
746 the model – there is a fine degree of control over predictions made by the inferred parameters.

747 In EPI, deep probability distributions are optimized to learn the inferred distribution. In deep
748 probability distributions, a simple random variable $\mathbf{z}_0 \sim q_0(\mathbf{z}_0)$ is mapped deterministically via a
749 sequence of deep neural network layers (g_1, \dots, g_l) parameterized by weights and biases $\boldsymbol{\theta}$ to the

750 support of the distribution of interest:

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

751 Such deep probability distributions embed the inferred distribution in a deep network. Once opti-
752 mized, this deep network representation has remarkably useful properties: fast sampling, probability
753 evaluations, and also first- and second-order probability gradient evaluations.

754 By choosing a neural circuit model, often represented as a system of differential equations, we
755 implicitly define a model likelihood $p(\mathbf{x} | \mathbf{z})$, which may be unknown or intractable for our purposes.
756 Given this model choice and that of an emergent property \mathcal{X} , $q_{\theta}(\mathbf{z})$ is optimized via the neural
757 network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep variational
758 family \mathcal{Q} producing the emergent property \mathcal{X} :

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

759 Entropy is chosen as the normative selection principle to match that of Bayesian inference (see
760 Section 5.1.5). However, a key difference is that variational inference and other Bayesian methods
761 do not constrain the predictions of their inferred parameter distribution. This optimization is
762 executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [48].

763 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.
764 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-
765 ond, we describe the special class of deep probability distributions used in EPI called normalizing
766 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-
767 tion 13 (Section 5.1.3). Then, we demonstrate the details of this optimization in a toy example
768 (Section 5.1.4). Finally, we establish the known relationship between maximum entropy distribu-
769 tions and exponential families (Section 5.1.5), which is used to explain how EPI can be viewed as
770 a form of variational inference (Section 5.1.6).

771 5.1.1 Related approaches

772 When Bayesian inference problems lack conjugacy, scientists use approximate inference methods like
773 variational inference (VI) [81] and Markov chain Monte Carlo (MCMC) [82, 83]. After optimization,
774 variational methods return a parameterized posterior distribution, which we can analyze. Also, the
775 variational approximating distribution class is often chosen such that it permits fast sampling. In

776 contrast MCMC methods only produce samples from the approximated posterior distribution. No
777 parameterized distribution is estimated, and additional samples are always generated with the same
778 sampling complexity. Inference in models defined by systems of differential has been demonstrated
779 with MCMC [84], although this approach requires tractable likelihoods. Advancements have lever-
780 aged structure in stochastic differential equation models to improve likelihood approximations, thus
781 expanding the domain of applicable models [85].

782 Likelihood-free (or ‘simulation-based’) inference (LFI) [68] is model parameter inference in the
783 absence of a tractable likelihood function. The most prevalent approach to LFI is approximate
784 Bayesian computation [79], in which satisfactory parameter samples are kept from random prior
785 sampling according to a rejection heuristic. The obtained set of parameters do not have a prob-
786 abilities, and further insight about the model must be gained from examination of the parameter
787 set and their generated activity. Methodological advances to ABC methods have come through the
788 use of Markov chain Monte Carlo (MCMC-ABC) [80] and sequential Monte Carlo (SMC-ABC) [43]
789 sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still strug-
790 gles to scale in dimensionality (cf. Fig. 2). Furthermore, once a parameter set has been obtained by
791 SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again from scratch
792 with a new population of initialized particles to obtain additional samples.

793 For scientific model analysis, we seek a parameter distribution exhibiting the properties of a well-
794 chosen variational approximation: a parametric form conferring analytic calculations, and trivial
795 sampling time. For this reason, ABC and MCMC techniques are unattractive, since they only
796 produce a set of parameter samples and have unchanging sampling rate. EPI infers parameters
797 in likelihood-free models using the MEFN [48] algorithm using a deep variational approximation.
798 The deep neural network of EPI defines the parametric form of the distribution approximation.
799 Furthermore, the EPI distribution is constrained to produce an emergent property. In other words,
800 the summary statistics of the posterior predictive distribution are fixed to have certain first and
801 second moments. EPI optimization is enabled using stochastic gradient techniques in the spirit
802 of likelihood-free variational inference [86]. The analytic relationship between EPI and variational
803 inference is explained in Secton 5.1.6.

804 We note that, during our preparation and early presentation of this work [87, 88], another work
805 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
806 circuits ([45, 89, 90]). We are encouraged by this general problem being recognized by others in the
807 community, and we emphasize that these works offer complementary neuroscientific contributions

808 (different theoretical models of focus) and use different technical methodologies (ours is built on
809 our prior work [48], theirs similarly [91]).

810 The method EPI differs from SNPE in some key ways. SNPE belongs to a “sequential” class of
811 recently developed LFI methods in which two neural networks are used for posterior inference.
812 This first neural network is a deep probability distribution (normalizing flow) used to estimate the
813 posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE) or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural likelihood (SNL [46])). A
814 recent advance uses an unconstrained neural network to estimate the likelihood ratio (sequential
815 neural ratio estimation (SNRE [47])). In SNL and SNRE, MCMC sampling techniques are used to
816 obtain samples from the approximated posterior. This contrasts with EPI and SNPE, which use
817 deep probability distributions to model parameters, which facilitates immediate measurements of
818 sample probability, gradient, or Hessian for system analysis. The second neural network in this
819 sequential class of methods is the amortizer. This unconstrained deep network maps data \mathbf{x} (or
820 statistics $f(\mathbf{x}; \mathbf{z})$ or model parameters \mathbf{z} to the weights and biases of the first neural network. These
821 methods are optimized on a conditional density (or ratio) estimation objective. The data used to
822 optimize this objective are generated via an adaptive procedure, in which training data pairs $(\mathbf{x}_i,$
823 $\mathbf{z}_i)$ become sequentially closer to the true data and posterior.

824 The approximating fidelity of the deep probability distribution in sequential approaches is optimized
825 to generalize across the training distribution of the conditioning variable. This generalization prop-
826 erty of the sequential methods can reduce the accuracy at the singular posterior of interest. Whereas
827 in EPI, the entire expressivity of the deep probability distribution is dedicated to learning a single
828 distribution as well as possible. Amortization is not possible in EPI, since EPI learns an expo-
829 nential family distribution parameterized by its mean (see Section 5.1.5). Since EPI distributions
830 are defined by the mean μ of their statistics, there is the well-known inverse mapping problem of
831 exponential families [92] that prohibits an amortization based approach. However, we have shown
832 that the same two-network architecture of the sequential LFI methods can be used for amortized
833 inference in intractable exponential family posteriors using their natural parameterization [93].

834 Finally, one important differentiating factor between EPI and sequential LFI methods is that EPI
835 leverages gradients $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ during optimization. These gradients can improve convergence time
836 and scalability, as we have shown on an example conditioning low-rank RNN connectivity on the
837 property of stable amplification (see Section 3.3). With EPI, we prove out the suggestion that a
838 deep inference technique can improve efficiency by leveraging these model gradients when they are
839 tractable. Sequential LFI techniques may be better suited for scientific problems where $\nabla_{\mathbf{z}} f(\mathbf{x}; \mathbf{z})$ is

840 intractable or unavailable: when there is a nondifferentiable model or it requires lengthy simulations.
841 However, the sequential LFI techniques cannot constrain the predictions of the inferred distribution
842 in the manner of EPI.

843 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in
844 natural models. Around a point, one can measure the Jacobian. One approach that scales well is
845 EAR [94]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [95] and
846 its stochastic adaptation [96]. Casting identifiability as a statistical estimation problem, the profile
847 likelihood can assess via iterated optimization while holding parameters fixed [97]. An exciting
848 recent method is capable of recovering the functional form of such unidentifiabilities away from a
849 point by following degenerate dimensions of the fisher information matrix [98]. Global structural
850 non-identifiabilities can be found for models with polynomial or rational dynamics equations using
851 DAISY [99]. With EPI, we have all the benefits given by a statistical inference method plus the
852 ability to query the first- or second-order gradient of the probability of the inferred distribution at
853 any chosen parameter value. The second-order gradient of the log probability (the Hessian), which
854 is directly afforded by EPI distributions, produces salient information about parametric sensitivity
855 of the emergent property. For example, the eigenvector with most negative eigenvalue of the Hessian
856 shows parametric combinations away from a parameter choice that decrease the in EPI distribution
857 probability the fastest. We refer to this eigenvector as the sensitivity dimension, and it is used to
858 generate scientific insight about a model of superior colliculus connectivity (see Section 3.5).

859 **5.1.2 Deep probability distributions and normalizing flows**

860 Deep probability distributions are comprised of multiple layers of fully connected neural networks
861 (Equation 12). When each neural network layer is restricted to be a bijective function, the sample
862 density can be calculated using the change of variables formula at each layer of the network. For
863 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

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

864 However, this computation has cubic complexity in dimensionality for fully connected layers. By
865 restricting our layers to normalizing flows [49, 100] – bijective functions with fast log determinant
866 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log
867 probability calculation confers efficient optimization of the maximum entropy objective (see Section
868 5.1.3). We use the Real NVP [60] normalizing flow class, because its coupling architecture confers

both fast sampling (forward) and fast log probability evaluation (backward). Fast probability evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout parameter space. Glow permutations were used in between coupling stages [101]. This is in contrast to autoregressive architectures [61, 102], in which only one of the forward or backward passes can be efficient. In this work, normalizing flows are used as flexible posterior approximations $q_{\theta}(\mathbf{z})$ having weights and biases θ . We specify the architecture used in each application by the number of Real-NVP affine coupling stages, and the number of neural network layers and units per layer of the conditioning functions.

