

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

¹Department of Neuroscience, Columbia University,

²Princeton Neuroscience Institute,

³Princeton University,

⁴Allen Institute for Brain Science,

⁵Institute of Neuroscience, Chinese Academy of Sciences,

⁶Howard Hughes Medical Institute,

⁷Department of Statistics, Columbia University

¹ 1 Abstract

² A cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures
³ a hypothesized neural mechanism. Such models are valuable when they give rise to an experimen-
⁴ tally observed phenomenon – whether behavioral or a pattern of neural activity – and thus can
⁵ offer insights into neural computation. The operation of these mechanistic circuits, like all models,
⁶ critically depends on the choices of model parameters. A key process in neuroscientific modeling
⁷ is then to identify the model parameters consistent with observed phenomena: to solve the in-
⁸ verse problem. While statistical inference has proven effective on a broad variety of neuroscientific
⁹ datasets, we clarify an important incongruity between theoretical approaches to neuroscience and
¹⁰ this probabilistic methodology. Theoretical neuroscience is focused on computational properties
¹¹ and how they emerge from biological mechanisms, rather than noisy experimental datasets and
¹² their quantified structure. In this work, we present a novel technique to directly infer circuit model
¹³ parameters producing these computational properties. This method tailors deep inference, the
¹⁴ use of deep neural networks for statistical inference, to the nature of theoretical inverse problems,
¹⁵ enabling scaling to high dimensions and considerable technical simplifications. With this method,
¹⁶ we bring deep inference to bear on important questions in theoretical neuroscience, and demon-
¹⁷ strate the broad range of insightful analyses this approach allows. First, we emphasize the general
¹⁸ applicability of this approach by inferring channel conductance parameters that produce a distinc-
¹⁹ tive spiking frequency in a biophysical model of the stomatogastric ganglion. Then, in a model
²⁰ of primary visual cortex with multiple neuron-types, where analysis becomes untenable as more

21 neuron-types are included, we use this technique to discover how noise properties govern excitatory population variability. Next, in a model of superior colliculus, we identify and characterize
22 two distinct regimes of connectivity that facilitate switching between opposite tasks throughout
23 interleaved trials. Finally, we scale inference to 1,000-dimensional parameter spaces of RNN
24 connectivity that exhibit stable, yet amplified responses. These analyses illustrate how we can further
25 leverage the power of deep learning towards solving inverse problems in theoretical neuroscience.
26

27 2 Introduction

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

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

47 One preferred formalism for parameter identification is statistical inference, which has been used
48 to great success in neuroscience through the stipulation of statistical generative models [15, 16, 17,
49 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29] (see review, [30]). Recent work has used variational
50 autoencoders (VAEs) [31, 32] to interrogate hidden states in models of both cortical population

51 activity [33, 34, 35, 36] and animal behavior [37, 38, 39], thus expanding the domain of neural
52 data sets amenable to statistical modeling. However, most neural circuit models in theoretical
53 neuroscience are noisy systems of differential equations that can only be sampled or realized through
54 forward simulation; they lack the explicit likelihood necessary for statistical inference. Therefore,
55 the most popular approaches to theoretical inverse problems have been likelihood-free inference
56 (LFI) methods [40, 41], in which reasonable parameters are obtained via simulation and rejection.
57 A flourishing new class of techniques [42, 43, 44] use deep learning to improve upon traditional LFI
58 approaches. However, as we detail, all of these approaches require good datasets for the scientific
59 question at hand.

60 This work seeks to clarify an important incongruity between theoretical approaches to neuroscience
61 and existing statistical inference methodology. In theoretical neuroscience, we are concerned with
62 the computational properties – the *emergent phenomena* – of our models [7, 8, 9, 10], not noisy
63 observed datasets [30]. To use the aforementioned inference paradigm, scientists must shoehorn
64 such mathematical criteria into an artificial dataset compatible with existing statistical approaches.
65 Theorists are therefore barred from using the probabilistic modeling toolkit for science, unless they
66 reformulate their inverse problem to fit an evidence accumulation framework.

67 These challenges motivate the development of a novel inference framework called emergent property
68 inference (EPI). As an adaption of variational inference [45], EPI infers parameter distributions
69 that produce an emergent property: not a singular dataset, but a collection of datasets exhibiting
70 some mathematical criteria. EPI fixates on the inverse problem by constraining the predictions
71 of the inferred parameter distribution to produce the emergent property exactly. Conditioning on
72 an emergent property requires a variant of deep probabilistic inference methods, which we have
73 previously introduced [46]. By using deep probability distributions EPI flexibly captures para-
74 metric structure in mechanistic models. This technique is designed to identify the full parameter
75 space producing an emergent property and facilitate the seamless structural analysis of the inferred
76 parameter distribution. Thus, EPI provides neuroscientists with an efficient, versatile probabilistic
77 modeling toolkit designed explicitly for theoretical inverse problems.

78 Equipped with this method, we bring deep inference to bear on theoretical neuroscience to an
79 unprecedented extent. Throughout this work, we showcase the capabilities of EPI on four neural
80 circuit models across ranges of biological realism, neural system function, and network scale. First,
81 we show EPI’s ability to capture subtle, nonlinear parametric structure in a stomatogastric ganglion
82 subcircuit model [47]. In a model of primary visual cortex [48], we show how to gain insight from

multiple inferred distributions. Next, we used EPI to identify and structurally characterize multiple parametric regimes of superior colliculus activity in a model of task switching [49]. Finally, we emphasize the superior scalability of EPI compared to other LFI techniques by inferring high-dimensional distributions of RNN connectivities that exhibit amplified, yet stable responses – a hallmark of cortical sensory systems [?, 50].

Most importantly of all in this work, we present novel theories of neural computation borne from EPI analysis. We identified an unknown parametric rule of variability with respect to inhibitory neuron type in a V1 model, where analytic techniques became untenable. Furthermore, we identified multiple regimes of SC connectivity which confer rapid task switching, and used the structural analytic tools of EPI to gain a mechanistic understanding of circuit responses. These valuable theoretical insights 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 [51]. 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 neural activity [52, 12]. A subcircuit model of the STG [47] is shown schematically in Figure 1A, and 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 (f_1 and f_2) mutually inhibit one another, and oscillate at a faster frequency than the mutually inhibiting slow neurons (s_1 and s_2). The hub neuron (hub) couples with either the fast or slow population or both.

Second, once the model is selected, one defines the emergent phenomena of scientific interest. In the STG example, we are concerned with neural spiking frequency, which emerges from the dynamics of the circuit model 1B. An interesting emergent property of this stochastic model is when the hub neuron fires at an intermediate frequency between the intrinsic spiking rates of the fast and slow

113 populations. This emergent property is shown in Figure 1C at an average frequency of 0.55Hz.
 114 Third, parameter analyses ensue: brute-force parameter sweeps, ABC sampling [41], and sensitivity
 115 analyses [53] are all routinely used to reason about what parameter configurations lead to an
 116 emergent property. In this last step lies the opportunity for a precise quantification of the emergent
 117 property as a statistical feature of the model. Once we have such a methodology, we can infer a
 118 probability distribution over parameter configurations that produce this emergent property.
 119 Before presenting technical details (in the following section), let us understand emergent property
 120 inference schematically: EPI (Fig. 1D) takes, as input, the model and the specified emergent
 121 property, and as its output, produces the parameter distribution EPI (Fig. 1E). This distribution
 122 – represented for clarity as samples from the distribution – is then the most random parameter
 123 distribution producing the emergent property. In the STG model, this distribution can be specif-
 124 ically queried to reveal the prototypical parameter configuration for intermediate hub frequency
 125 (the mode; Figure 1E yellow star), and how it decays based on changes away from the mode. In-
 126 deed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity (v_1)
 127 and degeneracy (v_2) (Fig. 1E, arrows) agree with error contours (Fig. 1E contours) and have
 128 diminished or preserved hub frequency, respectively (Fig. 1F activity traces) (see Section 5.2.1).

129 **3.2 A deep generative modeling approach to emergent property inference**

130 Emergent property inference (EPI) formalizes the three-step procedure of the previous section with
 131 deep probability distributions. First, we consider the model as a coupled set of differential equations
 132 [47]. In the running STG example, the model activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{\text{hub}}, x_{s1}, x_{s2}]$ is the membrane
 133 potential for each neuron, which evolves according to the biophysical conductance-based equation:

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

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

137 Second, we define the emergent property, which as above is “intermediate hub frequency” (Figure
 138 1C). Quantifying this phenomenon is straightforward: we stipulate that the hub neuron’s spiking
 139 frequency – denoted $\omega_{\text{hub}}(\mathbf{x})$ is close to a frequency of 0.55Hz. Mathematically, we achieve this

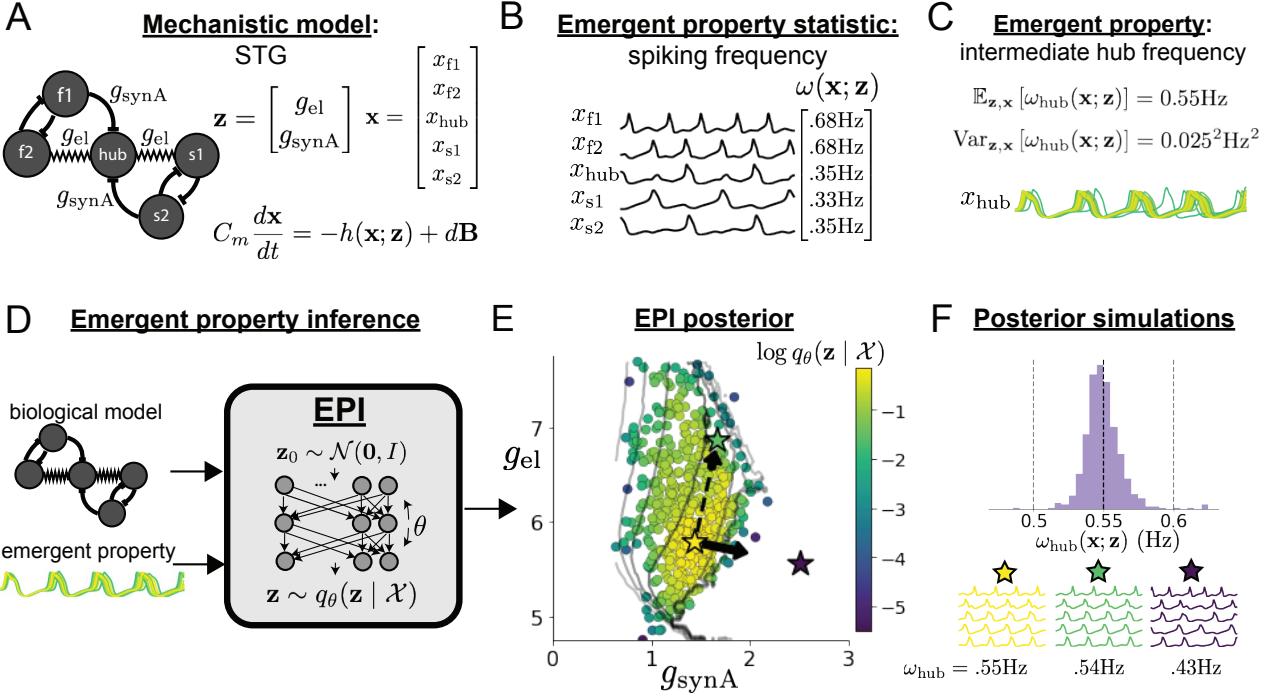


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based biophysical model of the STG subcircuit. In the STG model, jagged connections indicate electrical coupling having electrical conductance g_{el} . Other connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto the hub neuron, and $g_{\text{synB}} = 5\text{nS}$ for mutual inhibitory connections. Parameters are represented by the vector \mathbf{z} and membrane potentials by the vector \mathbf{x} . The evolution of this model's activity $\mathbf{x}(t)$ is predicated by differential equations. **B.** Spiking frequency $\omega(\mathbf{x}; \mathbf{z})$ is an emergent property statistic. In this example, spiking frequency is measured from simulated activity of the STG model at parameter choices of $g_{\text{el}} = 4.5\text{nS}$ and $g_{\text{synA}} = 3\text{nS}$. **C.** The emergent property of intermediate hub frequency, in which the hub neuron fires at a rate between the fast and slow frequencies. This emergent property is defined by a mean and variance on the emergent property statistic. Simulated activity traces are colored by log probability density of their generating parameters in the EPI-inferred distribution (Panel E). **D.** For a choice of model and emergent property, emergent property inference (EPI) learns a deep probability distribution of parameters \mathbf{z} . Deep probability distributions map a simple random variable $\mathbf{z}_0 \sim \mathcal{N}(0, I)$ through a deep neural network with weights and biases $\boldsymbol{\theta}$ to parameters $\mathbf{z} = q_{\boldsymbol{\theta}}(\mathbf{z}_0)$. In EPI optimization, stochastic gradient steps in $\boldsymbol{\theta}$ are taken such that entropy is maximized, and the emergent property \mathcal{X} is produced. The EPI posterior distribution is denoted $q_{\boldsymbol{\theta}}(\mathbf{z} | \mathcal{X})$. **E.** The EPI posterior producing intermediate hub frequency. Samples are colored by log probability density. Distribution contours of average hub neuron frequency from mean of .55 Hz are shown at levels of .525, .53,575 Hz (dark to light gray away from mean). Eigenvectors of the Hessian at the mode of the inferred distribution are indicated as \mathbf{v}_1 (solid) and \mathbf{v}_2 (dashed) with lengths scaled by the square root of the absolute value of their eigenvalues. **F** Simulations from parameters in E. (Top) The predictive distribution of the posterior obeys the emergent property. The black and gray dashed lines show the mean and two standard deviations according the emergent property, respectively. (Bottom) Simulations at the starred parameter values.

140 with two constraints: by setting the mean hub frequency of model generated activity to 0.55Hz.

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

141 and requiring that the variance around this mean is small

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

142 This level of variance was chosen to be low enough to exclude the fast and slow frequencies of the
143 two populations, but large enough to allow structural examination of the compatible parameter
144 space. In general, an emergent property

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

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

149 Third, we perform emergent property inference: we find a distribution over parameter configura-
150 tions \mathbf{z} , and insist that samples from this distribution produce the emergent property; in other
151 words, they obey the constraints introduced in Equation 4. This distribution will be chosen from a
152 family of probability distributions $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$, defined by a deep generative distribution
153 of the normalizing flow class [54, 55, 56] – neural networks which transform a simple distribution
154 into a suitably complicated distribution (as is needed here). This deep distribution is represented
155 in Figure 1D (see Section 5.1). Many distributions in \mathcal{Q} will respect the emergent property con-
156 straints, so we choose a normative selection principle imposing no additional structure beyond the
157 production of the emergent property [57, 58, 46, 59], which is the same normative principle of
158 Bayesian inference (see Section 5.1.6).

159 The probabilities of the distribution inferred from EPI are the densities of these parameters in the
160 most random distribution producing the emergent property. While existing approaches to proba-
161 bilistic structural identifiability analysis use the lens of evidence accumulation [53, 42], sensitivity
162 and robustness of parameter space dimensions with respect to emergent phenomena can be directly
163 quantified with EPI. Sensitivity quantifications are measured by the second order derivative of EPI
164 probability along the parameteric dimension of interest: to what extent is the emergent property
165 maintained or diminished along this dimension? Once an EPI distribution has been inferred, this
166 second order derivative requires trivial computation (as long as the correct architecture class is

167 chosen, see Section 5.1.2). Equipped with this method, we may examine structure in the resulting
 168 parameter distributions or make comparisons between distributions conditioned at different levels
 169 of the same emergent property statistic. In Sections 3.3 and 3.4, we prove out the value of EPI by
 170 using these techniques to investigate and produce novel scientific insight.