5.1.3 Augmented Lagrangian optimization

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

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

where average constraint violations $R(\theta) = \mathbb{E}_{\mathbf{z} \sim q_{\theta}} [\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [T(\mathbf{x}; \mathbf{z}) - \mu_{\text{opt}}]]$, $\eta_{\text{opt}} \in \mathbb{R}^m$ are the Lagrange multipliers where $m = |\mu_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})| = 2|f(\mathbf{x}; \mathbf{z})|$, and c is the penalty coefficient. The sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter μ_{opt} are determined by the means μ and variances σ^2 of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 13 (see Section 5.1.6). Specifically, $T(\mathbf{x}; \mathbf{z})$ is a concatenation of the first and second moments, μ_{opt} is a concatenation of μ and σ^2 (see section 5.1.5), and the Lagrange multipliers are closely related to the natural parameters η of exponential families (see Section 5.1.5). Weights and biases θ of the deep probability distribution are optimized according to Equation 15 using the Adam optimizer with learning rate 10^{-3} [103].

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

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

Thus, the gradient of the entropy of the deep probability distribution can be estimated as an average with respect to the base distribution z_0 :

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

The lagrangian parameters η_{opt} are initialized to zero and adapted following each augmented Lagrangian epoch, which is a period of optimization with fixed (η_{opt}, c) for a given number of

895 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
896 after each epoch based on constraint error reduction. The penalty coefficient is updated based
897 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
898 $\mathbb{E}[|R(\boldsymbol{\theta}_{k+1})|] > \gamma \mathbb{E}[|R(\boldsymbol{\theta}_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
899 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
900 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
901 according to application.

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

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

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

923 When assessing the suitability of EPI for a particular modeling question, there are some important
924 technical considerations. First and foremost, as in any optimization problem, the defined emergent
925 property should always be appropriately conditioned (constraints should not have wildly different
926 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution

grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no parameter set producing the emergent property, and EPI optimization will fail (appropriately). Next, one should consider the computational cost of the gradient calculations. In the best circumstance, there is a simple, closed form expression (e.g. Section 5.2.2) for the emergent property statistic given the model parameters. On the other end of the spectrum, many forward simulation iterations may be required before a high quality measurement of the emergent property statistic is available (e.g. Section 5.2.1). In such cases, backpropagating gradients through the SDE evolution will be expensive.

5.1.4 Example: 2D LDS

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

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

with

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

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

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

947

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

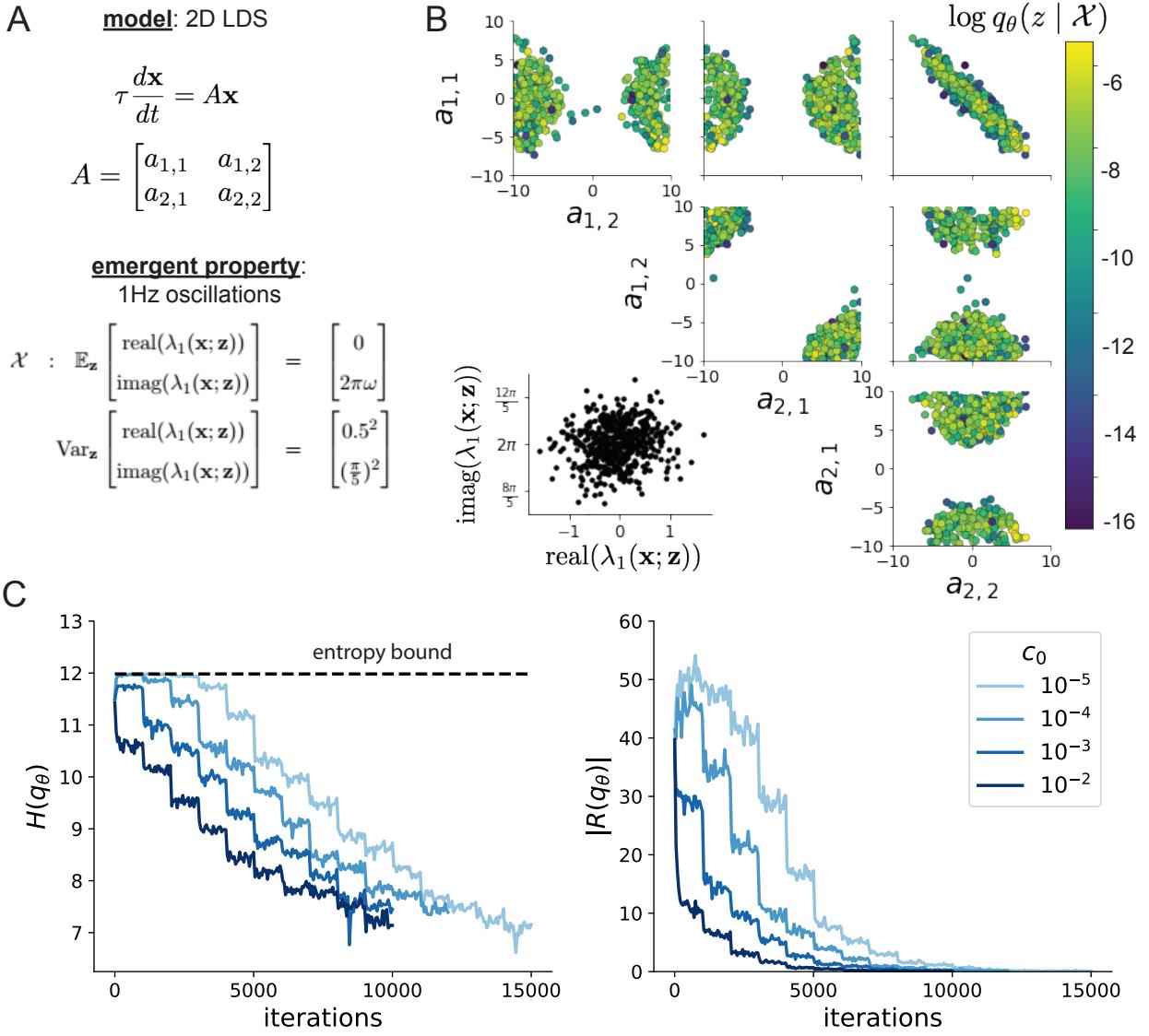


Figure 6: (LDS1): **A.** Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. **B.** The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance.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

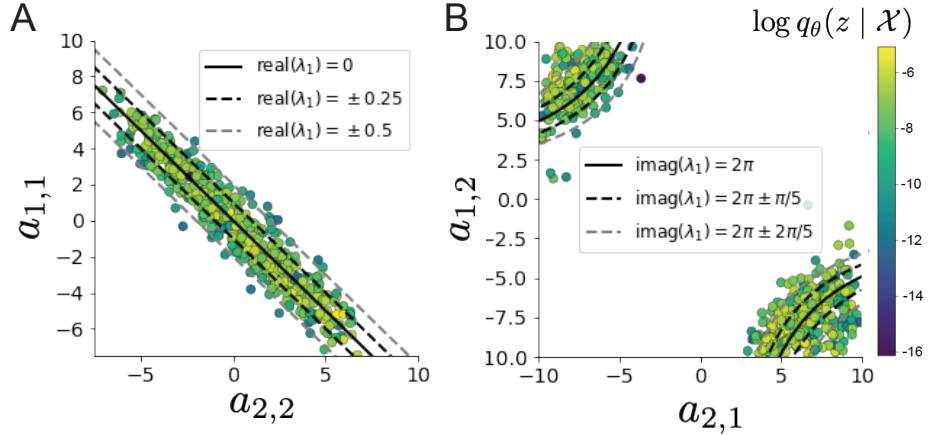


Figure 7: (LDS2): **A.** Probability contours in the a_1 - a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. **B.** Probability contours in the a_2 - a_3 plane were derived from the emergent property statistic of oscillation frequency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

950 the quadratic formula:

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

951 Importantly, even though $\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})}[T(\mathbf{x})]$ is calculable directly via a closed form function and
952 does not require simulation, we cannot derive the distribution q_θ^* directly. This fact is due to the
953 formally hard problem of the backward mapping: finding the natural parameters η from the mean
954 parameters μ of an exponential family distribution [92]. Instead, we used EPI to approximate this
955 distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two
956 neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped
957 onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

958 Even this relatively simple system has nontrivial (though intuitively sensible) structure in the
959 parameter distribution. To validate our method, we analytically derived the contours of the prob-
960 ability density from the emergent property statistics and values. In the a_1 - a_4 plane, the black
961 line at $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.25$,
962 and the dotted gray line at twice the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 0.5$ follow the contour
963 of probability density of the samples (Fig. S2A). The distribution precisely reflects the desired
964 statistical constraints and model degeneracy in the sum of a_1 and a_4 . Intuitively, the parameters
965 equivalent with respect to emergent property statistic $\text{real}(\lambda_1)$ have similar log densities.

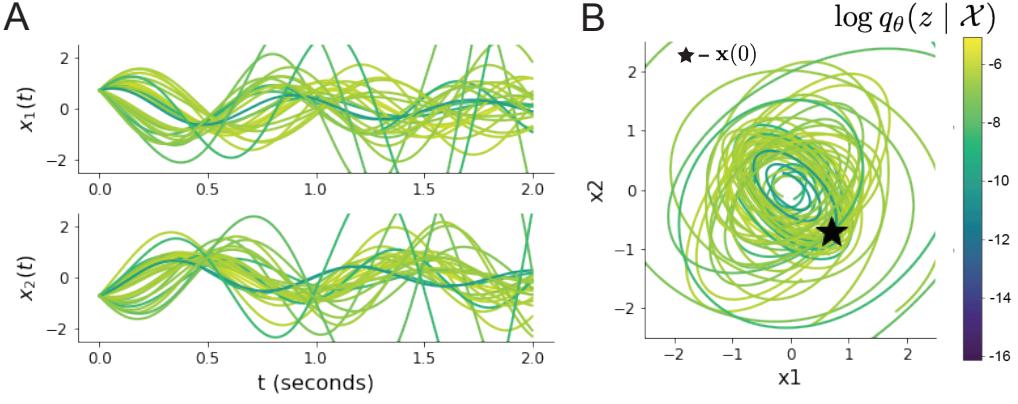


Figure 8: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\theta}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. **A.** Each dimension of the simulated trajectories throughout time. **B.** The simulated trajectories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

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1 a_4 - a_2 a_3}{\tau}}, & \text{if } a_1 a_4 < a_2 a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (22)$$

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

$$\text{imag}(\lambda_1)^2 = a_1 a_4 - a_2 a_3 \quad (23)$$

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

979 **5.1.5 Maximum entropy distributions and exponential families**

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

$$\begin{aligned} p^*(\mathbf{z}) &= \operatorname{argmax}_{p \in \mathcal{P}} H(p(\mathbf{z})) \\ \text{s.t. } \mathbb{E}_{\mathbf{z} \sim p}[T(\mathbf{z})] &= \boldsymbol{\mu}_{\text{opt}}. \end{aligned} \quad (24)$$

982 will have probability density in the exponential family:

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

983 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
 984 formally hard to identify except in special cases [92].

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

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

987 As a general maximum entropy distribution (Equation 24), the sufficient statistics vector contains
 988 both first and second order moments of $f(\mathbf{x}; \mathbf{z})$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (27)$$

989 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (28)$$

990 **5.1.6 EPI as variational inference**

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

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

995 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
 996 However, conjugacy is rare in practice, and alternative methods, such as variational inference [81],
 997 are utilized.

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

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

1000 The KL divergence can be written in terms of entropy of the variational approximation:

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

1001

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

1002 Since the marginal distribution of the data $p(\mathbf{x})$ (or ‘‘evidence’’) is independent of θ , variational
 1003 inference is executed by optimizing the remaining expression. This is usually framed as maximizing
 1004 the evidence lower bound (ELBO)

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

1005 Now, consider the setting where we have chosen a uniform prior, and stipulate a mean-field gaussian
 1006 likelihood on a chosen statistic of the data $f(\mathbf{x}; \mathbf{z})$

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

1007 where $\Sigma_f = \operatorname{diag}(\sigma_f^2)$. The log likelihood is then proportional to a dot product of the natural
 1008 parameter of this mean-field gaussian distribution and the first and second moment statistics.

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

1009 where

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

1010

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

1011 The variational objective is then

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

1012 Comparing this to the Lagrangian objective (without augmentation) of EPI, we see they are the
 1013 same

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

1014 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
 1015 (Equation 27). Thus, EPI is implicitly executing variational inference with a uniform prior and a
 1016 mean-field gaussian likelihood on the emergent property statistics. The mean and variances of the
 1017 mean-field gaussian likelihood are predicated by $\boldsymbol{\eta}_{\text{opt}}$ (Equations 36 and 38), which is adapted after
 1018 each EPI optimization epoch based on \mathcal{X} (see Section 5.1.3). In EPI, the inferred distribution is
 1019 not conditioned on a finite dataset as in variational inference, but rather the emergent property
 1020 \mathcal{X} dictates the likelihood parameterization such that the inferred distribution will produce the
 1021 emergent property. As a note, we could not simply choose $\boldsymbol{\mu}_f$ and $\boldsymbol{\sigma}_f$ directly from the outset, since
 1022 we do not know which of these choices will produce the emergent property \mathcal{X} , which necessitates
 1023 the EPI optimization routine that adapts $\boldsymbol{\eta}_{\text{opt}}$. Accordingly, we replace the notation of $p(\mathbf{z} | \mathbf{x})$
 1024 with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys emergent property \mathcal{X} (see Section
 1025 5.1).

1026 5.2 Theoretical models

1027 In this study, we used emergent property inference to examine several models relevant to theoretical
 1028 neuroscience. Here, we provide the details of each model and the related analyses.

1029 5.2.1 Stomatogastric ganglion

1030 We analyze how the parameters $\mathbf{z} = [g_{\text{el}}, g_{\text{synA}}]$ govern the emergent phenomena of intermediate
 1031 hub frequency in a model of the stomatogastric ganglion (STG) [51] shown in Figure 1A with
 1032 activity $\mathbf{x} = [x_{\text{f1}}, x_{\text{f2}}, x_{\text{hub}}, x_{\text{s1}}, x_{\text{s2}}]$, using the same hyperparameter choices as Gutierrez et al.
 1033 Each neuron's membrane potential $x_{\alpha}(t)$ for $\alpha \in \{\text{f1}, \text{f2}, \text{hub}, \text{s1}, \text{s2}\}$ is the solution of the following
 1034 stochastic differential equation:

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

1035 The input current of each neuron is the sum of the leak, calcium, potassium, hyperpolarization,
 1036 electrical and synaptic currents as well as gaussian noise dB . Each current component is a function

1037 of all membrane potentials and the conductance parameters \mathbf{z} .

1038 The capacitance of the cell membrane was set to $C_m = 1nF$. Specifically, the currents are the
 1039 difference in the neuron's membrane potential and that current type's reversal potential multiplied
 1040 by a conductance:

$$1041 \quad h_{leak}(\mathbf{x}; \mathbf{z}) = g_{leak}(x_\alpha - V_{leak}) \quad (41)$$

$$1042 \quad h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{post} - x_\alpha^{pre}) \quad (42)$$

$$1043 \quad h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (43)$$

$$1044 \quad h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (44)$$

$$1045 \quad h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (45)$$

$$1046 \quad h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (46)$$

1046 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 1047 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$, g_{Ca} ,
 1048 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 1049 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
 1050 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
 1051 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

1052 Furthermore, the Calcium, Potassium, and hyperpolarization channels have time-dependent gating
 1053 dynamics dependent on steady-state gating variables M_∞ , N_∞ and H_∞ , respectively:

$$1054 \quad M_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_1}{v_2} \right) \right) \quad (47)$$

$$1055 \quad \frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (48)$$

$$1056 \quad N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_3}{v_4} \right) \right) \quad (49)$$

$$1057 \quad \lambda_N = \phi_N \cosh \left(\frac{x_\alpha - v_3}{2v_4} \right) \quad (50)$$

$$1058 \quad \frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (51)$$

$$1059 \quad H_\infty = \frac{1}{1 + \exp \left(\frac{x_\alpha + v_5}{v_6} \right)} \quad (52)$$

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

1060 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 1061 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

1062 Finally, there is a synaptic gating variable as well:

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

1063 When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear
 1064 dynamical system. Gaussian noise $d\mathbf{B}$ of variance $(1 \times 10^{-12})^2 \text{ A}^2$ makes the model stochastic, and
 1065 introduces variability in frequency at each parameterization \mathbf{z} .

1066 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for
 1067 $T = 300$ time steps of $dt = 25\text{ms}$. The chosen dt and T were the most computationally convenient
 1068 choices yielding accurate frequency measurement. We used a basis of complex exponentials with
 1069 frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

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

1070 To measure spiking frequency, we processed simulated membrane potentials with a relu (spike
 1071 extraction) and low-pass filter with averaging window of size 20, then took the frequency with the
 1072 maximum absolute value of the complex exponential basis coefficients of the processed time-series.
 1073 The first 20 temporal samples of the simulation are ignored to account for initial transients.

1074 To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in$
 1075 $\mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in$
 1076 $\{\text{f1}, \text{f2}, \text{hub}, \text{s1}, \text{s2}\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

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

1077 where $i = [0, 1, \dots, 100]$. The frequency is then calculated as

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

1078 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
 1079 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
 1080 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be $(0.025\text{Hz})^2$ to
 1081 capture variation in frequency between 0.5Hz and 0.6Hz (Equation 4). As a maximum entropy dis-
 1082 tribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments of the hub neuron frequency

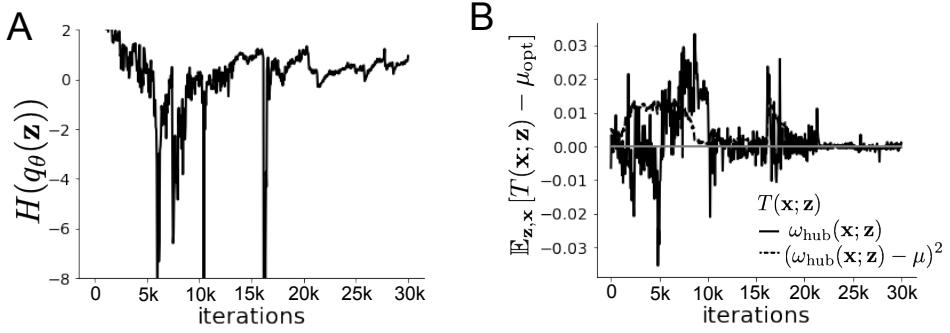


Figure 9: (STG1): EPI optimization of the STG model producing network syncing. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

1083 (as in Equations 27 and 28)

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

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

1084 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
1085 epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

1086 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-
1087 layer neural networks of 25 units per layer. The normalizing flow architecture mapped $z_0 \sim \mathcal{N}(\mathbf{0}, I)$
1088 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
1089 returned by a preliminary ABC search. We did not include $g_{\text{synA}} < 0.01$, for numerical stability.
1090 EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an
1091 augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The distribution
1092 shown is that of the architecture converging with criteria $N_{\text{test}} = 100$ at greatest entropy across
1093 random seeds.