171 **3.3 EPI reveals how neuron-type specific noise governs variability in the stochastic
 172 stabilized supralinear network**

173 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output
 174 function have succeeded in explaining a host of experimentally documented phenomena. In a
 175 regime characterized by inhibitory stabilization of strong recurrent excitation, these models give
 176 rise to paradoxical responses [9], selective amplification [60, 61], surround suppression [62] and
 177 normalization [63]. Despite their strong predictive power, E-I circuit models rely on the assumption
 178 that inhibition can be studied as an indivisible unit. However, experimental evidence shows
 179 that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) –
 180 composing 80% of GABAergic interneurons in V1 [64, 65, 66], and that these inhibitory cell types
 181 follow specific connectivity patterns (Fig. 2A) [67]. Recent theoretical advances [68, 69, 70], have
 182 only started to address the consequences of this multiplicity in the dynamics of V1, strongly relying
 183 on linear theoretical tools. Here, we use EPI to analyze V1 models of greater complexity in order
 184 to characterize properties of slow noise governing circuit variability.

185 We considered the response properties of a nonlinear dynamical V1 circuit model (Fig. 2A) with
 186 a state comprised of each neuron-type population’s rate $\mathbf{x} = [x_E, x_P, x_S, x_V]^\top$. Each population
 187 receives recurrent input $W\mathbf{x}$ from synaptic projections of effective connectivity W and an external
 188 input \mathbf{h} , which determine the population rate via supralinear nonlinearity $\phi = []_+^2$. The input is
 189 also comprised of a slow noise component $\epsilon \sim OU(\tau_{noise}, \sigma)$ of time scale $\tau_{noise} > \tau$ and variance
 190 parameters σ (see Section 5.2.2)

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \epsilon). \quad (5)$$

191 This model is the stochastic stabilized supralinear network (SSSN) [71] generalized to have in-
 192 hibitory multiplicity, and introduces stochasticity to previous four neuron-type models of V1 [68].
 193 Stochasticity and inhibitory multiplicity introduce substantial complexity to mathematical deriva-
 194 tions (see Section 5.2.3) motivating the treatment of this model with EPI. Here, we consider fixed
 195 weights W and input \mathbf{h} according to a fit of the deterministic model to contrast responses [48] (Fig.

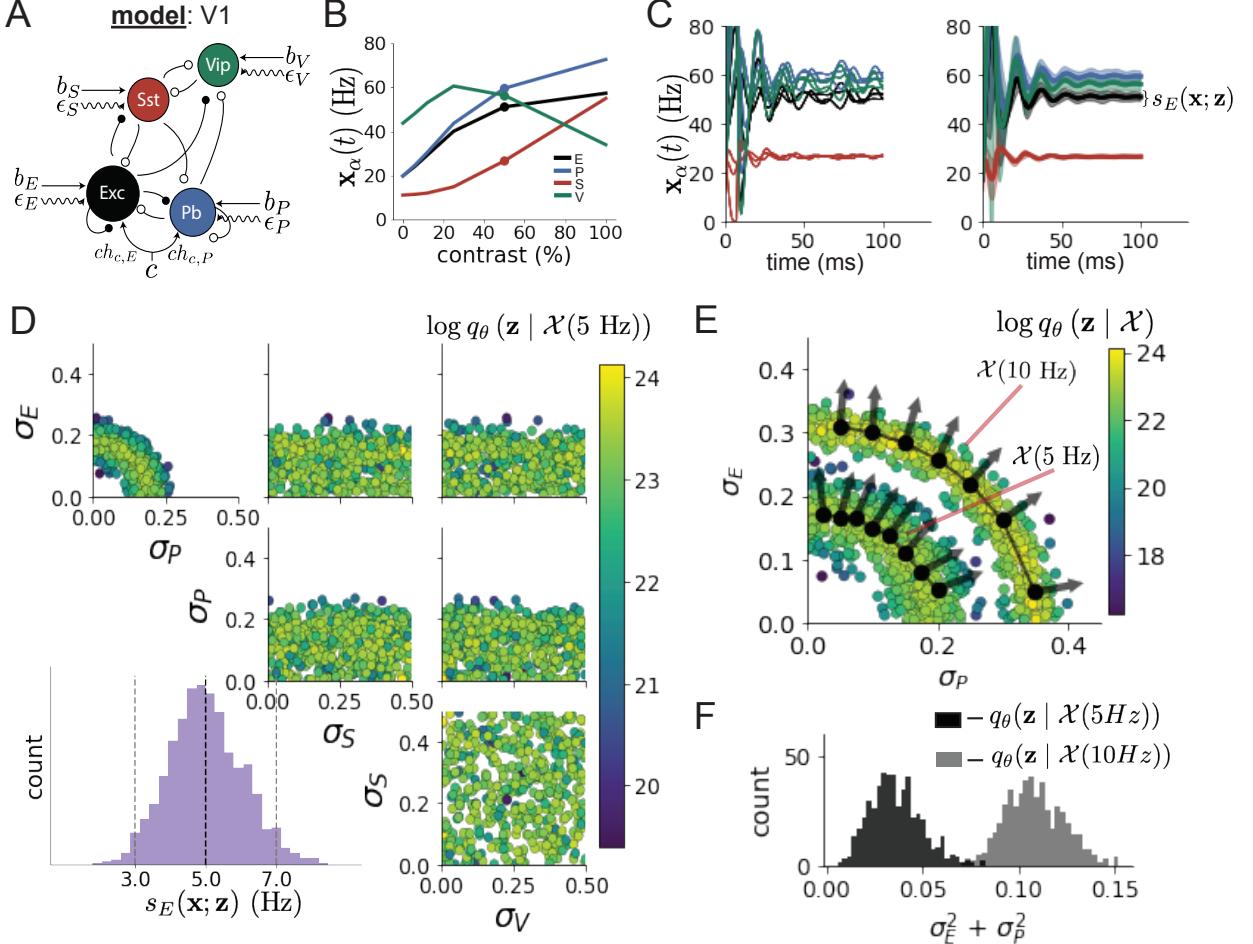


Figure 2: Emergent property inference in the stochastic stabilized supralinear network (SSSN) **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively). 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 posterior of noise parameters \mathbf{z} conditioned on E-population variability. The posterior 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 the posteriors $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ 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 posterior $q_\theta(\mathbf{z} | \mathcal{X}(5 \text{ Hz}))$ and $q_\theta(\mathbf{z} | \mathcal{X}(10 \text{ Hz}))$.

196 2B), and study the effect of noise parameterization $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on fluctuations at 50%
197 contrast.

198 For this SSSN, we are interested in how noise variability across neural populations governs stochastic
199 fluctuations in the E-population. Here, we quantify different levels y of E-population variability
200 with the emergent property

$$\begin{aligned}\mathcal{X}(y) &: \mathbb{E}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] = y \\ \text{Var}_{\mathbf{z}} [s_E(\mathbf{x}; \mathbf{z})] &= 1\text{Hz}^2,\end{aligned}\tag{6}$$

201 where $s_E(\mathbf{x}; \mathbf{z})$ is the standard deviation of the stochastic E-population response about its steady
202 state (Fig. 2C).

203 We ran EPI to obtain a posterior distribution $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz})$ producing E-population variability
204 around 5 Hz (Fig. 2D). From the marginal distribution of σ_E and σ_P (Fig. 2D, top-left), we can
205 see 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
206 are degenerate with respect to $s_E(\mathbf{x}; \mathbf{z})$ evidenced by the high variability in those dimensions of the
207 posterior (Fig. 2D, bottom-right). Together, these observations imply a parametric manifold of
208 degeneracy with respect to $s_E(\mathbf{x}; \mathbf{z})$ of 5 Hz, which is indicated by the modes along σ_P in the σ_E - σ_P
209 marginal (Fig. 2E). The dimensions of sensitivity conferred by EPI and this plain visual structure
210 suggest a quadratic relationship in the emergent property statistic $s_E(\mathbf{x}; \mathbf{z})$ and parameters \mathbf{z} , which
211 is preserved at a greater level of variability $\mathcal{X}(10 \text{ Hz})$ (Fig. 2E). Indeed, the sum of squares of σ_E
212 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 2F, $p = 0$), while the sum of squares
213 of σ_S and σ_V are not significantly different in the two posteriors (Fig. 11, $p = .402$).

214 While a quadratic relationship in $s_E(\mathbf{x}; \mathbf{z})$ and \mathbf{z} is potentially derivable by extending the derivation
215 in Section 5.2.2 to the case of $\tau \neq \tau_{\text{noise}}$, the coefficients in front of each quadratic term would be
216 unruly, and likely escape comprehensible analysis. This makes EPI an attractive tool for revealing
217 the characteristics of noise governing variability and for answering other questions in this complex
218 model. Intriguingly, this circuit exhibited a paradoxical effect in the P-population, and no other
219 inhibitory types at 50% contrast (Fig. 11) implying that the E-population is P-stabilized. Future
220 work motivated by our analysis here, may uncover a relationship between the neuron-type mediating
221 stability and the factors governing circuit variability.

222 **3.4 EPI identifies multiple regimes of rapid task switching**

223 In a rapid task switching experiment [72], rats were explicitly cued on each trial to either orient
 224 towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti
 225 (A) task (Fig. 3A). Neural recordings in the midbrain superior colliculus (SC) exhibited two
 226 populations of neurons that simultaneously represented both task context (Pro or Anti) and motor
 227 response (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons
 228 [49]. Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is
 229 a four-population dynamical system. We analyzed this model, where the neuron-type populations
 230 are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right
 231 (R)), their connectivity is parameterized geometrically (Fig. 3B). The input-output function of
 232 this model is chosen such that the population responses $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$ are bounded
 233 from 0 to 1 as a function ϕ of a dynamically evolving internal variable \mathbf{u} . The model responds to
 234 the side with greater Pro neuron activation; e.g. the response is left if $x_{LP} > x_{RP}$ at the end of
 235 the trial. The dynamics evolve with timescale $\tau = 90\text{ms}$ governed by connectivity weights W

$$\begin{aligned} \tau \frac{d\mathbf{u}}{dt} &= -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B} \\ \mathbf{x} &= \phi(\mathbf{u}) \end{aligned} \tag{7}$$

236 with white noise of variance 0.2^2 . The input \mathbf{h} is comprised of a cue-dependent input to the Pro
 237 or Anti populations, a stimulus orientation input to either the Left or Right populations, and a
 238 choice-period input to the entire network (see Section 5.2.4). Here, we use EPI to determine the
 239 network connectivity $\mathbf{z} = [sW, vW, dW, hW]^\top$ that produces rapid task switching behavior.

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

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 75\% \\ 75\% \end{bmatrix} \\ \text{Var}_{\mathbf{z}} \begin{bmatrix} p_P(\mathbf{x}; \mathbf{z}) \\ p_A(\mathbf{x}; \mathbf{z}) \end{bmatrix} &= \begin{bmatrix} 7.5\%^2 \\ 7.5\%^2 \end{bmatrix} \end{aligned} \tag{8}$$

245 A variance of $7.5\%^2$ in each task will confer a posterior producing performances ranging from about
 246 60% – 90%, allowing us to examine the properties of connectivity that yield better performance in

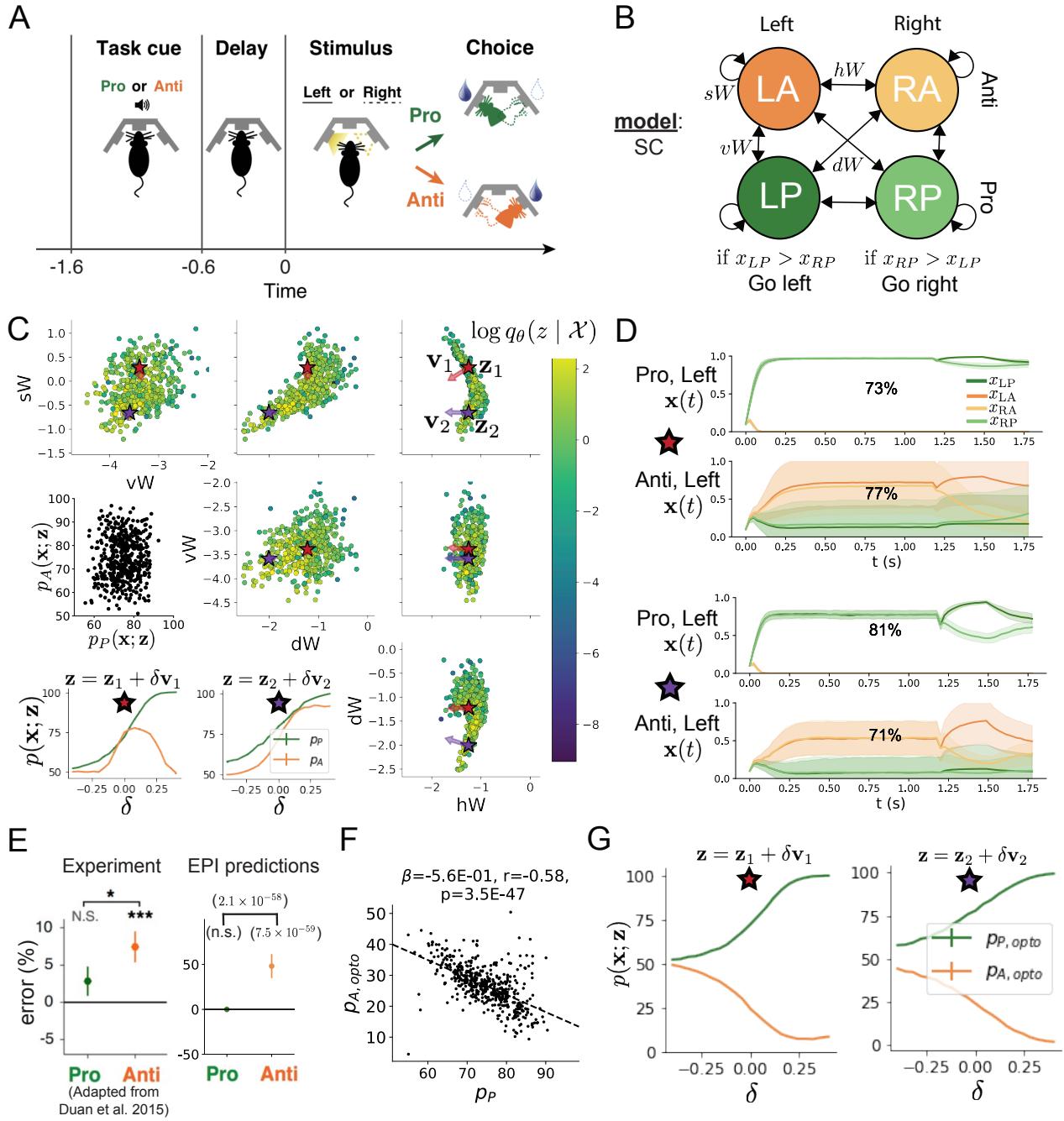


Figure 3: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. **C.** The EPI posterior distribution of rapid task switching networks. Red and purple stars (\mathbf{z}_1 and \mathbf{z}_2) indicate different connectivity regimes with different sensitivity vectors \mathbf{v}_1 and \mathbf{v}_2 . (Middle-left) Posterior predictive distribution of task accuracies. (Bottom-left) Task accuracy along dimensions of sensitivity in each connectivity regime. **D.** Means (solid) and standard deviations (shaded) of each population across random simulated trials. Top plots show Pro (top) and Anti (bottom) responses for connectivity \mathbf{z}_1 . Bottom rows show the same \mathbf{z}_2 . **E.** The EPI posterior predicts experimental results (left) showing no change in the Pro task, but larger error in the Anti task (right). **F.** Accuracy in the Anti task during delay period optogenetic inactivation $p_{A,\text{opto}}$ is strongly anticorrelated with accuracy in the Pro task. **G.** Accuracy with delay period inactivation along each connectivity regime's dimension of sensitivity.

247 each task. Notably, this is our first example using EPI to condition on multiple emergent property
248 statistics ($|f(\mathbf{x}; \mathbf{z})| = 2$).

249 The EPI inferred parameters (Fig. 3C) generate a distribution of task accuracies (Fig. 3C, middle-
250 left) according to our mathematical definition of rapid task switching (Equation 8). The nonlinear
251 patterns of connectivity that govern each task accuracy (Fig. 12A-B) are not fully captured by
252 linear prediction (Fig. 12C). For example, the patterns in connectivity increasing Pro accuracy
253 change dramatically after crossing a threshold of sW (Fig. 12A $sW-hW$ marginal). Not only has
254 EPI captured this complex nonlinear posterior, it offers probabilistic tools for understanding the
255 different regimes of model behavior.

256 To establish these two regimes of connectivity, we took gradient steps along $q_{\theta}(\mathbf{z} \mid \mathcal{X})$ to produce
257 modes \mathbf{z}_1 and \mathbf{z}_2 (Fig. 3C red and purple stars, Section 5.2.4). Simulations from these two regimes
258 reveal different responses in each task (Fig. 3D). We characterized these regimes by identifying
259 the dimensions of connectivity that rapid task switching is most sensitive to. The sensitivity
260 dimensions \mathbf{v}_1 and \mathbf{v}_2 (Fig. 3C, red and purple arrows) point in different directions, resulting
261 in different changes to task accuracy (Fig. 3D, bottom-left, 13). In regime 1, Anti accuracy
262 diminishes in either direction of sensitivity away from the mode, while in regime 2, Anti accuracy
263 tracks monotonic increases in Pro accuracy. These responses make intuitive sense, recognizing that
264 \mathbf{v}_1 (unlike \mathbf{v}_2) points strongly in the direction of connectivity eigenvalue λ_{diag} , which is strongly
265 anticorrelated with p_A (Fig. 14, 15, see Section 5.2.4).

266 In agreement with experimental results from Duan et al., we found optogenetic inactivation during
267 the delay period consistently decreased performance in the Anti task, but had no effect on the
268 Pro task (Fig. 3E)). This difference in resiliency across tasks to delay perturbation is a prediction
269 made by the inferred EPI distribution, rather than an emergent property that was conditioned
270 upon. Similarities across Pro and Anti trials in choice period responses following delay period
271 inactivation (Fig. 17A) suggested that connectivity patterns inducing greater Pro task accuracy
272 increase error in delay period inactivated Anti trials (Fig. 3F). The strong anticorrelation between
273 p_P and $p_{A,\text{opto}}$ across posterior connectivities led to the following hypothesis about each connectivity
274 regime: the sensitivity dimension of each regime decreases $p_{A,\text{opto}}$ irrespective of its effect on p_A ,
275 since both \mathbf{v}_1 and \mathbf{v}_2 increase p_P . Indeed, in regimes 1 and 2 where sensitivity dimensions elicit
276 different responses in p_A , $p_{A,\text{opto}}$ decreases since the connectivity changes enhancing p_P exacerbate
277 Anti trial error (Fig. 3F).

278 In summary, we used EPI to obtain the full distribution of connectivities that execute rapid task

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

3.5 EPI scales well to high-dimensional parameter spaces

Transient amplification is a hallmark of neural activity throughout cortex, and is often thought to be intrinisically generated by recurrent connectivity in the responding cortical area [61, 73, 74]. It has been shown that to generate such amplified, yet stabilized responses, the connectivity of RNNs must be non-normal [60, 61], and satisfy additional constraints [50]. In theoretical neuroscience, RNNs are optimized and then examined to show how dynamical systems could execute a given computation [75, 76], but such biologically realistic constraints on connectivity are ignored during optimization for practical reasons. In general, access to distributions of connectivity adhering to theoretical criteria like stable amplification, chaotic fluctuations [8], or low tangling [77] would add great scientific value and contextualization to existing research with RNNs. Here, we use EPI to learn RNN connectivities producing stable amplification, and demonstrate the superior scalability and efficiency of EPI to alternative approaches.

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

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}, \quad (9)$$

where $U = [\mathbf{u}_1 \ \mathbf{u}_2] + g\chi^{(W)}$, $V = [\mathbf{v}_1 \ \mathbf{v}_2] + g\chi^{(V)}$, $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2 \in [-1, 1]^N$, and $\chi_{i,j}^{(W)}, \chi_{i,j}^{(V)} \sim \mathcal{N}(0, 1)$. 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 amplification. Two conditions are necessary and sufficient for RNNs to exhibit stable amplification [50]: $\text{real}(\lambda_1) < 1$ and $\lambda_1^s > 1$, where λ_1 is the eigenvalue of W with greatest real part and λ^s is the maximum eigenvalue of $W^s = \frac{W+W^\top}{2}$. RNNs with $\text{real}(\lambda_1) = 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$ will be stable with modest decay rate ($\text{real}(\lambda_1)$ close to its upper bound of 1) and exhibit modest amplification (λ_1^s close to its lower bound of 1). EPI can naturally condition on this emergent

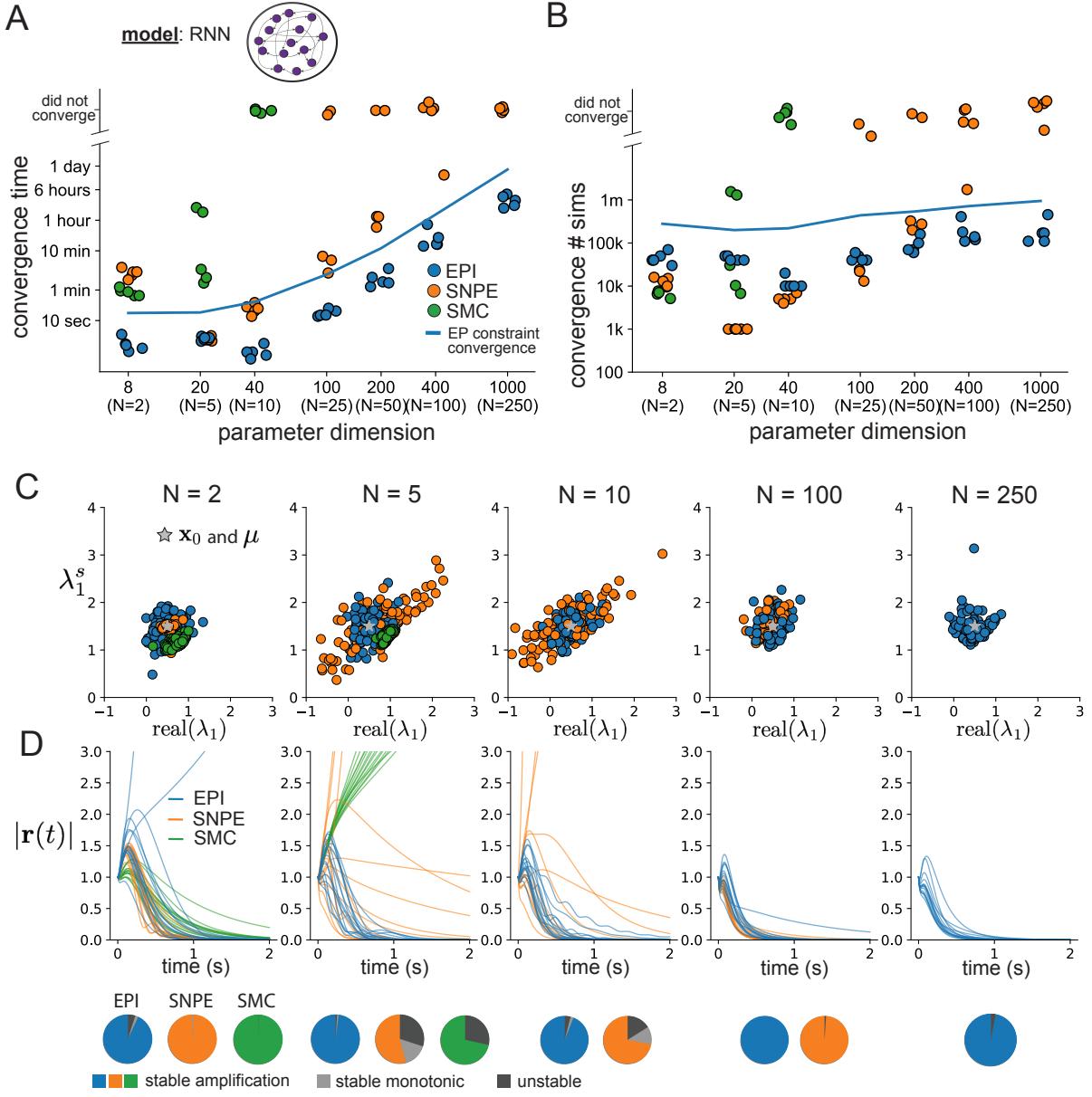


Figure 4: **A.** Wall time of EPI (blue), SNPE (orange), and SMC-ABC (green) to converge on RNN connectivities producing stable amplification. Each dot shows convergence time for an individual random seed. For reference, the mean wall time for EPI to achieve its full constraint convergence (means and 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.

304 property

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}, \end{aligned} \quad (10)$$

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

307 For comparison, we infer the parameters \mathbf{z} likely to produce stable amplification using two alter-
 308 native likelihood-free inference approaches. We ran sequential monte-carlo approximate Bayesian
 309 computation (SMC-ABC) [40] and sequential neural posterior estimation (SNPE) [42] with obser-
 310 vation $\mathbf{x}_0 = \boldsymbol{\mu}$. SMC-ABC is a rejection sampling approach that uses sophisticated techniques to
 311 improve efficiency, and SNPE approximates posteriors with deep probability distributions using a
 312 two-network architecture (see Section 5.1.1). Unlike EPI, these statistical inference techniques do
 313 not control the mean or variance of the predictive distribution, and these predictions of the inferred
 314 posteriors are typically affected by model characteristics (e.g. N and g , Fig. 18). To compare the
 315 efficiency of these different techniques, we measured the time and number of simulations necessary
 316 for the distance of the predictive mean to be less than 0.25 from $\boldsymbol{\mu} = \mathbf{x}_0$ (see Section 5.2.5).

317 As the number of neurons N in the RNN are scaled, and thus the dimension of the parameter space
 318 $\mathbf{z} \in [-1, 1]^{4N}$, we see that EPI converges at greater speed and at greater dimension than SMC-ABC
 319 and SNPE (Fig. 4A). It also becomes most efficient to use EPI in terms of simulation count at
 320 $N = 50$ (Fig. 4B). It is well known that ABC techniques struggle mightily in dimensions greater
 321 than about 30 [78], yet we were careful to assess the scalability of the more comparable approach
 322 SNPE. Between EPI and SNPE, we closely controlled the number of parameters in deep probability
 323 distributions by dimensionality (Fig. 19), and tested more aggressive SNPE hyperparameterizations
 324 when SNPE failed to converge (Fig. 20).

325 No matter the number of neurons, EPI always produces connectivity distributions with mean
 326 and variance of $\text{real}(\lambda_1)$ and λ_1^s according to \mathcal{X} (Fig. 4C, blue). For the dimensionalities in
 327 which SMC-ABC is tractable, the inferred parameters always exhibit stable amplification (Fig.
 328 4C, green), but are concentrated and offset from \mathbf{x}_0 . When using SNPE the inferred parameters
 329 are highly concentrated at some RNN sizes and widely varied in others (Fig. 4C, orange). We see
 330 these properties reflected in simulations from the inferred distributions: EPI produces a consistent
 331 variety of stable amplified activity norms $|r(t)|$, SMC produces a limited variety in responses, and the

332 changing variety of responses from SNPE emphasizes the control of EPI on parameter predictions.
333 From this analysis, we see that deep inference techniques EPI and SNPE are far more amenable to
334 high dimensional parameter distributions than rejection sampling techniques like SMC-ABC, and
335 have the additional benefit of fast sampling after optimization. EPI outperforms SNPE in high
336 dimensions by leveraging gradient information (from $\nabla_{\mathbf{z}}f(\mathbf{x}; \mathbf{z}) = \nabla_{\mathbf{z}}[\text{real}(\lambda_1), \lambda_1^s]^\top$) on each opti-
337 mization iteration and constantly adapting the approximation $q_{\theta}(\mathbf{z})$ to the parameter distribution.
338 While EPI can be used for scientific insight into RNNs, this analysis shows that when $\nabla_{\mathbf{z}}f(\mathbf{x}; \mathbf{z})$ is
339 tractable, EPI can be used to efficiently infer high dimensional parameter distributions in mechani-
340 stic models of neural computation.

341 4 Discussion

342 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets
343 [30]. Such careful inference procedures are developed for these statistical models allowing precise,
344 quantitative reasoning, which clarifies the way data informs beliefs about the model parameters.
345 However, these statistical models lack resemblance to the underlying biology, making it unclear
346 how to go from the structure revealed by these methods, to the neural mechanisms giving rise
347 to it. In contrast, theoretical neuroscience has focused on careful mechanistic modeling and the
348 production of emergent properties of computation. The careful steps of *i.*) model design and
349 *ii.*) emergent property definition, are followed by *iii.*) practical inference methods resulting in an
350 opaque characterization of the way model parameters govern computation. In this work, we improve
351 upon parameter inference techniques in theoretical neuroscience with emergent property inference,
352 harnessing deep learning towards careful inference in careful models of neural computation (see
353 Section 5.1.1).

354 Specifically, approximate Bayesian computation [79, 80, 40] has been the standard approach to
355 parameter inference in neural circuit models lacking tractable likelihoods. ABC methods do not
356 confer probabilities on accepted parameters, require an acceptance threshold chosen to trade-off
357 inference quality with tractability, do not scale efficiently to high-dimensional parameter spaces, and
358 require independent techniques to analyze sensitivity for local parameter choices [78]. In contrast,
359 EPI allows probability evaluations at any point in parameter space, conditions posteriors on the
360 natural quantification of emergent properties, scales to high dimensional parameter spaces, and
361 naturally admits sensitivity quantification via fast evaluations of the posterior Hessian.

362 Technically, EPI is a maximum entropy method, which learns parameter distributions that are
363 as random as possible given that they produce the emergent property. Conceptually, maximally
364 random distributions given some constraints are useful for understanding parametric sensitivity.

365 This is well understood in Bayesian inference, where maximum entropy is the chosen normative
366 principle. This is emphasized by an innovative formalism unifying top-down maximum entropy
367 normative models with bottom-up statistical models [81]. Indeed, EPI is an adaptive variational
368 inference program, and may be considered to have a Bayesian uniform prior (see Section 5.1.6).