1094 We calculated the Hessian at the mode of the inferred EPI distribution. The Hessian of a probability
1095 model is the second order gradient of the log probability density $\log q_{\boldsymbol{\theta}}(\mathbf{z})$ with respect to the
1096 parameters \mathbf{z} : $\frac{\partial^2 \log q_{\boldsymbol{\theta}}(\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^T}$. With EPI, we can examine the Hessian, which is analytically available
1097 throughout distribution, to indicate the dimensions of parameter space that are sensitive (strongly
1098 negative eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the
1099 negative eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the

emergent property produced. In Figure 1D, the eigenvectors of the Hessian v_1 (solid) and v_2 (dashed) are shown evaluated at the mode of the distribution. The length of the arrows is inversely proportional to the square root of absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$. Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D parameter space are chosen arbitrarily.

5.2.2 Scaling EPI for stable amplification in RNNs

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

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

In this analysis, we inferred connectivity parameterizations $\mathbf{z} = [\mathbf{u}_1^\top, \mathbf{u}_2^\top, \mathbf{v}_1^\top, \mathbf{v}_2^\top]^\top \in [-1, 1]^{(4N)}$ that produced stable amplification using EPI, SMC-ABC [43], and SNPE [45] (see Section Related Methods).

For this RNN model to be stable, all real eigenvalues of W must be less than 1: $\text{real}(\lambda_1) < 1$, where λ_1 denotes the greatest real eigenvalue of W . For a stable RNN to amplify at least one input pattern, the symmetric connectivity $W^s = \frac{W+W^\top}{2}$ must have an eigenvalue greater than 1: $\lambda_1^s > 1$, where λ^s is the maximum eigenvalue of W^s . These two conditions are necessary and sufficient for stable amplification in RNNs [63]. We defined the emergent property of stable amplification with means of these eigenvalues (0.5 and 1.5, respectively) that satisfy these conditions. To complete the emergent property definition, we chose variances (0.25^2) about those means such that samples rarely violate the eigenvalue constraints. In terms of the EPI optimization variables, this is written as

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

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

1123 Gradients of maximum eigenvalues of Hermitian matrices like W^s are available with modern auto-
 1124 automatic differentiation tools. To differentiate through the $\text{real}(\lambda_1)$, we solved the following equation
 1125 for eigenvalues of rank-2 matrices using the rank reduced matrix $W^r = V^\top U$

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

1126 For EPI in Fig. 2, we used a real NVP architecture with three coupling layers of affine transfor-
 1127 mations parameterized by two-layer neural networks of 100 units per layer. The initial distribution
 1128 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
 1129 an augmented Lagrangian coefficient of $c_0 = 10^3$, a batch size $n = 200$, $\beta = 4$, and chose to use
 1130 500 iterations per augmented Lagrangian epoch and emergent property constraint convergence was
 1131 evaluated at $N_{\text{test}} = 200$ (Fig. 2B blue line, and Fig. 2C-D blue).

1132 We compared EPI to two alternative likelihood-free inference (LFI) techniques, since the likelihood
 1133 of these eigenvalues given \mathbf{z} is not available. Approximate Bayesian computation (ABC) [79] is a
 1134 rejection sampling technique for obtaining sets of parameters \mathbf{z} that produce activity \mathbf{x} close to some
 1135 observed data \mathbf{x}_0 . Sequential Monte Carlo approximate Bayesian computation (SMC-ABC) is the
 1136 state-of-the-art ABC method, which leverages SMC techniques to improve sampling speed. We ran
 1137 SMC-ABC with the pyABC package [104] to infer RNNs with stable amplification: connectivities
 1138 having eigenvalues within an ϵ -defined l_2 distance of

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

1139 SMC-ABC was run with a uniform prior over $\mathbf{z} \in [-1, 1]^{(4N)}$, a population size of 1,000 particles
 1140 with simulations parallelized over 32 cores, and a multivariate normal transition model.

1141 SNPE, the next LFI approach in our comparison, is far more similar to EPI. Like EPI, SNPE
 1142 treats parameters in mechanistic models with deep probability distributions, yet the two learning
 1143 algorithms are categorically different. SNPE uses a two-network architecture to approximate the
 1144 posterior distribution of the model conditioned on observed data \mathbf{x}_0 . The amortizing network maps
 1145 observations \mathbf{x}_i to the parameters of the deep probability distribution. The weights and biases of the
 1146 parameter network are optimized by sequentially augmenting the training data with additional pairs
 1147 $(\mathbf{z}_i, \mathbf{x}_i)$ based on the most recent posterior approximation. This sequential procedure is important
 1148 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.
 1149 For the deep probability distribution architecture, we chose a masked autoregressive flow with affine
 1150 couplings (the default choice), three transforms, 50 hidden units, and a normalizing flow mapping

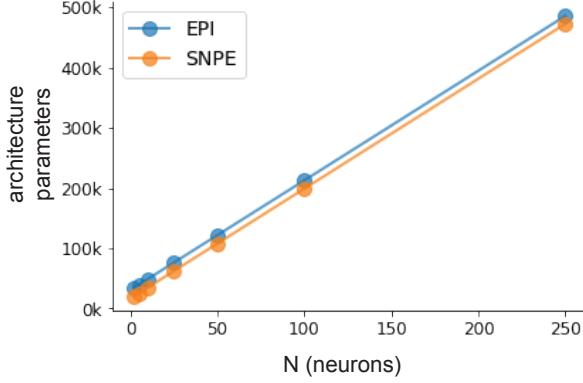


Figure 10: (RNN1): Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size (N).

1151 to the support as in EPI. This architectural choice closely tracked the size of the architecture used
 1152 by EPI (Fig. 10). As in SMC-ABC, we ran SNPE with $\mathbf{x}_0 = \mu$. All SNPE optimizations were
 1153 run for a limit of 1.5 days on a Tesla V100 GPU, or until two consecutive rounds resulted in a
 1154 validation log probability lower than the maximum observed for that random seed.

1155 To clarify the difference in objectives of EPI and SNPE, we show their results on RNN models
 1156 with different numbers of neurons N and random strength g . The parameters inferred by EPI
 1157 consistently produces the same mean and variance of $\text{real}(\lambda_1)$ and λ_1^s , while those inferred by
 1158 SNPE change according to the model definition (Fig. 11A). For $N = 2$ and $g = 0.01$, the SNPE
 1159 posterior has greater concentration in eigenvalues around \mathbf{x}_0 than at $g = 0.1$, where the model has
 1160 greater randomness (Fig. 11B top, orange). At both levels of g when $N = 2$, the posterior of SNPE
 1161 has lower entropy than EPI at convergence (Fig. 11B top). However at $N = 10$, SNPE results in
 1162 a predictive distribution of more widely dispersed eigenvalues (Fig. 11A bottom), and an inferred
 1163 posterior with greater entropy than EPI (Fig. 11B bottom). We highlight these differences not
 1164 to focus on an insightful trend, but to emphasize that these methods optimize different objectives
 1165 with different implications.

1166 Note that SNPE converges when it's validation log probability has saturated after several rounds
 1167 of optimization (Fig. 11C), and that EPI converges after several epochs of its own optimization
 1168 to enforce the emergent property constraints (Fig. 11D blue). Importantly, as SNPE optimizes
 1169 its posterior approximation, the predictive means change, and at convergence may be different
 1170 than \mathbf{x}_0 (Fig. 11D orange, left). It is sensible to assume that predictions of a well-approximated
 1171 SNPE posterior should closely reflect the data on average (especially given a uniform prior and