369 Biologically realistic models of neural circuits often prove formidable to analyze for two main rea-
370 sons. A primary challenge is that the number of parameters scales dramatically with the number of
371 neurons, limiting analysis of its parameter space. We see in Section 3.5 that EPI scales seemlessly
372 to high dimensional parameter spaces of RNN connectivities, while maintaining the production
373 of the specified emergent property. EPI strongly outperforms the standard likelihood-free infer-
374 ence technique (SMC-ABC [40]), and a recently developed deep likelihood-free inference technique
375 (SNPE [42]), most likely because of it's ability to leverage the gradient information of the emer-
376 gent property statistics and to adapt it's paramter sampling distribution at every step of gradient
377 descent.

378 A secondary challenge is that the structure of the parametric regimes governing emergent properties
379 is intricate. For example, even in low dimensional circuits, models can support more than one steady
380 state [82] and non-trivial dynamics on strange attractors [83]. With EPI, we use deep probabillity
381 distributions to capture the complex nonlinear parameter distributions governing model behavior.

382 In Section 3.3, we used EPI to reveal a curved parametric manifolds governing curcuit variability
383 in the stochastic stabilized supralinear network, and used hypothesis testing techniques to validate
384 our findings. In Section 3.4, we identified two regimes of SC connectivity resulting in rapid task
385 switching, and found that the full distribution of rapid task switching networks reproduced an
386 experimental result.

387 EPI leverages deep learning technology for neuroscientific inquiry in a categorically different way
388 than approaches focused on training neural networks to execute behavioral tasks [84]. These works
389 focus on examining optimized deep neural networks while considering the objective function, learn-
390 ing rule, and architecture used. This endeavor efficiently obtains sets of parameters that can be
391 reasoned about with respect to such considerations, but lacks the careful probabilistic treatment of
392 parameter inference in EPI. All of these approaches can be used complementarily to enhance the
393 practice of theoretical neuroscience.

394 **Acknowledgements:**

395 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
396 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
397 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
398 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
399 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James
400 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

401 **Data availability statement:**

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

404 **Code availability statement:**

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

406 **References**

- 407 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
408 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 409 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
410 21(1):25–45, 1998.
- 411 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 412 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in
413 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 414 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
415 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
416 models. *PLoS Comput Biol*, 3(10):e189, 2007.
- 417 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
418 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
419 sion model. *Neuron*, 82(4):809–821, 2014.
- 420 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-
421 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

- 422 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
423 networks. *Physical review letters*, 61(3):259, 1988.
- 424 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
425 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
426 17(11):4382–4388, 1997.
- 427 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration
428 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 429 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
430 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 431 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
432 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 433 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
434 model. *PLoS computational biology*, 2(7):e94, 2006.
- 435 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
436 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 437 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
438 13(8):1713–1720, 2001.
- 439 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
440 A statistical paradigm for neural spike train decoding applied to position prediction from
441 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
442 7425, 1998.
- 443 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
444 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- 445 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.
446 A point process framework for relating neural spiking activity to spiking history, neural
447 ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 448 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise corre-
449 lations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–
450 1012, 2006.

- 451 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan
452 Segev. A novel multiple objective optimization framework for constraining conductance-based
453 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 454 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature
455 analysis. *Neural computation*, 19(4):1022–1038, 2007.
- 456 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
457 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of
458 neural population activity. In *Advances in neural information processing systems*, pages
459 1881–1888, 2009.
- 460 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and
461 Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural*
462 *information processing systems*, 24:1350–1358, 2011.
- 463 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In
464 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 465 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-
466 dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,
467 9(3):e1002922, 2013.
- 468 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
469 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
470 *Science*, 349(6244):184–187, 2015.
- 471 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq
472 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in
473 visual path integration. *Neuron*, 99(1):194–206, 2018.
- 474 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
475 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th In-*
476 *ternational Conference on Machine Learning*, 2019.
- 477 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.
478 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.
479 *Nature Communications*, 10(4933), 2019.

- 480 [30] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
481 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
482 2018.
- 483 [31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Con-*
484 *ference on Learning Representations*, 2014.
- 485 [32] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropaga-
486 tion and variational inference in deep latent gaussian models. *International Conference on*
487 *Machine Learning*, 2014.
- 488 [33] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
489 neural population models through nonlinear embeddings. In *Advances in neural information*
490 *processing systems*, pages 163–171, 2016.
- 491 [34] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-
492 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 493 [35] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-
494 encoders. *bioRxiv*, page 399246, 2018.
- 495 [36] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
496 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R
497 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-
498 encoders. *Nature methods*, page 1, 2018.
- 499 [37] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
500 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
501 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 502 [38] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
503 Datta. Composing graphical models with neural networks for structured representations and
504 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 505 [39] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
506 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
507 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural*
508 *Information Processing Systems*, 2019.

- 509 [40] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
510 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 511 [41] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
512 A framework for parameter estimation and model selection from experimental data in systems
513 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 514 [42] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
515 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P
516 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
517 dynamics. *bioRxiv*, page 838383, 2019.
- 518 [43] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
519 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on*
520 *Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 521 [44] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
522 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
523 4248. PMLR, 2020.
- 524 [45] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
525 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 526 [46] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
527 networks. *International Conference on Learning Representations*, 2017.
- 528 [47] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
529 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
530 *Neuron*, 77(5):845–858, 2013.
- 531 [48] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
532 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
533 operating regime of cortex. *bioRxiv*, 2020.
- 534 [49] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
535 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
536 routing. *bioRxiv*, page 245613, 2018.

- 537 [50] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural
538 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 539 [51] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
540 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 541 [52] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-
542 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,
543 2001.
- 544 [53] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
545 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
546 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
547 1929, 2009.
- 548 [54] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
549 *International Conference on Machine Learning*, 2015.
- 550 [55] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
551 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 552 [56] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for
553 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,
554 2017.
- 555 [57] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
556 1957.
- 557 [58] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
558 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.
- 559 [59] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
560 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 561 [60] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
562 2009.
- 563 [61] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
564 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

- 565 [62] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
566 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
567 592, 2009.
- 568 [63] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear-
569 ear network: a unifying circuit motif underlying multi-input integration in sensory cortex.
570 *Neuron*, 85(2):402–417, 2015.
- 571 [64] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
572 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
573 5(10):793, 2004.
- 574 [65] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
575 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-
576 robiology*, 71(1):45–61, 2011.
- 577 [66] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-
578 tex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 579 [67] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
580 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
581 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 582 [68] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and
583 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysi-
584 ology*, 115(3):1399–1409, 2016.
- 585 [69] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing
586 Wang. Paradoxical response reversal of top- down modulation in cortical circuits with three
587 interneuron types. *Elife*, 6:1–15, 2017.
- 588 [70] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying
589 the response of mouse cortical networks to optogenetic manipulation. 2019.
- 590 [71] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
591 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
592 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.

- 593 [72] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and
594 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,
595 2015.
- 596 [73] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. Optimal control of transient dy-
597 namics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–
598 1406, 2014.
- 599 [74] Giulio Bondanelli, Thomas Deneux, Brice Bathellier, and Srdjan Ostojic. Population coding
600 and network dynamics during off responses in auditory cortex. *BioRxiv*, page 810655, 2019.
- 601 [75] David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in*
602 *neurobiology*, 25:156–163, 2014.
- 603 [76] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current*
604 *opinion in neurobiology*, 46:1–6, 2017.
- 605 [77] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H
606 Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. Motor cortex
607 embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966,
608 2018.
- 609 [78] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian com-*
610 *putation*. CRC Press, 2018.
- 611 [79] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-
612 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 613 [80] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
614 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
615 2003.
- 616 [81] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkacik. Statistical
617 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.
- 618 [82] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give
619 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of*
620 *Sciences*, 115(13):3464–3469, 2018.

- 621 [83] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emerg-
622 ent dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint*
623 *arXiv:1605.04463*, 2016.
- 624 [84] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,
625 2019.
- 626 [85] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
627 Edward Teller. Equation of state calculations by fast computing machines. *The journal of*
628 *chemical physics*, 21(6):1087–1092, 1953.
- 629 [86] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
630 1970.
- 631 [87] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
632 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
633 73(2):123–214, 2011.
- 634 [88] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-
635 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
636 820, 2011.
- 637 [89] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based infer-
638 ence. *Proceedings of the National Academy of Sciences*, 2020.
- 639 [90] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
640 free variational inference. In *Advances in Neural Information Processing Systems*, pages
641 5523–5533, 2017.
- 642 [91] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
643 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience*
644 *Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 645 [92] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
646 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience
647 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.

- 648 [93] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
649 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
650 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 651 [94] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-
652 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,*
653 *Germany*, 2019.
- 654 [95] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
655 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
656 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 657 [96] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
658 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 659 [97] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
660 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 661 [98] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
662 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
663 2012.
- 664 [99] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
665 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
666 2018.
- 667 [100] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
668 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 669 [101] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
670 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 671 [102] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
672 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 673 [103] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-
674 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv*
675 *preprint arXiv:1912.02762*, 2019.

- 676 [104] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
677 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 678 [105] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
679 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-*
680 *tion processing systems*, 29:4743–4751, 2016.
- 681 [106] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Internation-*
682 *al Conference on Learning Representations*, 2015.
- 683 [107] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for
684 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

685 **5 Methods**

686 **5.1 Emergent property inference (EPI)**

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

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

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

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

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

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

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

716 Given a choice of model $p(\mathbf{x} \mid \mathbf{z})$ and emergent property of interest \mathcal{X} , $q_{\theta}(\mathbf{z})$ is optimized via
717 the neural network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep
718 variational family \mathcal{Q} producing the emergent property \mathcal{X} :

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

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

719 Entropy is chosen as the normative selection principle, since we want the posterior to only contain
720 structure predicated by the emergent property [57, 58]. This choice of selection principle is also
721 that of standard Bayesian inference, and we derive an exact relation between EPI and variational
722 inference (see Section 5.1.5). However, a key difference is that variational inference and other
723 Bayesian methods do not constrain the predictions of their inferred posteriors. This optimization
724 is executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [46].

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

733 5.1.1 Related approaches

734 When Bayesian inference problems lack conjugacy, scientists use approximate inference methods
735 like variational inference (VI) [45] and Markov chain Monte Carlo (MCMC) [85, 86]. After opti-
736 mization, variational methods return a parameterized posterior distribution, which we can analyze.
737 Also, the variational approximating distribution class is often chosen such that it permits fast
738 sampling. In contrast MCMC methods only produce samples from the approximated posterior dis-
739 tribution. No parameterized distribution is estimated, and additional samples are always generated
740 with the same sampling complexity. Inference in models defined by systems of differential has been

741 demonstrated with MCMC [87], although this approach requires tractable likelihoods. Advances
742 have leveraged structure in stochastic differential equation models to improve likelihood
743 approximations, thus expanding the domain of applicable models [88].

744 Likelihood-free (or “simulation-based”) inference (LFI) [89] is model parameter inference in the
745 absence of a tractable likelihood function. The most prevalent approach to LFI is approximate
746 Bayesian computation [79], in which satisfactory parameter samples are kept from random prior
747 sampling according to a rejection heuristic. The obtained set of parameters do not have a prob-
748 abilities, and further insight about the model must be gained from examination of the parameter
749 set and their generated activity. Methodological advances to ABC methods have come through
750 the use of Markov chain Monte Carlo (MCMC-ABC) [80] and sequential Monte Carlo (SMC-ABC)
751 [40] sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still
752 struggles to scale in dimensionality (cf. Fig. 4). Furthermore, once a parameter set has been
753 obtained by SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again
754 with a new population of initialized particles to obtain additional samples.

755 For scientific model analysis, we seek a posterior distribution exhibiting the properties of a well-
756 chosen variational approximation: a parametric form conferring analytic calculations, and trivial
757 sampling time. For this reason, ABC and MCMC techniques are unattractive, since they only
758 produce a set of parameter samples and have unchanging sampling rate. EPI executes likelihood-
759 free inference using the MEFN [46] algorithm using a deep variational posterior approximation.
760 The deep neural network of EPI defines the parametric form of the posterior approximation. Fur-
761 thermore, the EPI distribution is constrained to produce an emergent property. In other words,
762 the summary statistics of the posterior predictive distribution are fixed to have certain first and
763 second moments. EPI optimization is enabled using stochastic gradient techniques in the spirit
764 of likelihood-free variational inference [90]. The analytic relationship between EPI and variational
765 inference is explained in Secton 5.1.6.

766 We note that, during our preparation and early presentation of this work [91, 92], another work
767 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
768 circuits ([93, 94, 42]). We are encouraged by this general problem being recognized by others in the
769 community, and we emphasize that these works offer complementary neuroscientific contributions
770 (different theoretical models of focus) and use different technical methodologies (ours is built on
771 our prior work [46], theirs similarly [95]).

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

773 recently developed LFI methods in which two neural networks are used for posterior inference.
774 This first neural network is a normalizing flow used to estimate the posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE)
775 or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural likelihood (SNL [43])). A recent advance uses an
776 unconstrained neural network to estimate the likelihood ratio (sequential neural ratio estimation
777 (SNRE [44])). In SNL and SNRE, MCMC sampling techniques are used to obtain samples from
778 the approximated posterior. This contrasts with EPI and SNPE, which afford a normalizing flow
779 approximation to the posterior, which facilitates immediate measurements of sample probability,
780 gradient, or Hessian for system analysis. The second neural network in this sequential class of
781 methods is the amortizer. This network maps data \mathbf{x} (or statistics $f(\mathbf{x}; \mathbf{z})$ or model parameters \mathbf{z})
782 to the weights and biases of the first neural network. These methods are optimized on a conditional
783 density (or ratio) estimation objective on a sequentially adapting finite sample-based approximation
784 to the posterior.

785 The approximating fidelity of the first neural network in sequential approaches is optimized to
786 generalize across the entire distribution it is conditioned upon. This optimization towards gen-
787 eralization of sequential methods can reduce the accuracy at the singular posterior of interest.
788 Whereas in EPI, the entire expressivity of the normalizing flow is dedicated to learning a single
789 distribution as well as possible. While amortization is not possible in EPI parameterized by the
790 mean parameter μ (due to the inverse mapping problem [96]), we have shown this two-network
791 amortization approach to be effective in exponential family distributions defined by their natural
792 parameterization [97].

793 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in
794 natural models. Around a point, one can measure the Jacobian. One approach that scales well is
795 EAR [98]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [99] and
796 its stochastic adaptation [100]. Casting identifiability as a statistical estimation problem, the profile
797 likelihood can assess via iterated optimization while holding parameters fixed [53]. An exciting
798 recent method is capable of recovering the functional form of such unidentifiabilities away from a
799 point by following degenerate dimensions of the fisher information matrix [101]. Global structural
800 non-identifiabilities can be found for models with polynomial or rational dynamics equations using
801 DAISY [102]. With EPI, we have all the benefits given by a statistical inference method plus the
802 ability to query the gradient or Hessian of the inferred distribution at any chosen parameter value.

803 **5.1.2 Normalizing flows**

804 Deep probability distributions are comprised of multiple layers of fully connected neural networks
 805 (Equation). When each neural network layer is restricted to be a bijective function, the sample
 806 density can be calculated using the change of variables formula at each layer of the network. For
 807 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

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

808 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 809 restricting our layers to normalizing flows [54, 103] – bijective functions with fast log determinant
 810 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log
 811 probability calculation confers efficient optimization of the maximum entropy objective (see Section
 812 5.1.3). We use the Real NVP [55] normalizing flow class, because its coupling architecture confers
 813 both fast sampling (forward) and fast log probability evaluation (backward). Fast probability
 814 evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout
 815 parameter space. Glow permutations were used in between coupling stages [104]. This is in contrast
 816 to autoregressive architectures [56, 105], in which only forward or backward passes are efficient. In
 817 this work, normalizing flows are used as flexible posterior approximations $q_{\theta}(\mathbf{z})$ having weights and
 818 biases θ . We specify the architecture used in each application by the number of Real-NVP affine
 819 coupling stages, and the number of neural network layers and units per layer of the conditioning
 820 functions.