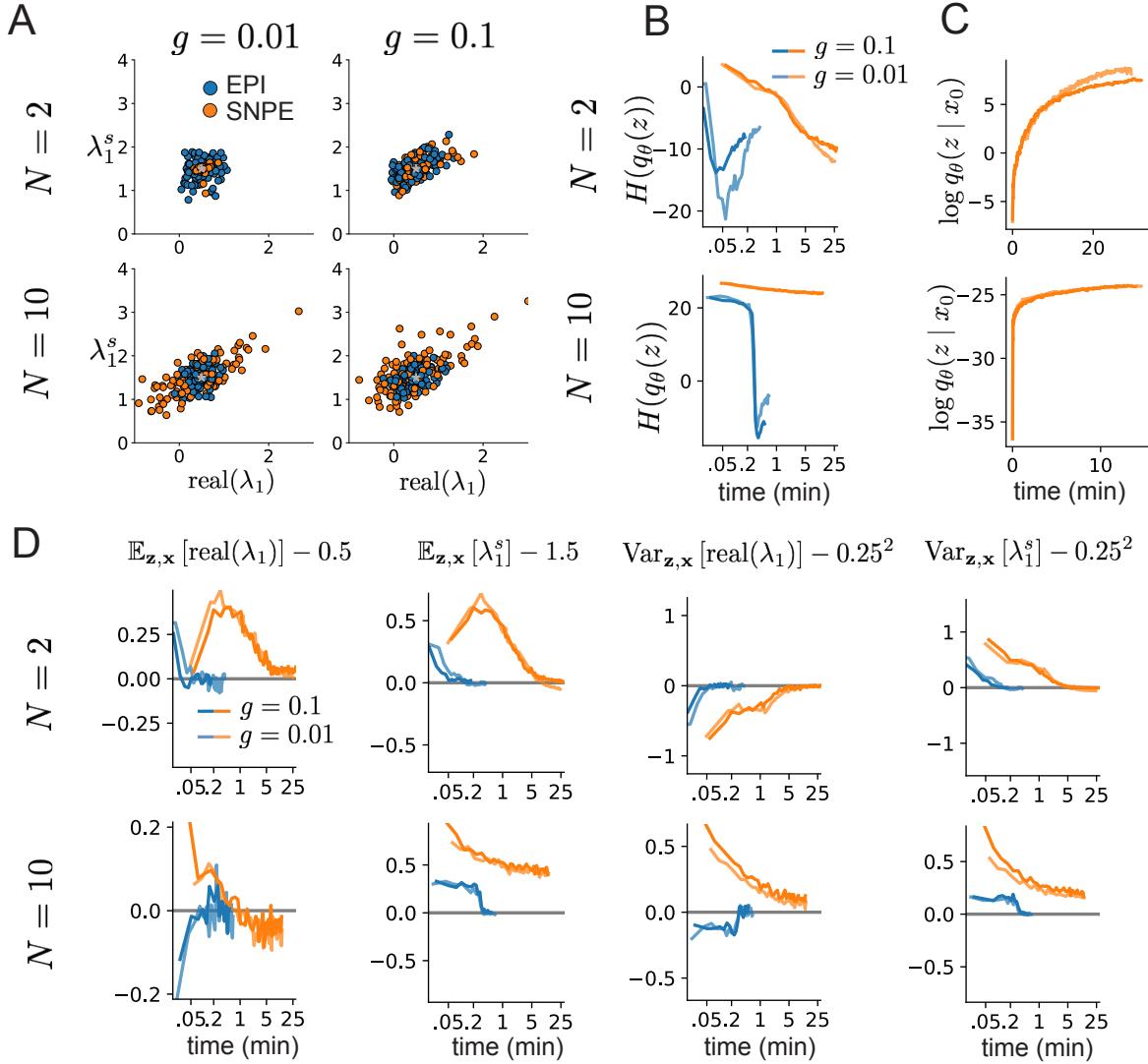


Figure 11: (RNN2): Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A.** Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with $N = 2$ (top), $N = 10$ (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 low degree of stochasticity), however this is not a given. Furthermore, no aspect of the SNPE
1173 optimization controls the variance of the predictions (Fig. 11D orange, right).

1174 To compare the efficiency of these algorithms for inferring RNN connectivity distributions producing
1175 stable amplification, we develop a convergence criteria that can be used across methods. While EPI
1176 has its own hypothesis testing convergence criteria for the emergent property, it would not make
1177 sense to use this criteria on SNPE and SMC-ABC which do not constrain the means and variances
1178 of their predictions. Instead, we consider EPI and SNPE to have converged after completing its
1179 most recent optimization epoch (EPI) or round (SNPE) in which the distance

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

1180 is less than 0.5. We consider SMC-ABC to have converged once the population produces samples
1181 within the $\epsilon = 0.5$ ball ensuring stable amplification.

1182 When assessing the scalability of SNPE, it is important to check that alternative hyperparameter-
1183izations could not yield better performance. Key hyperparameters of the SNPE optimization are
1184 the number of simulations per round n_{round} , the number of atoms used in the atomic proposals of
1185 the SNPE-C algorithm [105], and the batch size n . To match EPI, we used a batch size of $n = 200$
1186 for $N \leq 25$, however we found $n = 1,000$ to be helpful for SNPE in higher dimensions. While
1187 $n_{\text{round}} = 1,000$ yielded SNPE convergence for $N \leq 25$, we found that a substantial increase to
1188 $n_{\text{round}} = 25,000$ yielded more consistent convergence at $N = 50$ (Fig. 12A). By increasing n_{round} ,
1189 we also necessarily increase the duration of each round. At $N = 100$, we tried two hyperparameter
1190 modifications. As suggested in [105], we increased n_{atom} by an order of magnitude to improve
1191 gradient quality, but this had little effect on the optimization (much overlap between same random
1192 seeds) (Fig. 12B). Finally, we increased n_{round} by an order of magnitude, which yielded convergence
1193 in one case, but no others. We found no way to improve the convergence rate of SNPE without
1194 making more aggressive hyperparameter choices requiring high numbers of simulations.

1195 In Figure 2C-D, we show samples from the random seed resulting in emergent property convergence
1196 at greatest entropy (EPI), the random seed resulting in greatest validation log probability (SNPE),
1197 and the result of all converged random seeds (SMC).

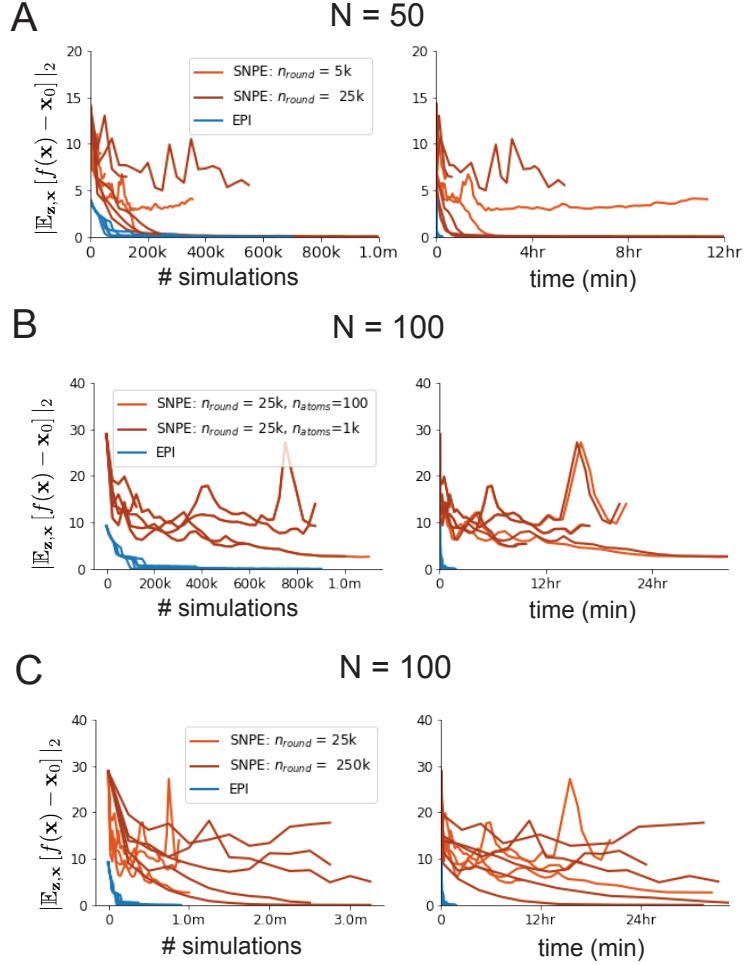


Figure 12: (RNN3): SNPE convergence was enabled by increasing n_{round} , not n_{atom} . **A.** Difference of mean predictions \mathbf{x}_0 throughout optimization at $N = 50$ with by simulation count (left) and wall time (right) of SNPE with $n_{\text{round}} = 5,000$ (light orange), SNPE with $n_{\text{round}}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

1198 **5.2.3 Primary visual cortex**

1199 In the stochastic stabilized supralinear network [77], population rate responses \mathbf{x} to input \mathbf{h} , recur-
 1200 rent input $W\mathbf{x}$ and slow noise ϵ are governed by

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

1201 where the noise is an Ornstein-Uhlenbeck process $\epsilon \sim OU(\tau_{\text{noise}}, \sigma)$

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

1202 with $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$. The noisy process is parameterized as

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

1203 so that σ parameterizes the variance of the noisy input in the absence of recurrent connectivity
 1204 ($W = \mathbf{0}$). As contrast increases, input to the E- and P-populations increases relative to a baseline
 1205 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
 1206 deterministic V1 circuit model [56]