821 **5.1.3 Augmented Lagrangian optimization**

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

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

824 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
 825 Lagrange multipliers where $m = |\mu_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})| = 2|f(\mathbf{x}; \mathbf{z})|$, and c is the penalty coefficient.
 826 The sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter μ_{opt} are determined by the means μ and
 827 variances σ^2 of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 13. Specifically, $T(\mathbf{x}; \mathbf{z})$ is
 828 a concatenation of the first and second moments, μ_{opt} is a concatenation of μ and σ^2 (see section
 829 5.1.5), and the Lagrange multipliers are closely related to the natural parameters η of exponential

830 families (see Section 5.1.6). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution are optimized
831 according to Equation 15 using the Adam optimizer with learning rate 10^{-3} [106].

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

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

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

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

837 The lagrangian parameters $\boldsymbol{\eta}_{\text{opt}}$ are initialized to zero and adapted following each augmented
838 Lagrangian epoch, which is a period of optimization with fixed $(\boldsymbol{\eta}_{\text{opt}}, c)$ for a given number of
839 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
840 after each epoch based on constraint error reduction. The penalty coefficient is updated based
841 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
842 $\mathbb{E}[|R(\boldsymbol{\theta}_{k+1})|] > \gamma \mathbb{E}[|R(\boldsymbol{\theta}_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
843 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
844 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
845 according to application.

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

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

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

867 When assessing the suitability of EPI for a particular modeling question, there are some important
 868 technical considerations. First and foremost, as in any optimization problem, the defined emergent
 869 property should always be appropriately conditioned (constraints should not have wildly different
 870 units). Furthermore, if the program is underconstrained (not enough constraints), the distribution
 871 grows (in entropy) unstably unless mapped to a finite support. If overconstrained, there is no pa-
 872 rameter set producing the emergent property, and EPI optimization will fail (appropriately). Next,
 873 one should consider the computational cost of the gradient calculations. In the best circumstance,
 874 there is a simple, closed form expression (e.g. Section 5.2.5) for the emergent property statistic
 875 given the model parameters. On the other end of the spectrum, many forward simulation iterations
 876 may be required before a high quality measurement of the emergent property statistic is available
 877 (e.g. Section 5.2.1). In such cases, backpropagating gradients through the SDE evolution will be
 878 expensive.

879 5.1.4 Example: 2D LDS

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

$$882 \quad \tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (18)$$

882 with

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

883 To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fix-
 884 ing $\tau = 1$), the emergent property statistics $T(\mathbf{x})$ were chosen to contain the first and second
 885 moments of the oscillatory frequency, $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscil-
 886 lating 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 0.25², and the oscillation frequency $2\pi\text{imag}(\lambda_1)$ having mean $\omega = 1$ Hz with variance (0.1Hz)²:

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

891

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 the quadratic formula:

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

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 does not require simulation, we cannot derive the distribution q_{θ}^* directly. This fact is due to the formally hard problem of the backward mapping: finding the natural parameters η from the mean parameters $\boldsymbol{\mu}$ of an exponential family distribution [96]. Instead, we used EPI to approximate this distribution (Fig. S1B). We used a real-NVP normalizing flow architecture with four masks, two neural network layers of 15 units per mask, with batch normalization momentum 0.99, mapped onto a support of $z_i \in [-10, 10]$. (see Section 5.1.2).

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

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

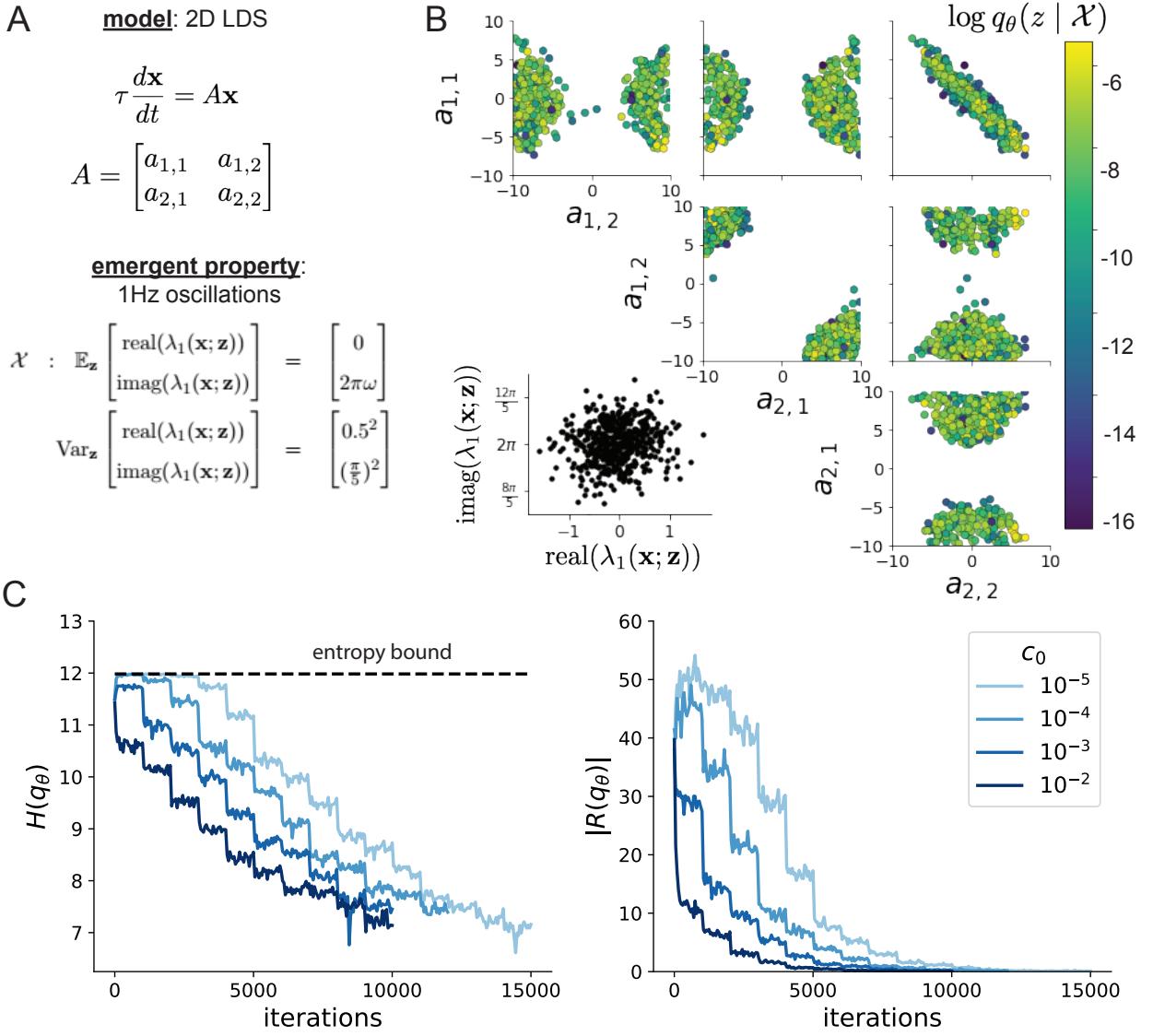


Figure 5: (LDS1): **A.** Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. **B.** The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed lines indicate the parameter axes. **C.** Entropy throughout the optimization. At the beginning of each augmented Lagrangian epoch (2,000 iterations), the entropy dipped due to the shifted optimization manifold where emergent property constraint satisfaction is increasingly weighted. **D.** Emergent property moments throughout optimization. At the beginning of each augmented Lagrangian epoch, the emergent property moments adjust closer to their constraints.

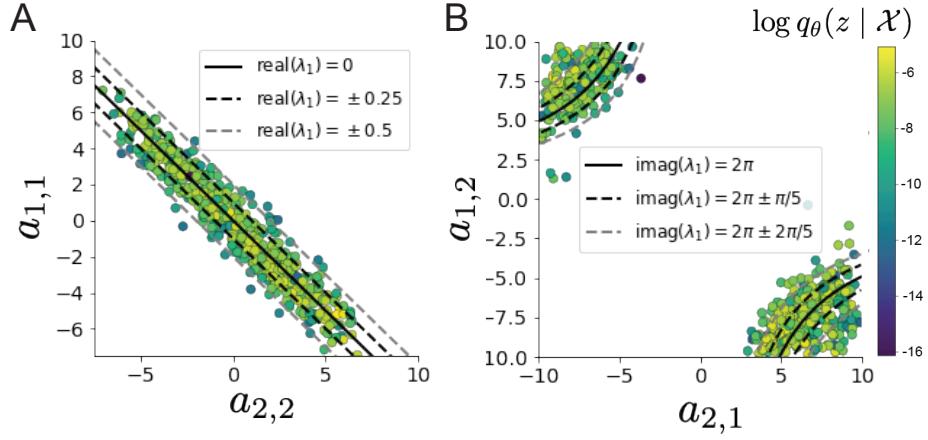


Figure 6: (LDS2): **A.** Probability contours in the a_1 - a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. **B.** Probability contours in the a_2 - a_3 plane were derived from the emergent property statistic of oscillation frequency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

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

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

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

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

914 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
915 $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_\theta} [|a_1a_4 - E_{q_\theta}[a_1a_4]|]$,
916 we show the contours as $a_1a_4 = 0$ (black), $a_1a_4 = -\sigma_{1,4}$ (black dotted), and $a_1a_4 = -2\sigma_{1,4}$ (grey
917 dotted). This validates the curved structure of the inferred distribution learned through EPI. We
918 took steps in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive
919 values a_1a_4 in the learned distribution. Subtler combinations of model and emergent property will
920 have more complexity, further motivating the use of EPI for understanding these systems. As we
921 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz
922 (Fig. S3).

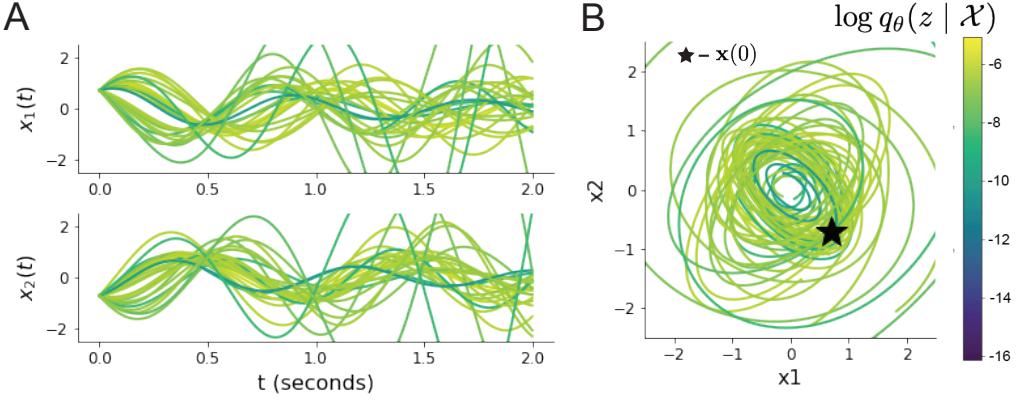


Figure 7: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_\theta(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. **A.** Each dimension of the simulated trajectories throughout time. **B.** The simulated trajectories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

923 5.1.5 Maximum entropy distributions and exponential families

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

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

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

926 will have probability density in the exponential family:

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

927 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
928 formally hard to identify [96].

929 In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equation
930 4

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

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

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

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

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

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

934 which are constrained to the chosen means and variances

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

935 5.1.6 EPI as variational inference

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

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

940 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
 941 However, conjugacy is rare in practice, and alternative methods, such as variational inference [107],
 942 are utilized.

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

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

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

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

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

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

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

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

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

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

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

954 where

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

$$955 \quad T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \quad (38)$$

956 The variational objective is then

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

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

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

959 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
 960 (Equation 28). Thus, EPI is implicitly executing variational inference with a uniform prior and a
 961 mean-field gaussian likelihood on the emergent property statistics. The data \mathbf{x} used by this implicit
 962 variational inference program would be that generated by the adapting variational approximation
 963 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{z})q_\theta(\mathbf{z})$, and the likelihood parameters $\boldsymbol{\eta}_f$ of EPI optimization epoch k are predicated
 964 by $\boldsymbol{\eta}_{\text{opt},k}$. However, in EPI we have not specified a prior distribution, or collected data, which can
 965 inform us about model parameters. Instead we have a mathematical specification of an emergent
 966 property, which the model must produce, and a maximum entropy selection principle. Accordingly,
 967 we replace the notation of $p(\mathbf{z} | \mathbf{x})$ with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys
 968 emergent property \mathcal{X} (see Section 5.1).

969 5.2 Theoretical models

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

972 **5.2.1 Stomatogastric ganglion**

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

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

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

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

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

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

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

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

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

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

984 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 985 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 986 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 987 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
 988 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
 989 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

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

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

997
$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (49)$$

998
$$N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_3}{v_4} \right) \right) \quad (50)$$

999
$$\lambda_N = \phi_N \cosh \left(\frac{x_\alpha - v_3}{2v_4} \right) \quad (51)$$

1000
$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (52)$$

1001
$$H_\infty = \frac{1}{1 + \exp \left(\frac{x_\alpha + v_5}{v_6} \right)} \quad (53)$$

1002
$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp \left(\frac{-x_\alpha + v_7}{v_8} \right)} \right). \quad (54)$$

1003 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,
 1004 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

1005 Finally, there is a synaptic gating variable as well:

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

1006 When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear
 1007 dynamical system. Gaussian noise $d\mathbf{B}$ of variance $(1 \times 10^{-12})^2 \text{ A}^2$ makes the model stochastic, and
 1008 introduces variability in frequency at each parameterization \mathbf{z} .

1009 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for
 1010 $T = 300$ time steps of $dt = 25\text{ms}$. The chosen dt and T were the most computationally convenient
 1011 choices yielding accurate frequency measurement. We used a basis of complex exponentials with
 1012 frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

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

1013 To measure spiking frequency, we processed simulated membrane potentials with a relu (spike
 1014 extraction) and low-pass filter with averaging window of size 20, then took the frequency with the
 1015 maximum absolute value of the complex exponential basis coefficients of the processed time-series.
 1016 The first 20 temporal samples of the simulation are ignored to account for initial transients.