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

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

1207 and

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

1208 To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equiv-
 1209 alence class

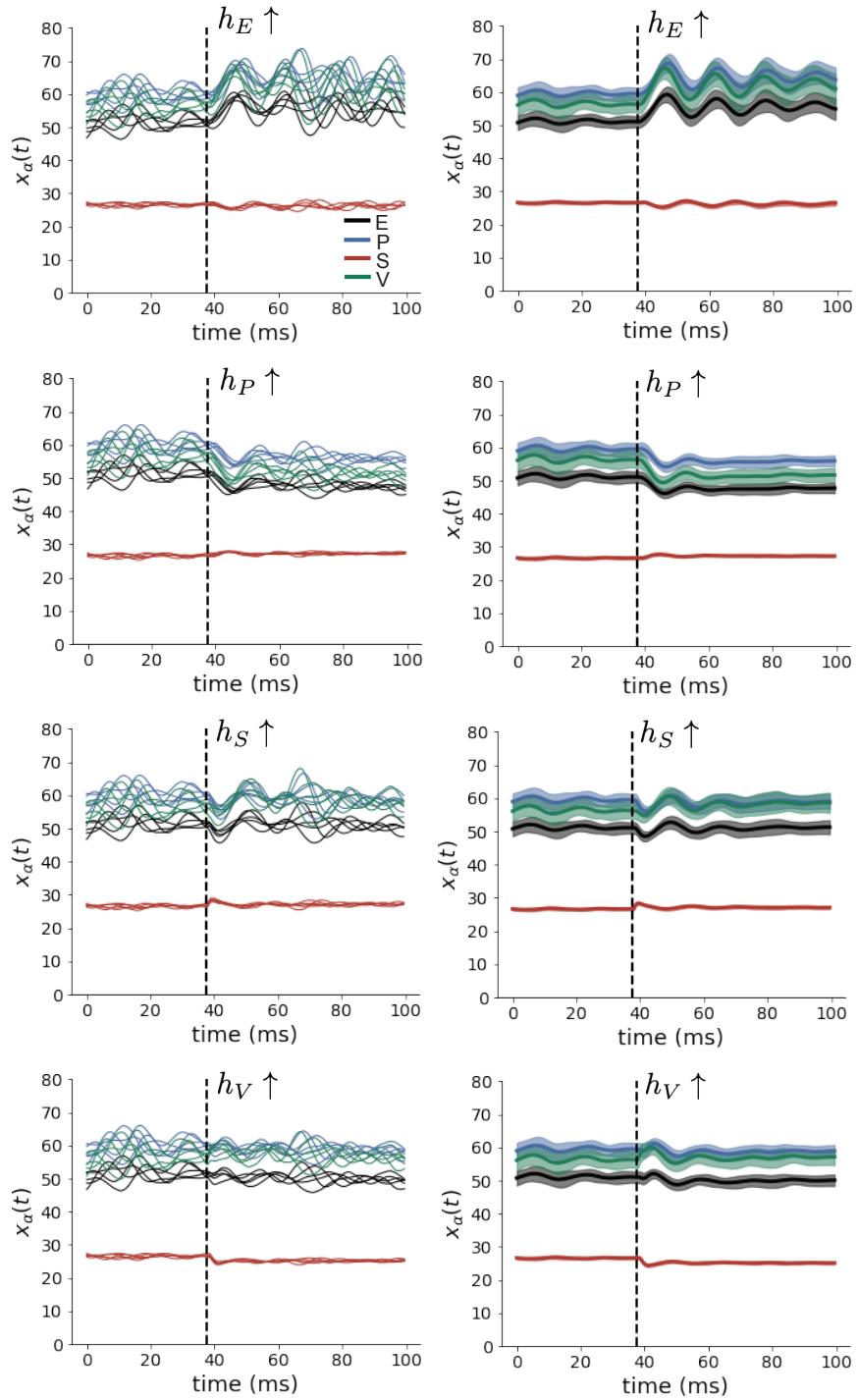


Figure 13: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

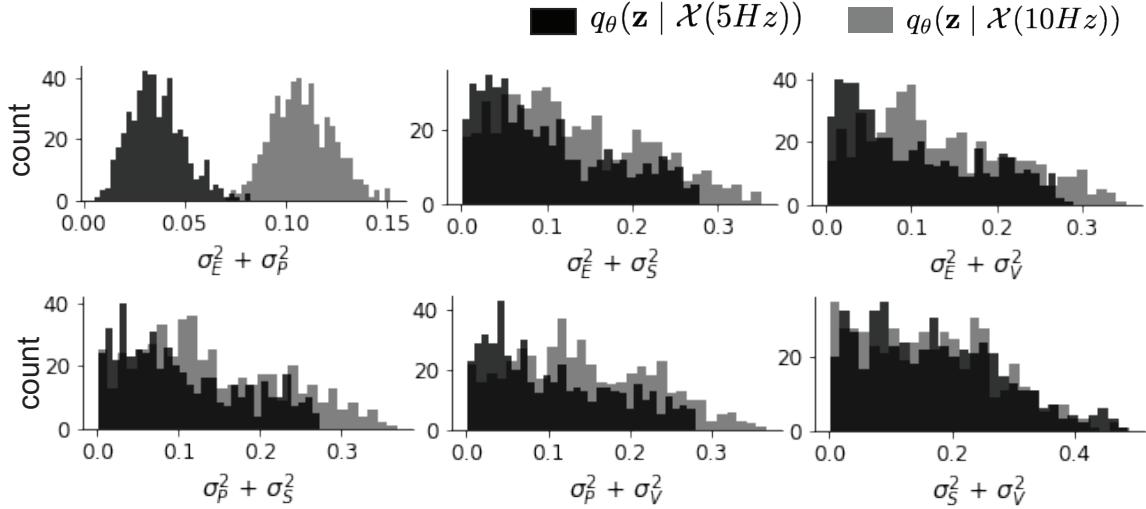


Figure 14: (V1 2) EPI predictive distributions of the sum of squares of each pair of noise parameters.

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

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

1210 and

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

1211 Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5\text{ms}$ from an initial condition
 1212 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
 1213 as the square root of the temporal variance from $t_{ss} = 75\text{ms}$ to $Tdt = 100\text{ms}$ averaged over 100

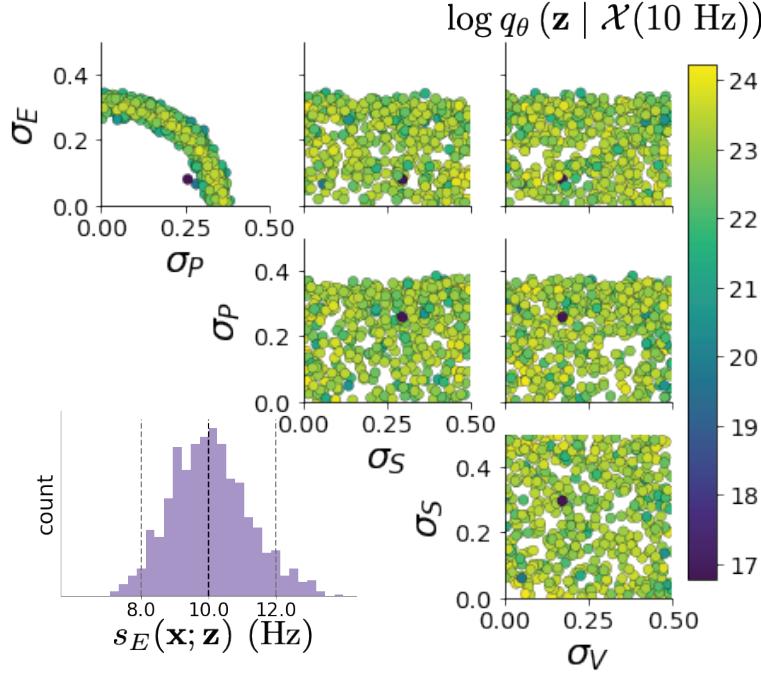


Figure 15: (V1 3) EPI inferred distribution for $\mathcal{X}(10 \text{ Hz})$.

1214 independent trials.

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

1215 For EPI in Fig 3D-E, we used a real NVP architecture with three Real NVP coupling layers
 1216 and two-layer neural networks of 50 units per layer. The normalizing flow architecture mapped
 1217 $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ 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
 1218 different random seeds for architecture initialization θ with an augmented Lagrangian coefficient of
 1219 $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the architectures
 1220 converging with criteria $N_{\text{test}} = 100$ at greatest entropy across random seeds.

1221 In Fig. 3E, we visualize the modes of $q_\theta(\mathbf{z} | \mathcal{X})$ throughout the σ_E - σ_P marginal. Specifically, we
 1222 calculated

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

1223 At each mode \mathbf{z}^* , we calculated the Hessian and visualized the sensitivity dimension in the direction
 1224 of positive σ_E .

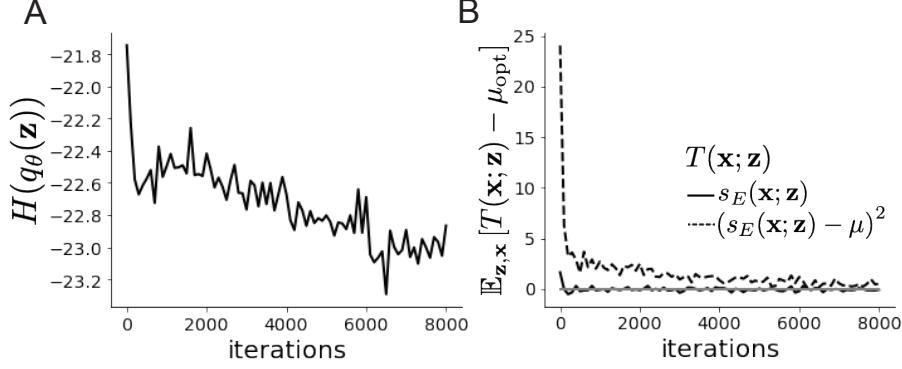


Figure 16: (V1 4) Optimization for V1

1225 **5.2.4 Primary visual cortex: challenges to analysis**

1226 TODO Agostina and I are putting this together now.

1227 **5.2.5 Superior colliculus**

1228 The ability to switch between two separate tasks throughout randomly interleaved trials, or “rapid
1229 task switching,” has been studied in rats, and midbrain superior colliculus (SC) has been shown to
1230 play an important role in this computation [78]. Neural recordings in SC exhibited two populations of
1231 neurons that simultaneously represented both task context (Pro or Anti) and motor response (con-
1232 tralateral or ipsilateral to the recorded side), which led to the distinction of two functional classes:
1233 the Pro/Contra and Anti/Ipsi neurons [57]. Given this evidence, Duan et al. proposed a model
1234 with four functionally-defined neuron-type populations: two in each hemisphere corresponding to
1235 the Pro/Contra and Anti/Ipsi populations. We study how the connectivity of this neural circuit
1236 governs rapid task switching ability.

1237 The four populations of this model are denoted as left Pro (LP), left Anti (LA), right Pro (RP)
1238 and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh \left(\frac{u_\alpha - a}{b} \right) + \frac{1}{2} \right), \quad (77)$$

1239 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-

1240 earity. We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (78)$$

1241 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (79)$$

1242 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2 . These
1243 hyperparameter values are motivated by modeling choices and results from [57].