1017 To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in$
 1018 $\mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in$

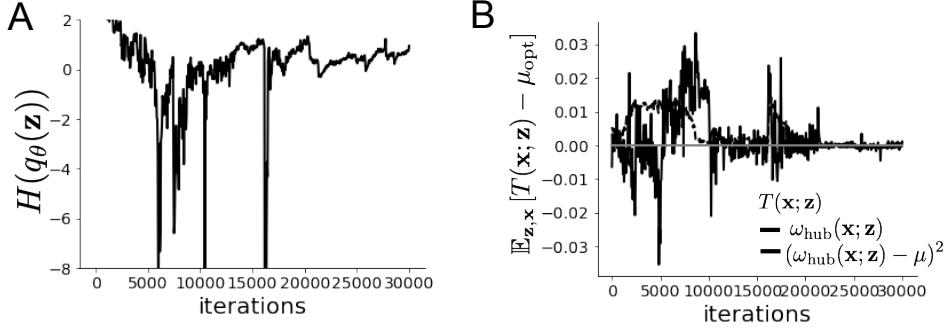


Figure 8: (STG1): EPI optimization of the STG model producing network syncing. **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

1019 $\{f1, f2, \text{hub}, s1, s2\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

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

1020 where $i = [0, 1, \dots, 100]$. The frequency is then calculated as

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

1021 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
 1022 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
 1023 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be $(0.025\text{Hz})^2$ to
 1024 capture variation in frequency between 0.5Hz and 0.6Hz (Equation 4). As a maximum entropy dis-
 1025 tribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments of the hub neuron frequency
 1026 (as in Equations 28 and 29)

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

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

1027 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 1028 epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

1029 For EPI in Fig 1E, we used a real NVP architecture with three Real NVP coupling layers and two-
 1030 layer neural networks of 25 units per layer. The normalizing flow architecture mapped $z_0 \sim \mathcal{N}(\mathbf{0}, I)$

1032 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
 1033 returned by a preliminary ABC search. We did not include $g_{\text{synA}} < 0.01$, for numerical stability.
 1034 EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an
 1035 augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The distribution
 1036 shown is that of the architecture converging with criteria $N_{\text{test}} = 100$ at greatest entropy across
 1037 random seeds.

1038 We calculated the Hessian at the mode of the inferred EPI distribution. The Hessian of a probability
 1039 model is the second order gradient of the log probability density $\log q_{\boldsymbol{\theta}}(\mathbf{z})$ with respect to the
 1040 parameters \mathbf{z} : $\frac{\partial^2 \log q_{\boldsymbol{\theta}}(\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^\top}$. With EPI, we can examine the Hessian, which is analytically available
 1041 throughout distribution, to indicate the dimensions of parameter space that are sensitive (strongly
 1042 negative eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the
 1043 emergent property produced. In Figure 1D, the eigenvectors of the Hessian v_1 (solid) and v_2
 1044 (dashed) are shown evaluated at the mode of the distribution. The length of the arrows is inversely
 1045 proportional to the square root of absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$.
 1046 Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D parameter
 1047 space are chosen arbitrarily.

1048 5.2.2 Primary visual cortex

1049 In the stochastic stabilized supralinear network [71], population rate responses \mathbf{x} to input \mathbf{h} , recur-
 1050 rent input $W\mathbf{x}$ and slow noise $\boldsymbol{\epsilon}$ are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \quad (61)$$

1051 where the noise is an Ornstein-Uhlenbeck process $\boldsymbol{\epsilon} \sim OU(\tau_{\text{noise}}, \boldsymbol{\sigma})$

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (62)$$

1052 with $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$. The noisy process is parameterized as

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (63)$$

1053 so that $\boldsymbol{\sigma}$ parameterizes the variance of the noisy input in the absence of recurrent connectivity
 1054 ($W = \mathbf{0}$). As contrast increases, input to the E- and P-populations increases relative to a baseline
 1055 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
 1056 deterministic V1 circuit model [48]

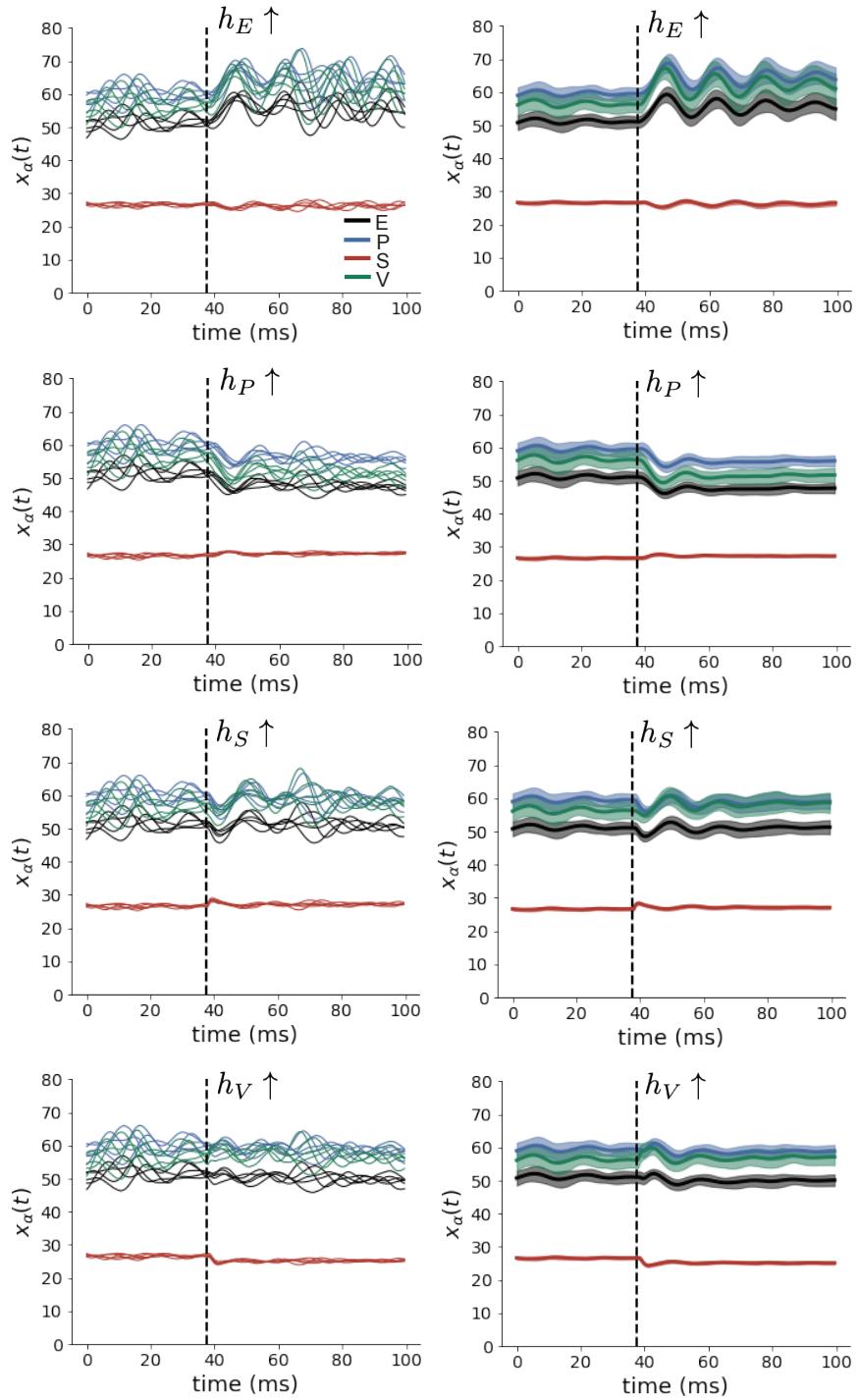


Figure 9: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

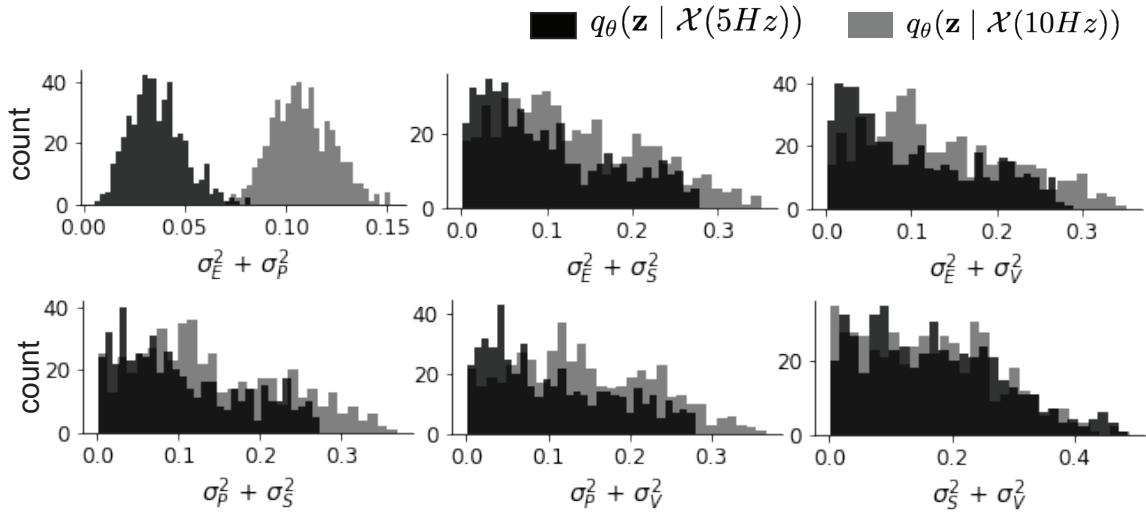


Figure 10: (V1 2) Posterior predictive distributions of the sum of squares of each pair of noise parameters.

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (64)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (65)$$

¹⁰⁵⁷ and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (66)$$

¹⁰⁵⁸ To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equivalence class
¹⁰⁵⁹

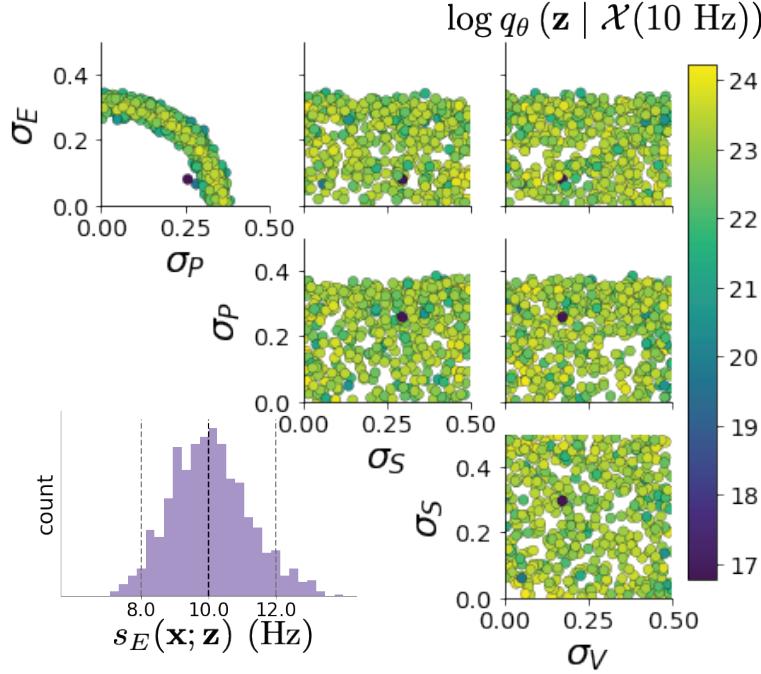


Figure 11: (V1 3) EPI posterior for $\mathcal{X}(10 \text{ Hz})$.

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (67)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (68)$$

¹⁰⁶⁰ and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (69)$$

¹⁰⁶¹ Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5\text{ms}$ from an initial condition

1062 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
 1063 as the square root of the temporal variance from $t_{ss} = 75\text{ms}$ to $Tdt = 100\text{ms}$ averaged over 100
 1064 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[\sqrt{\mathbb{E}_{t > t_{ss}} \left[(x_E(t) - \mathbb{E}_{t > t_{ss}} [x_E(t)])^2 \right]} \right] \quad (70)$$

1065 For EPI in Fig 2D-E, we used a real NVP architecture with three Real NVP coupling layers
 1066 and two-layer neural networks of 50 units per layer. The normalizing flow architecture mapped
 1067 $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ 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
 1068 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an augmented Lagrangian coefficient of
 1069 $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the architectures
 1070 converging with criteria $N_{\text{test}} = 100$ at greatest entropy across random seeds.

1071 In Fig. 2E, we visualize the modes of $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ throughout the σ_E - σ_P marginal. Specifically, we
 1072 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (71)$$

1073 At each mode \mathbf{z}^* , we calculated the Hessian and visualized the sensitivity dimension in the direction
 1074 of positive σ_E .

1075 5.2.3 Primary visual cortex: challenges to analysis

1076 TODO Agostina and I are putting this together now.

1077 5.2.4 Superior colliculus

1078 In the model of Duan et al [49], there are four total units: two in each hemisphere corresponding to
 1079 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right
 1080 Pro (RP) and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related
 1081 by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh \left(\frac{u_\alpha - a}{b} \right) + \frac{1}{2} \right) \quad (72)$$

1082 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-
 1083 earity, respectively. During periods of optogenetic inactivation, activity was decreased proportional
 1084 to the optogenetic strength γ

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (73)$$

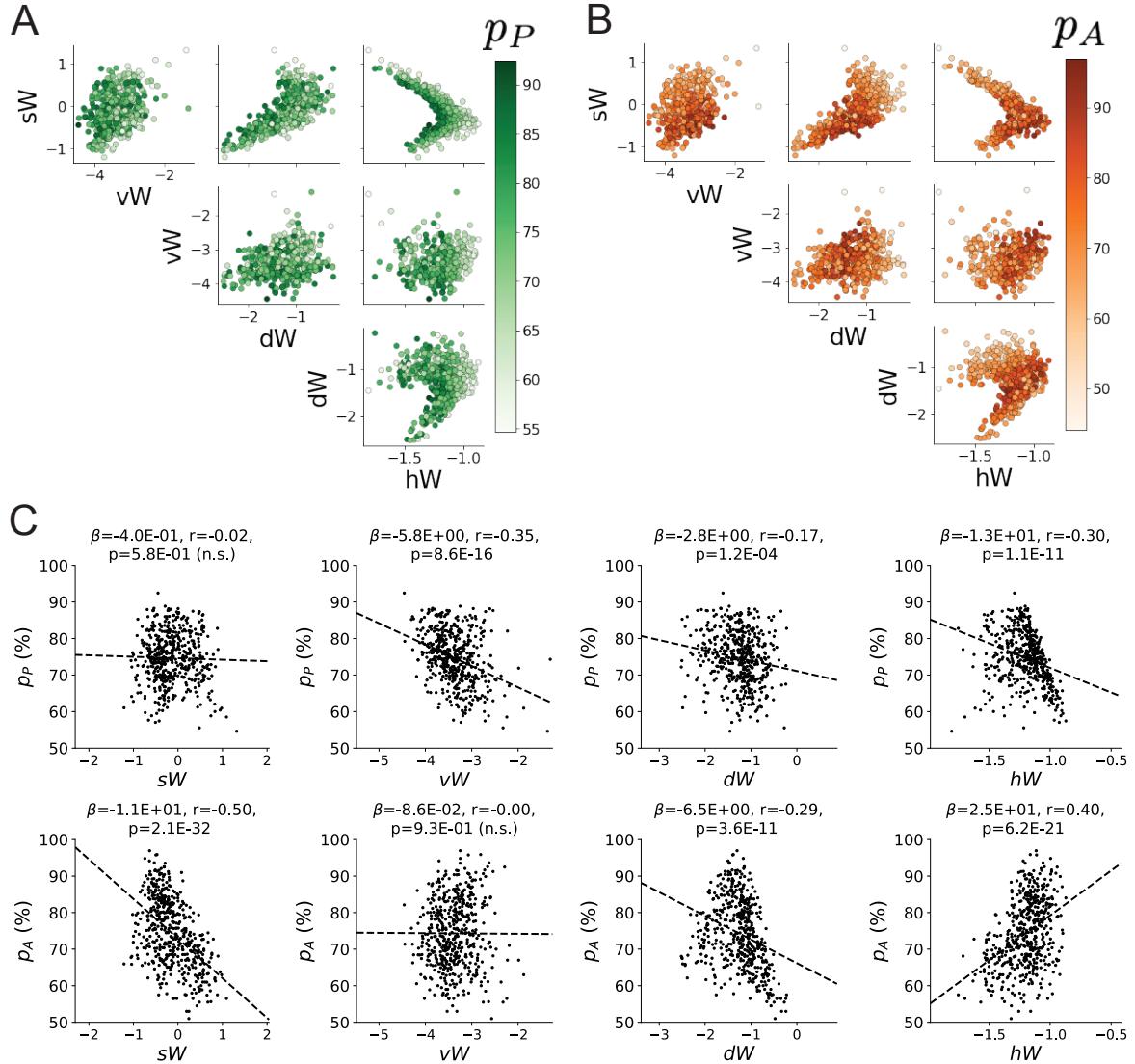


Figure 12: (SC1): **A.** Same pairplot as Fig. 3C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