1244 The weight matrix has 4 parameters for self sW , vertical vW , horizontal hW , and diagonal dW
1245 connections:

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (80)$$

1246 We study the role of parameters $\mathbf{z} = [sW, vW, hW, dW]^\top$ in rapid task switching.

1247 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (81)$$

1248 There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (82)$$

1249 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (83)$$

1250 rule-based input depending on the condition

$$\mathbf{h}_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}}[1, 0, 1, 0]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \quad (84)$$

1251

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

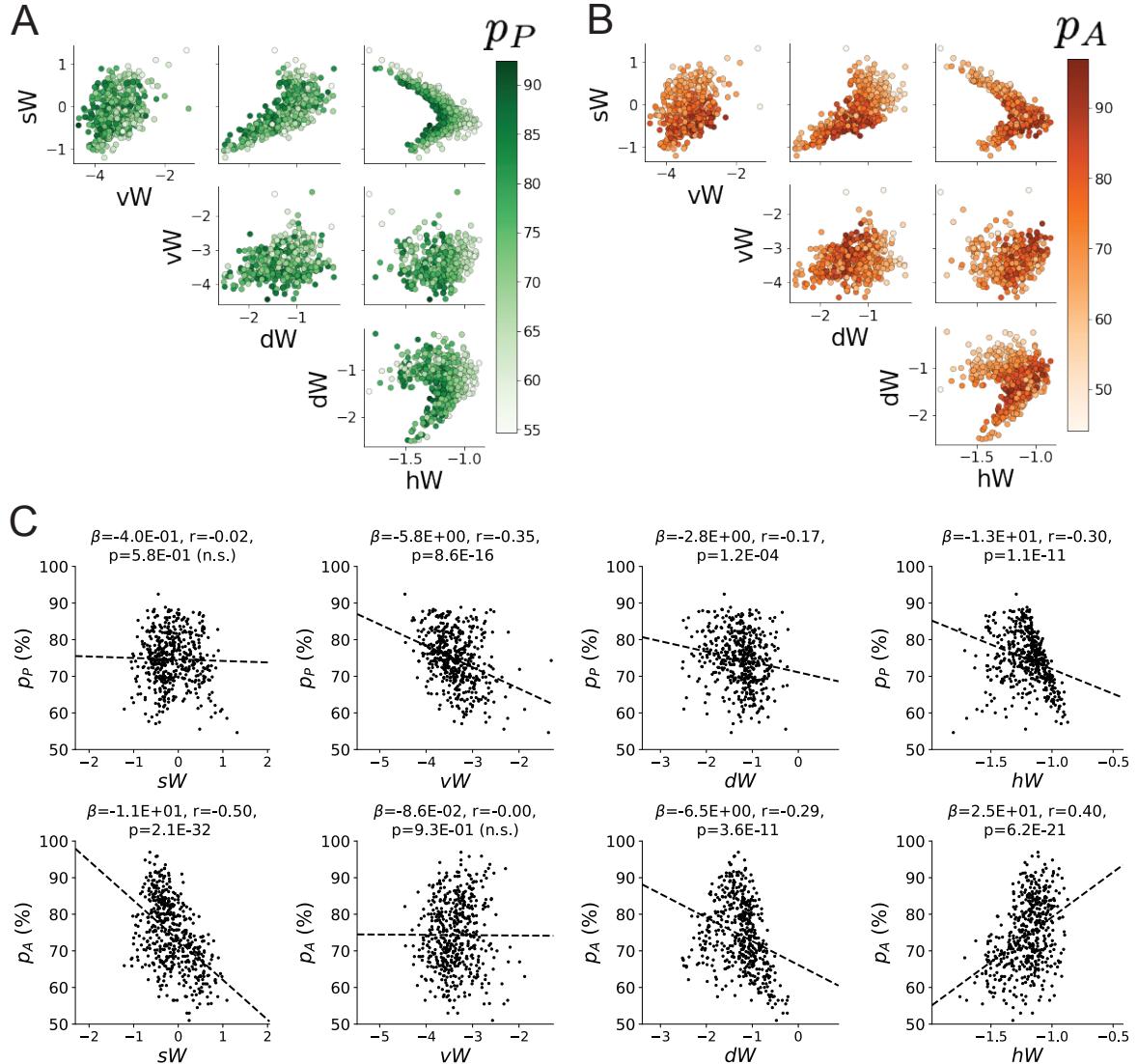


Figure 17: (SC1): **A.** Same pairplot as Fig. 4C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

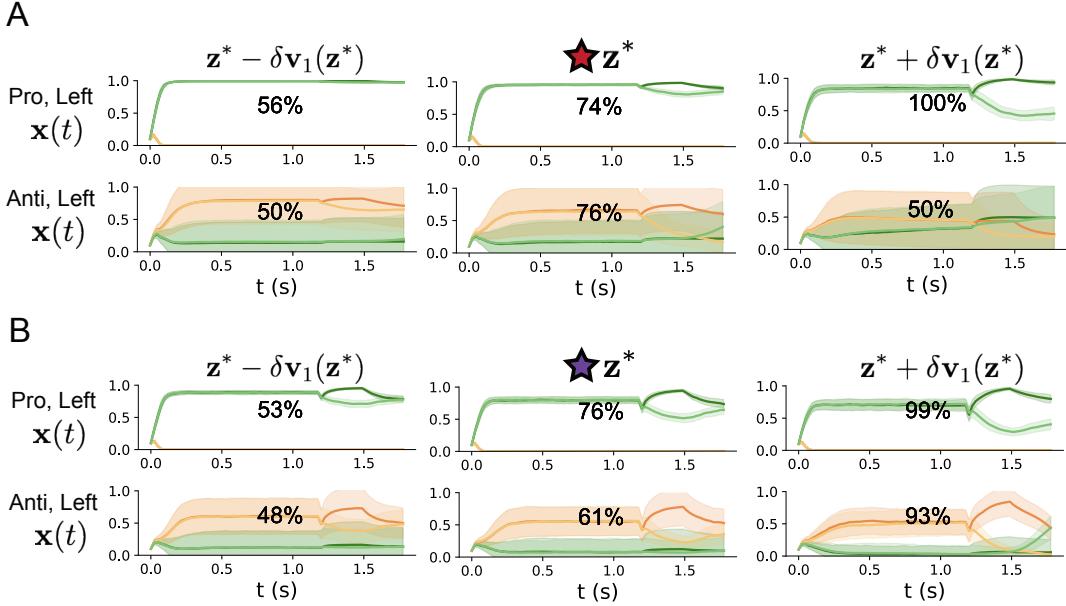


Figure 18: (SC2): **A.** Simulations in network regime 1 ($hW_{\text{fixed}} = -1.2$) (center) with simulations given connectivity perturbations in the negative direction of the sensitivity vector \mathbf{v}_1 (left) and positive direction (right). **B.** Same as A for network regime 2.

1252 a choice-period input

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

1253 and an input to the right or left-side depending on where the light stimulus is delivered

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^{\top}, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^{\top}, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases}. \quad (87)$$

1254 The input parameterization was fixed to $I_{\text{constant}} = 0.75$, $I_{\text{P,bias}} = 0.5$, $I_{\text{P,rule}} = 0.6$, $I_{\text{A,rule}} = 0.6$,

1255 $I_{\text{choice}} = 0.25$, and $I_{\text{light}} = 0.5$.

1256 The accuracies of each task p_P and p_A are calculated as

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x}} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (88)$$

1257 and

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

1258 given that the stimulus is on the left side, where Θ is the Heaviside step function, and the accuracy

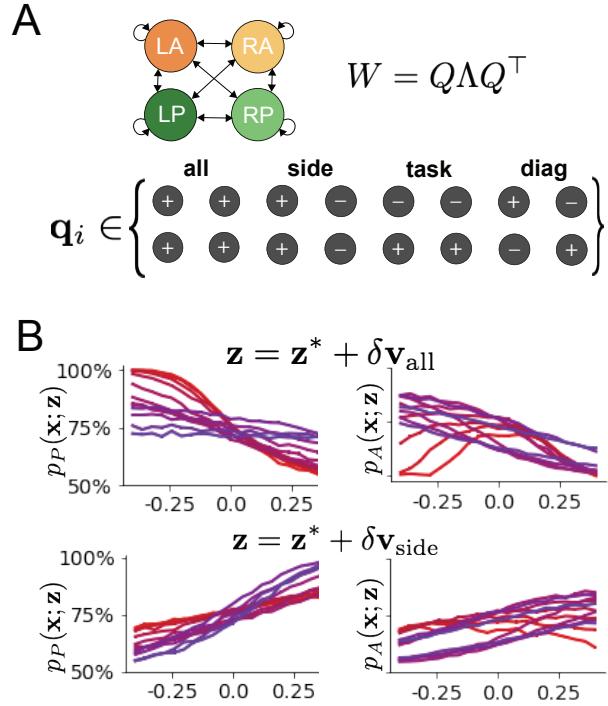


Figure 19: (SC3): **A.** Invariant eigenvectors of connectivity matrix W . **B.** Accuracies for connectivity perturbations for increasing λ_{all} and λ_{side} (rest shown in Fig. 4D).

1259 is averaged over 200 independent trials. The Heaviside step function is approximated as

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (90)$$

1260 where $\beta = 100$.

1261 Writing the EPI distribution as a maximum entropy distribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these
1262 first and second moments of the accuracy in each task (as in Equations 27 and 28)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \\ (p_P(\mathbf{x}; \mathbf{z}) - 75\%)^2 \\ (p_A(\mathbf{x}; \mathbf{z}) - 75\%)^2 \end{bmatrix}, \quad (91)$$

1263

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 7.5\%^2 \\ 7.5\%^2 \end{bmatrix}. \quad (92)$$

1264 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each

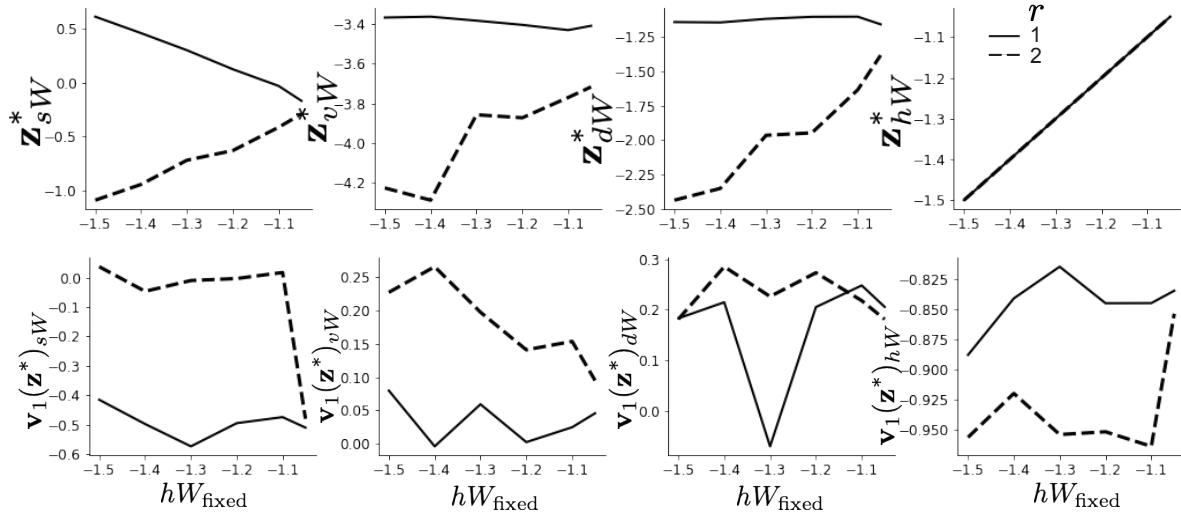


Figure 20: (SC4): **A.** The individual parameters of each mode throughout the two regimes. **B.** The individual sensitivities of parameters of each mode throughout the two regimes.

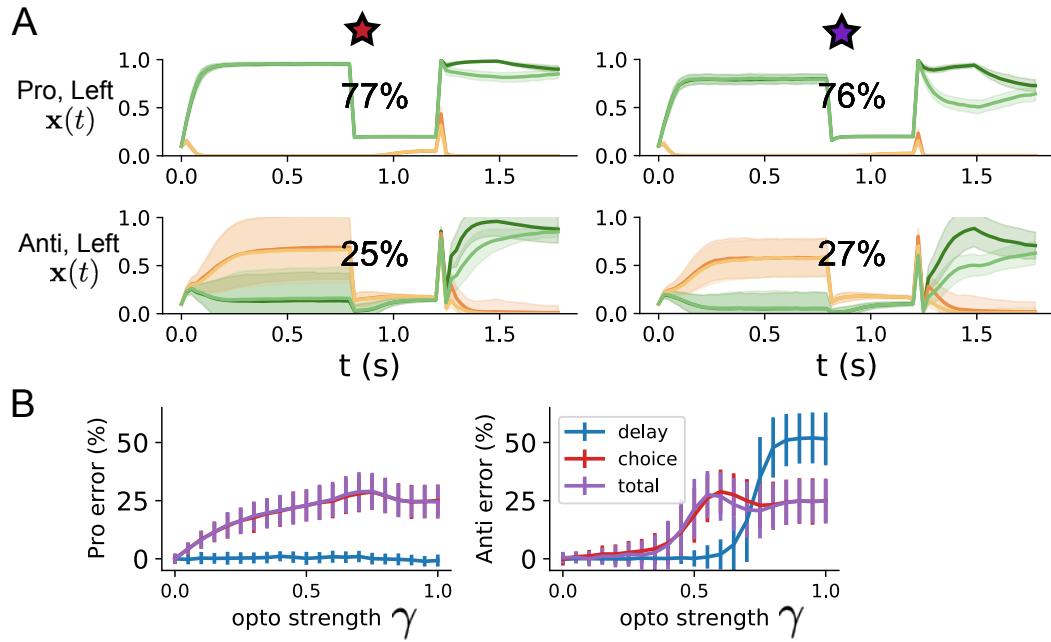


Figure 21: (SC5): **A.** Response of each parameter regime to optogenetic silencing during the delay period. **B.** Error induced by delay period inactivation with increasing optogenetic strength. Means and standard deviations are calculated across the entire EPI distribution.

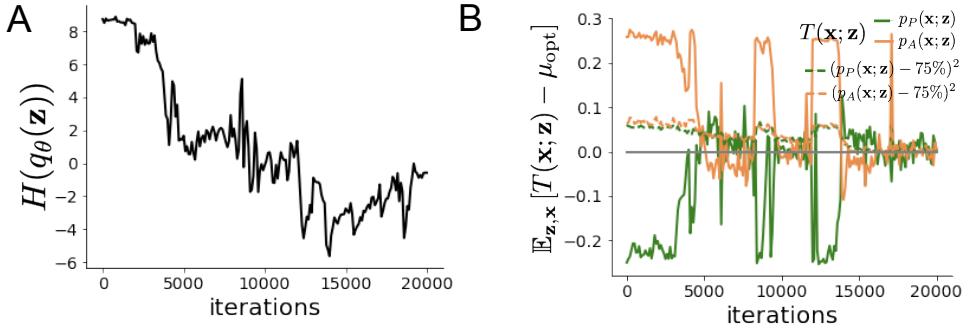


Figure 22: (SC6): **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

epoch of 2,000 iterations (see Section 5.1.3). The optimization converged after ten epochs (Fig. 22).

For EPI in Fig. 4C, we used a real NVP architecture with three coupling layers of affine transformations parameterized by two-layer neural networks of 50 units per layer. The initial distribution 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 augmented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution converged with criteria $N_{\text{test}} = 25$.

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

We sought two sets of parameters in this distribution representative of each regime, so that we could assess their implications on computation. For fixed values of hW , we hypothesized that there are two modes: one in each regime of greater and lesser sW . To begin, we found one mode for each regime at $hW_{\text{fixed}} = -1.5$ using 200 steps of gradient ascent of the deep probability distribution $q_{\theta}(\mathbf{z} \mid \mathcal{X})$. In regime 1, the initialization had positive sW , and the initialization had negative sW in regime 2, which led to disparate modes (Fig. 20 top). These modes were then used as the initialization to find the next mode at $hW_{\text{fixed}} = -1.4$ and so on. 200 steps of gradient ascent

1287 were always taken, and learning rates of 2.5×10^{-4} and 5×10^{-4} were used for regimes 1 and 2,
 1288 respectively. Each of these modes is denoted $\mathbf{z}^*(hW_{\text{fixed}}, r)$ for regime $r \in \{1, 2\}$.

1289 At each mode, we measure the sensitivity dimension (that of most negative eigenvalue in the Hessian
 1290 of the EPI distribution) $\mathbf{v}_1(\mathbf{z}^*)$. To resolve sign degeneracy in eigenvectors, we chose $\mathbf{v}_1(\mathbf{z}^*)$ to have
 1291 negative element in hW . This tells us what parameter combination rapid task switching is most
 1292 sensitive to at this parameter choice in the regime. We see that while the modes of each regime
 1293 gradually converge to similar connectivities at $hW_{\text{fixed}} = -1.05$ (Fig. 20 top), the sensitivity
 1294 dimensions remain categorically different throughout the two regimes (Fig. 20 bottom). Only at
 1295 $hW_{\text{fixed}} = -1.05$ is there a flip in sensitivity from regime 2 to regime 1 (in $\mathbf{v}_1(\mathbf{z}^*)_{sW}$ and $\mathbf{v}_1(\mathbf{z}^*)_{hW}$).
 1296 There is thus some ambiguity regarding the “regime” of $\mathbf{z}^*(-1.05, 2)$, since the mode is derived
 1297 from an initialization in regime 2, but has sensitivity like regime 1. We can consider this as an
 1298 intermediate transitional region of parameter space between the two regimes. To emphasize this,
 1299 $\mathbf{z}^*(-1.05, 1)$ and $\mathbf{z}^*(-1.05, 2)$ have the same color.

1300 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-
 1301 tion of the symmetric connectivity matrices $W = Q\Lambda Q^{-1}$, which results in the same basis vectors
 1302 \mathbf{q}_i for all W parameterized by \mathbf{z} (Fig. 19A). These basis vectors have intuitive roles in processing for
 1303 this task, and are accordingly named the *all* eigenmode - all neurons co-fluctuate, *side* eigenmode
 1304 - one side dominates the other, *task* eigenmode - the Pro or Anti populations dominate the other,
 1305 and *diag* mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair.
 1306 Due to the parametric structure of the connectivity matrix, the parameters \mathbf{z} are a linear function
 1307 of the eigenvalues $\boldsymbol{\lambda} = [\lambda_{\text{all}}, \lambda_{\text{side}}, \lambda_{\text{task}}, \lambda_{\text{diag}}]^\top$ associated with these eigenmodes.

$$\mathbf{z} = A\boldsymbol{\lambda} \quad (93)$$

1308

$$A = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}. \quad (94)$$

1309 We are interested in the effect of raising or lowering the amplification of each eigenmode in the
 1310 connectivity matrix. To test this, we calculate the unit vector of changes in the connectivity \mathbf{z} that
 1311 result from a change in the associated eigenvalues

$$\mathbf{v}_a = \frac{\frac{\partial \mathbf{z}}{\partial \lambda_a}}{\left| \frac{\partial \mathbf{z}}{\partial \lambda_a} \right|_2}, \quad (95)$$

1312 where

$$\frac{\partial \mathbf{z}}{\partial \lambda_a} = A \mathbf{e}_a, \quad (96)$$

1313 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 .

1314 While perturbations in the sensitivity dimension $\mathbf{v}_1(\mathbf{z}^*)$ adapt with the mode \mathbf{z}^* chosen, perturba-

1315 tions in \mathbf{v}_a for $a \in \{\text{all}, \text{side}, \text{text}, \text{diag}\}$ are invariant to \mathbf{z} (Equation 96).

1316 We tested whether the inferred SC model connectivities could reproduce experimental effects of

1317 optogenetic inactivation in rats [78]. During periods of simulated optogenetic inactivation, activity

1318 was decreased proportional to the optogenetic strength γ

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (97)$$

1319 Delay period inactivation was from $0.8 < t < 1.2$, choice period inactivation was for $t > 1.2$ and

1320 total inactivation was for the entire trial.