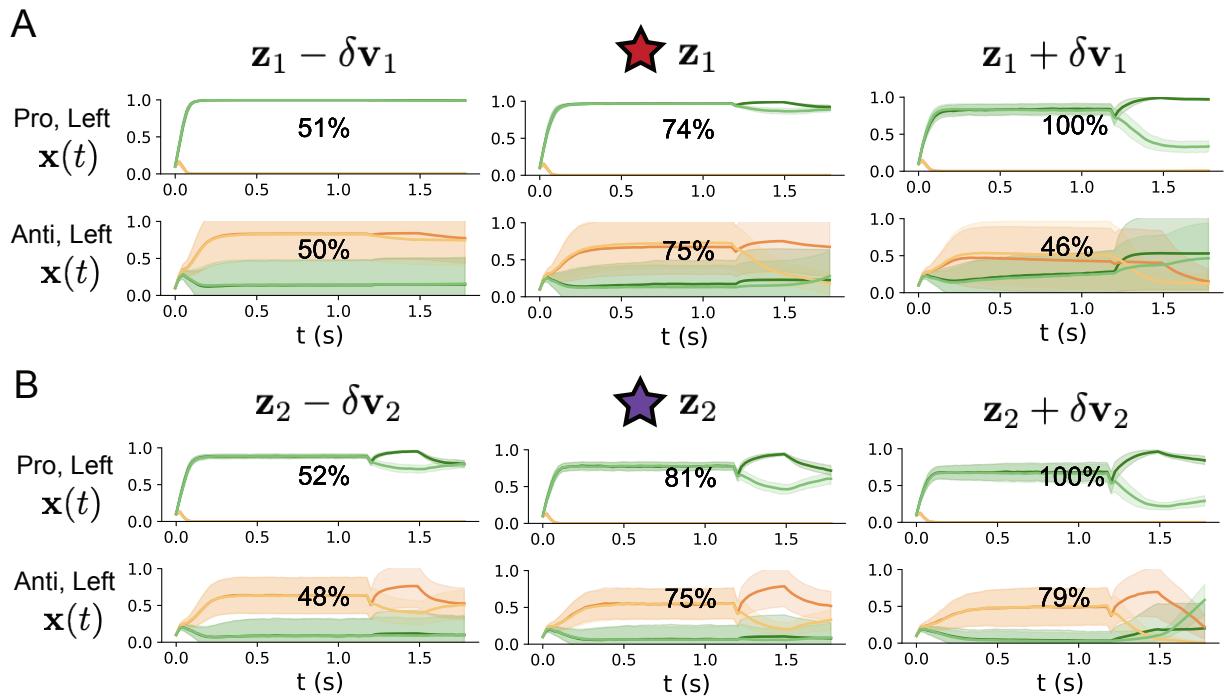


Figure 13: (SC2): **A.** Simulations in network regime \mathbf{z}_1 (center) with simulations given connectivity perturbations in the negative direction of the sensitivity vector \mathbf{v}_1 (left) and positive direction (right). **B.** Same as A for network regime \mathbf{z}_2 .

1085 We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (74)$$

1086 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (75)$$

1087 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2 . The weight
1088 matrix has 4 parameters sW , vW , hW , and dW :

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (76)$$

1089 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (77)$$

1090 There is a constant input to every population,

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

1091 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (79)$$

1092 rule-based input depending on the condition

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

1093

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

1094 a choice-period input

$$\mathbf{h}_{\text{choice}}(t) = \begin{cases} I_{\text{choice}}[1, 1, 1, 1]^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (82)$$

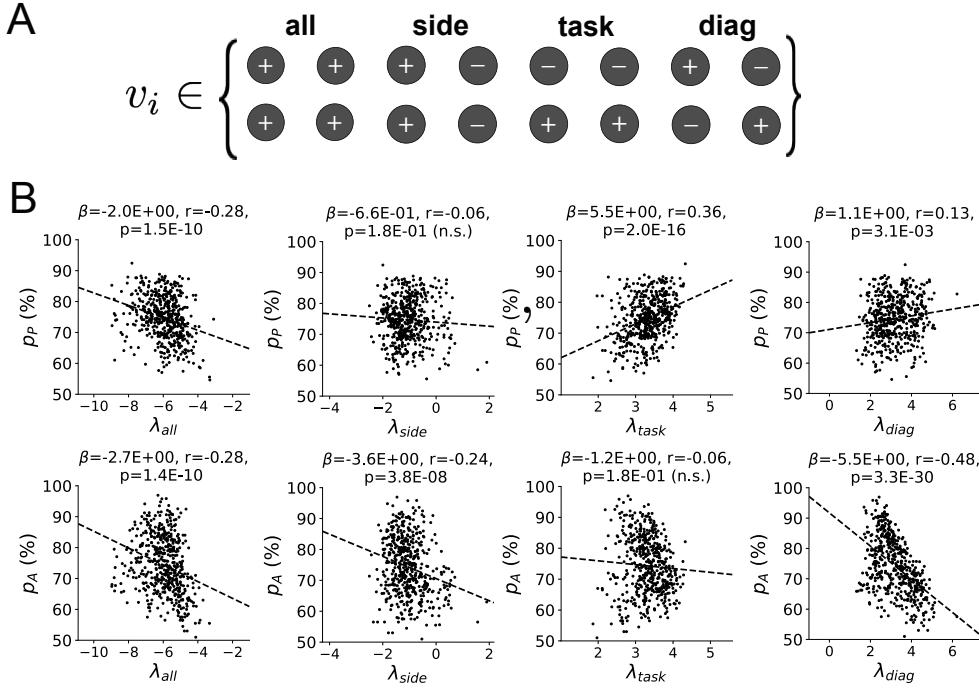


Figure 14: (SC3): **A.** Invariant eigenvectors of connectivity matrix W . **B.** Eigenvalues of connectivities of EPI distribution versus task accuracies.

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

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^T, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^T, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases} \quad (83)$$

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

The accuracies of p_P and p_A are calculated as

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

and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (85)$$

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

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (86)$$

where $\beta = 100$.

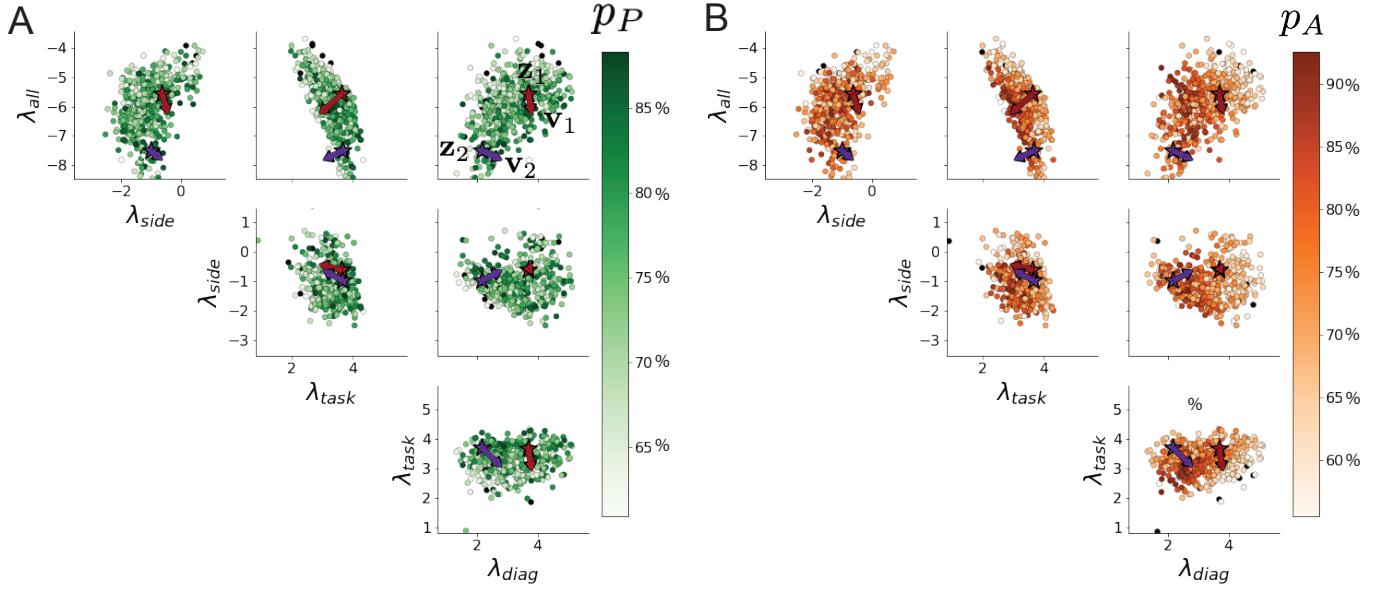


Figure 15: (SC4): **A.** Pairplots of eigenvalues of connectivity matrices in EPI distribution colored by Pro task accuracy. Red and purple stars and arrows correspond to eigenvalues and sensitivity directions \mathbf{z}_1 , \mathbf{z}_2 , \mathbf{v}_1 , and \mathbf{v}_2 . **B.** Same colored by Anti task accuracy.

1103 Writing the EPI posterior as a maximum entropy distribution, $T(\mathbf{x}; \mathbf{z})$ is comprised of both these
1104 first and second moments of the accuracy in each task (as in Equations 28 and 29)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p(\mathbf{x}; \mathbf{z})_P \\ p(\mathbf{x}; \mathbf{z})_A \\ (p(\mathbf{x}; \mathbf{z})_P - 75\%)^2 \\ (p(\mathbf{x}; \mathbf{z})_A - 75\%)^2 \end{bmatrix}, \quad (87)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 7.5\%^2 \\ 7.5\%^2 \end{bmatrix}. \quad (88)$$

1105 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each epoch of 2,000 iterations(see Section 5.1.3). The optimization converged after six epochs (Fig. 17).

1108 For EPI in Fig. 3C, 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 1109 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 1110 augmented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution 1111

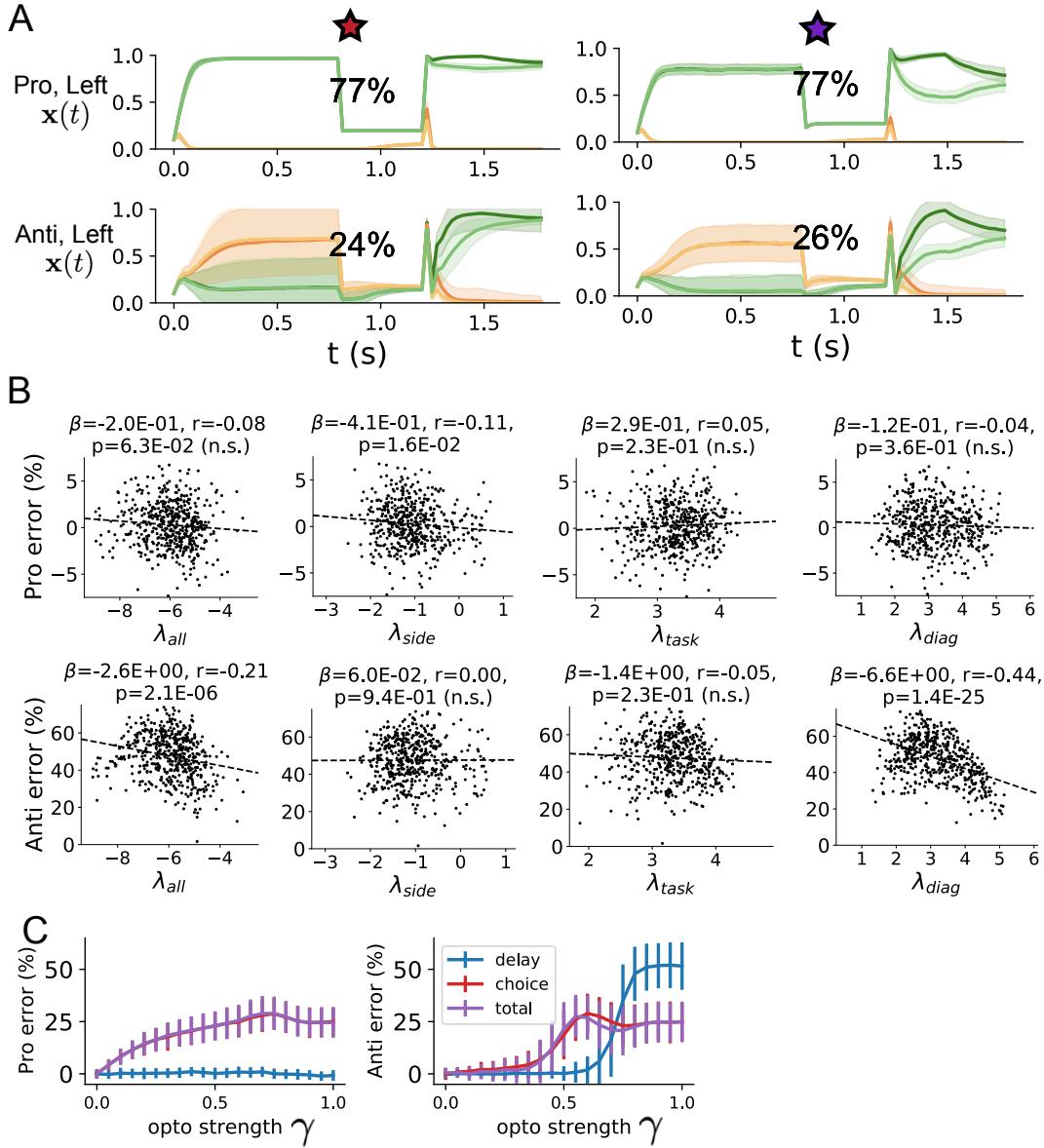


Figure 16: (SC5): **A.** Response of each parameter regime to optogenetic silencing during the delay period. **B.** Connectivity eigenvalues versus the task error induced by delay period inactivation. **C.** Error induced by delay period inactivation with increasing optogenetic strength. Means and standard deviations are calculated across the entire EPI posterior.

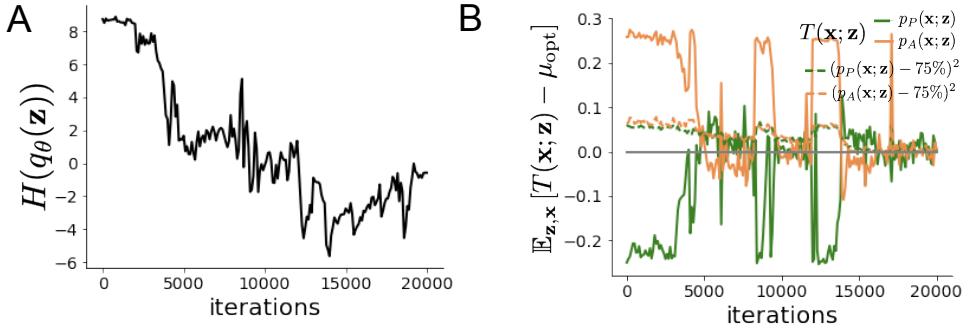


Figure 17: (SC6): **A.** Entropy throughout optimization. **B.** The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

1112 shown is that of the architecture converging with criteria $N_{\text{test}} = 25$ at greatest entropy across
1113 random seeds.

1114 To make sense of this inferred distribution, we identified two modes used to represent the two
1115 regimes of connectivity in this posterior:

$$\begin{aligned} \mathbf{z}_1 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW > 0 \end{aligned} \tag{89}$$

1116 and

$$\begin{aligned} \mathbf{z}_2 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW < 0 \end{aligned} \tag{90}$$

1117 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-
1118 tion of the symmetric connectivity matrices $W = V\Lambda V^{-1}$, which results in the same basis vectors
1119 \mathbf{v}_i for all W parameterized by \mathbf{z} (Fig. 14A). These basis vectors have intuitive roles in processing
1120 for this task, and are accordingly named the *all* mode - all neurons co-fluctuate, *side* mode - one
1121 side dominates the other, *task* mode - the Pro or Anti populations dominate the other, and *diag*
1122 mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair. We found
1123 significant trends across the EPI posterior connectivities: the eigenvalues λ_{task} and λ_{diag} were cor-
1124 related with p_P , while λ_{all} was anticorrelated with p_P . λ_{all} , λ_{side} , and λ_{diag} were all significantly
1125 anticorrelated with p_A .

1126 Under this decomposition, we can re-visualize the posterior in eigenvalue space (Fig. 15). Fur-
1127 thermore, we can project the dimensions of sensitivity into eigenvalue space as well, giving us a
1128 more intuitive sense of how connectivity affects computation in each regime. We see that sensitivity

1129 dimensions \mathbf{v}_1 and \mathbf{v}_2 , which cause p_P to increase and a regime dependent change in p_A , both point
 1130 in the direction of increasing λ_{side} and decreasing λ_{task} . These eigenvalue changes are evident in
 1131 the simulations of connectivity perturbations away from the modes (Fig. 13). As the component
 1132 of connectivity along \mathbf{v}_1 and \mathbf{v}_2 becomes stronger (left-to-right), there is less separation between
 1133 Pro an Anti populations (lower λ_{task}) and greater separation between Left and Right populations
 1134 following stimulus presentation (greater λ_{side}). A key differentiating factor is that \mathbf{v}_1 substantially
 1135 increases λ_{diag} , while \mathbf{v}_2 does not.

1136 During optogenetic silencing simulations, activations $x_\alpha(t)$ were set to a fraction of their values $(1 -$
 1137 $\gamma)$, where γ is the optogenetic perturbation strength. We found that λ_{all} and λ_{diag} were significantly
 1138 anticorrelated with Anti error during delay period inactivation. Delay period inactivation was from
 1139 $0.8 < t < 1.2$, choice period inactivation was for $t > 1.2$ and total inactivation was for the entire
 1140 trial.

1141 5.2.5 Scaling EPI for stable amplification in RNNs

1142 We examine the scaling properties of EPI by learning connectivities of RNNs of increasing size
 1143 that exhibit stable amplification. Rank-2 RNN connectivity is modeled as $W = UV^\top$, where
 1144 $U = [\mathbf{u}_1 \ \mathbf{u}_2] + g\chi^{(W)}$, $V = [\mathbf{v}_1 \ \mathbf{v}_2] + g\chi^{(V)}$, and $\chi_{i,j}^{(W)}, \chi_{i,j}^{(V)} \sim \mathcal{N}(0, 1)$. This RNN model has
 1145 dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x}. \quad (91)$$

1146 In this analysis, we infer 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
 1147 produce stable amplification.

1148 For this RNN model to be stable, all real eigenvalues of W must be less than 1: $\text{real}(\lambda_1) < 1$, where
 1149 λ_1 denotes the greatest real eigenvalue of W . For a stable RNN to amplify at least one input pattern,
 1150 the symmetric connectivity $W^s = \frac{W+W^\top}{2}$ must have an eigenvalue greater than 1: $\lambda_1^s > 1$, where
 1151 λ^s is the maximum eigenvalue of W^s . These two conditions are necessary and sufficient for stable
 1152 amplification in RNNs [50]. We define the emergent property of stable amplification with means
 1153 of these eigenvalues (0.5 and 1.5, respectively) that satisfy these conditions and variances (0.25^2)
 1154 about those means such that samples rarely violate the eigenvalue constraints. As a maximum

1155 entropy distribution, this is written as

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) \\ \lambda_1^s(\mathbf{x}; \mathbf{z}) \\ (\text{real}(\lambda_1)(\mathbf{x}; \mathbf{z}) - 0.5)^2 \\ (\lambda_1^s(\mathbf{x}; \mathbf{z}) - 1.5)^2 \end{bmatrix}, \quad (92)$$

1156

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.25^2 \\ 0.25^2 \end{bmatrix}. \quad (93)$$

1157 Gradients of maximum eigenvalues of hermitian matrices like W^s are available with modern auto-
 1158 automatic differentiation tools. To differentiate through the $\text{real}(\lambda_1)$, we solve the following equation
 1159 for eigenvalues of rank-2 matrices using the rank reduced matrix $W^r = V^\top U$

$$\lambda_{\pm} = \frac{\text{Tr}(W^r) \pm \sqrt{\text{Tr}(W^r)^2 - 4\text{Det}(W^r)}}{2}. \quad (94)$$

1160 For EPI in Fig. 4, we used a real NVP architecture with three coupling layers of affine transfor-
 1161 mations parameterized by two-layer neural networks of 100 units per layer. The initial distribution
 1162 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 an
 1163 augmented Lagrangian coefficient of $c_0 = 10^3$, a batch size $n = 200$, $\beta = 4$, and chose to use 500
 1164 iterations per augmented Lagrangian epoch.

1165 ABC methods define a tolerance ϵ from observed data x_0 for which we keep sampled parameters.
 1166 To make this ABC approach as similar as possible to the EPI program defined by Equation 10, we
 1167 chose $\epsilon = 0.5$, an l -2 distance metric, and

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \quad (95)$$

1168 located at the mean of our desired emergent property. We use sequential Monte Carlo ABC (SMC-
 1169 ABC), to compare efficiency, since it is considered the most efficient ABC approach. SNPE [42] is
 1170 another deep likelihood-free inference method that emerged along with this work. In contrast to
 1171 EPI, SNPE cannot condition on the variance of the posterior predictive distribution. Also, there
 1172 is no tolerance parameter for SNPE like ϵ in ABC, so the comparative SNPE approach simply
 1173 conditions on observation x_0 .

1174 How we did EPI

1175

1176 **How we did SNPE**

1177 SMC-ABC has many hyperparameters, of which pyABC selects automatically by running some ini-
1178 tial diagnostics upon initialization. In concurrence with the literature, SMC-ABC fails to converge
1179 around 25-30 dimensions, since it's proposal samples never get close enough to the target statistics.

1180 **How we did SNPE**

1181 We searched over many SNPE hyperparameter choices: $n_{\text{train}} \in [2,000, 10,000, 100,000]$ is the
1182 number of simulations run per training epoch, and $n_{\text{mades}} \in [2, 3]$ is the number of masked autore-
1183 gressive density estimators in the deep parameter distribution architecture. The greater n_{train} , the
1184 longer each epoch will take, but the more likely SNPE may converge during that epoch. Greater
1185 n_{mades} increases the flexibility of the deep parameter distribution of SNPE, but slows optimization.
1186 For the timing plot, we show the fastest among all of these choices, and for the convergence plot,
1187 we show the best convergence among all of these choices. During optimization, we used $n_{\text{atom}}=100$
1188 atomic proposals as is recommended.

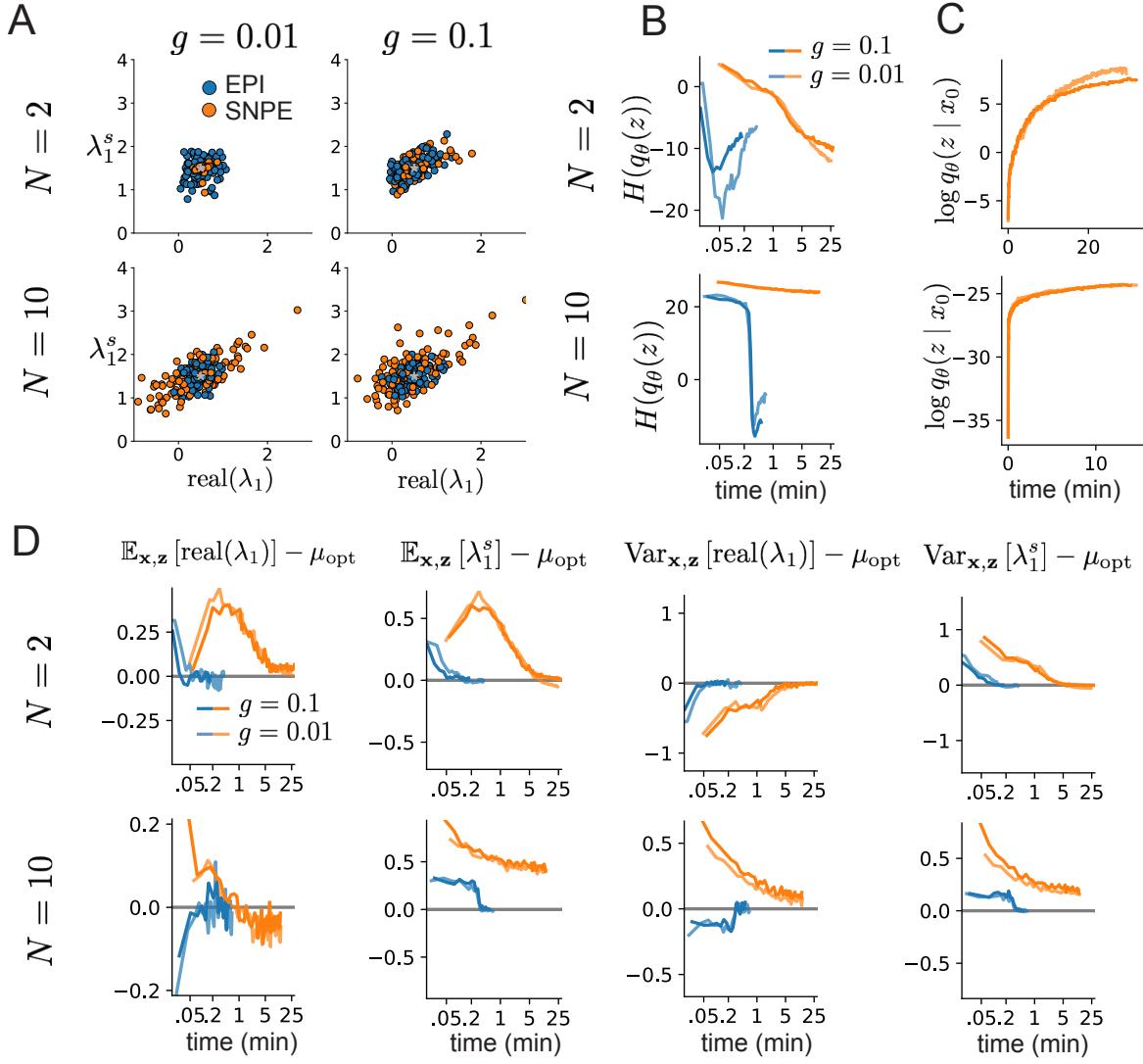


Figure 18: (RNN1): Model characteristics affect predictions of posteriors inferred by SNPE, while predictions of parameters inferred by EPI remain fixed. **A.** Predictive distribution of EPI (blue) and SNPE (orange) inferred connectivity of RNNs exhibiting stable amplification with $N = 2$ (top), $N = 10$ (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.

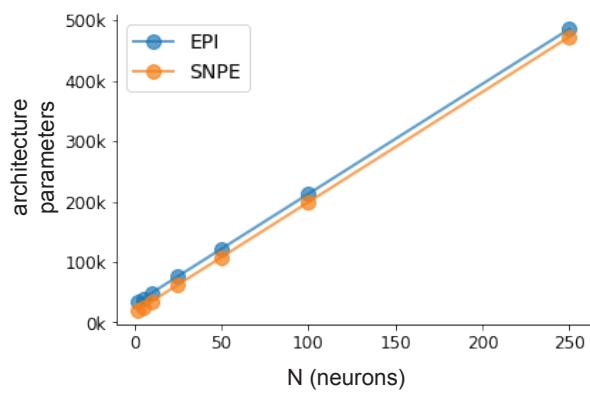


Figure 19: (RNN2): Number of parameters in deep probability distribution architectures of EPI (blue) and SNPE (orange) by RNN size (N).

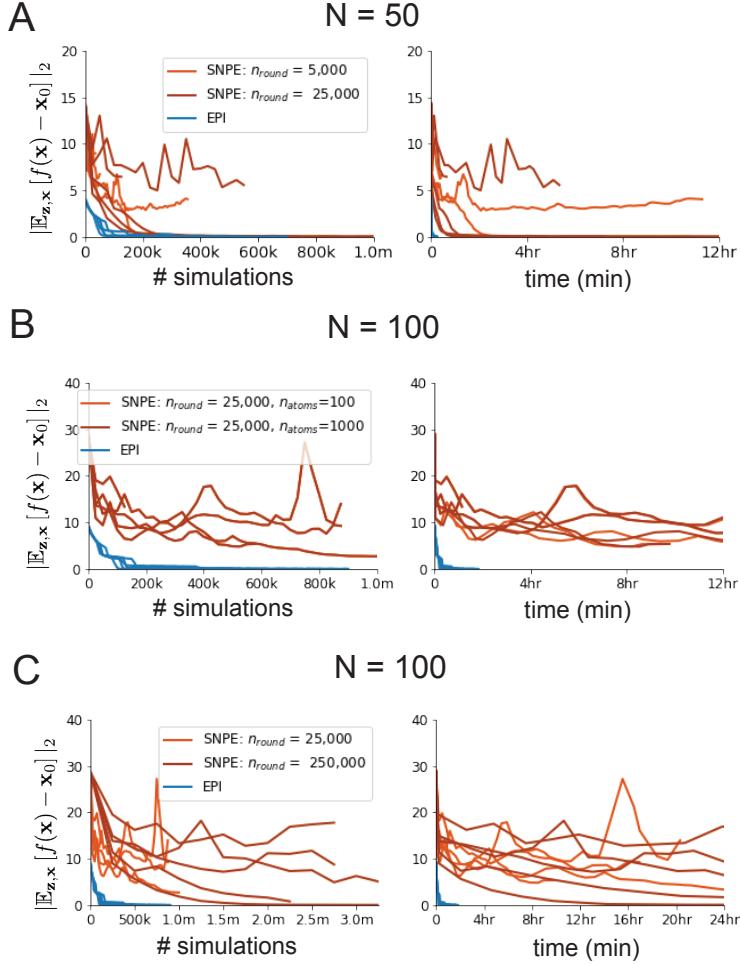


Figure 20: (RNN3): SNPE convergence was enabled by increasing n_{round} , not n_{atom} . **A.** Difference of mean predictions \mathbf{x}_0 throughout optimization at $N = 50$ with by simulation count (left) and wall time (right) of SNPE with $n_{\text{round}} = 5,000$ (light orange), SNPE with $n_{\text{round}} = 25,000$ (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).