

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 in terms of neural activity – and thus can offer
⁵ insights into neural computation. The operation of these mechanistic circuits, like all models, crit-
⁶ ically depends on the choices of model parameters. We present a novel technique designed to infer
⁷ parameter distributions that produce mathematically defined emergent properties of computation.
⁸ We detail how the parameter distributions of theoretical neuroscience problems are better defined
⁹ by the emergent properties they produce, rather than conditioned upon artificial datasets. This
¹⁰ approach is not only well-motivated, but allows technical simplifications to statistical inference in
¹¹ mechanistic models and scaling to high dimensions. With this method, we bring deep inference
¹² to bear on important questions in theoretical neuroscience, and demonstrate the broad range of
¹³ insightful analyses this approach allows. First, we show a worked example analyzing parametric
¹⁴ sensitivity in the stomatogastric ganglion, emphasizing the general applicability of this approach
¹⁵ to complex models with complex emergent properties. Then, in a model of primary visual cortex
¹⁶ where mathematical analysis becomes untenable, we identify a nonlinear manifold revealing the rule
¹⁷ of parametric degeneracy with respect to variability. Next, we gain a mechanistic understanding of
¹⁸ how superior colliculus connectivity produces rapid task switching in multiple parametric regimes.
¹⁹ Finally, we scale inference of RNN connectivity exhibiting stable amplification to 1,000-dimensional
²⁰ parameter spaces. These analyses illustrate how we can further leverage the power of deep learning

21 towards solving inverse problems in theoretical neuroscience.

22 2 Introduction

23 The fundamental practice of theoretical neuroscience is to use a mathematical model to understand
24 neural computation, whether that computation enables perception, action, or some intermediate
25 processing. A neural circuit is systematized with a set of equations – the model – and these
26 equations are motivated by biophysics, neurophysiology, and other conceptual considerations [1,
27 2, 3, 4]. The function of this system is governed by the choice of model *parameters*, which when
28 configured in a particular way, give rise to a measurable signature of a computation. The work
29 of analyzing a model then requires solving the inverse problem: given a computation of interest,
30 how can we reason about particular parameter configurations? The inverse problem is crucial for
31 reasoning about likely parameter values, uniquenesses and degeneracies, and predictions made by
32 the model [5, 6].

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

42 One preferred formalism for parameter identification is statistical inference, which has been used
43 to great success in neuroscience through the stipulation of statistical generative models [15, 16, 17,
44 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29] (see review, [30]). Recent work has used variational
45 autoencoders (VAEs) [33, 34] to interrogate hidden states in models of both cortical population
46 activity [35, 36, 37, 38] and animal behavior [39, 40, 41], thus expanding the domain of neural
47 data sets amenable to statistical modeling. However, most neural circuit models in theoretical
48 neuroscience are noisy systems of differential equations that can only be sampled or realized through
49 forward simulation; they lack the explicit likelihood necessary for statistical inference. Therefore,
50 the most popular approaches to theoretical inverse problems have been likelihood-free inference

51 (LFI) methods [31, 32], in which reasonable parameters are obtained via simulation and rejection.
52 A flourishing new class of techniques [75, 89, 90] use deep learning to improve upon traditional LFI
53 approaches. However, as we detail, all of these approaches require good datasets for the scientific
54 question at hand.

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

62 These challenges motivate the development of a novel inference framework called emergent property
63 inference (EPI). As an adaption of variational inference [46], EPI infers parameter distributions
64 that produce an emergent property: not a singular dataset, but a collection of datasets exhibiting
65 some mathematical criteria. EPI fixates on the inverse problem by constraining the predictions
66 of the inferred parameter distribution to produce the emergent property exactly. Conditioning on
67 an emergent property requires a variant of deep probabilistic inference methods, which we have
68 previously introduced [52]. By using deep probability distributions EPI flexibly captures para-
69 metric structure in mechanistic models. This technique is designed to identify the full parameter
70 space producing an emergent property and facilitate the seamless structural analysis of the inferred
71 parameter distribution. Thus, EPI provides neuroscientists with an efficient, versatile probabilistic
72 modeling toolkit designed explicitly for theoretical inverse problems.

73 Equipped with this method, we bring deep inference to bear on theoretical neuroscience to an
74 unprecedented extent. Throughout this work, we showcase the capabilities of EPI on four neural
75 circuit models across ranges of biological realism, neural system function, and network scale. First,
76 we show EPI’s ability to capture subtle, nonlinear parametric structure in a stomatogastric ganglion
77 subcircuit model [57]. In a model of primary visual cortex [72], we show how to gain insight from
78 multiple inferred distributions. Next, we used EPI to idenitfy and structurally characterize multiple
79 parametric regimes of superior colliculus activity in a model of task switching [55]. Finally, we
80 emphasize the superior scalability of EPI compared to other LFI techniques by inferring high-
81 dimensional distributions of RNN connectivities that exhibit amplified, yet stable responses – a
82 hallmark of cortical sensory systems [?, 74].

83 Most importantly of all in this work, we present novel theories of neural computation borne from EPI
84 analysis. We identified an unknown parametric rule of variability with respect to inhibitory neuron
85 type in a V1 model, where analytic techniques became untenable. Furthermore, we identified
86 multiple regimes of SC connectivity which confer rapid task switching, and used the structural
87 analytic tools of EPI to gain a mechanistic understanding of circuit responses. These valuable
88 theoretical insights illustrate the value of deep inference for the interrogation of neural circuit
89 models.

90 3 Results

91 3.1 Motivating emergent property inference of theoretical models

92 Consideration of the typical workflow of theoretical modeling clarifies the need for emergent prop-
93 erty inference. First, one designs or chooses an existing model that, it is hypothesized, captures
94 the computation of interest. To ground this process in a well-known example, consider the stom-
95 atogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic
96 muscle activation patterns for digestion [56]. Despite full knowledge of STG connectivity and a
97 precise characterization of its rhythmic pattern generation, biophysical models of the STG have
98 complicated relationships between circuit parameters and neural activity [53, 12]. A subcircuit
99 model of the STG [57] is shown schematically in Figure 1A, and note that the behavior of this
100 model will be critically dependent on its parameterization – the choices of conductance parameters
101 $\mathbf{z} = [g_{el}, g_{synA}]$. Specifically, the two fast neurons (f_1 and f_2) mutually inhibit one another, and
102 oscillate at a faster frequency than the mutually inhibiting slow neurons (s_1 and s_2). The hub
103 neuron (hub) couples with either the fast or slow population or both.

104 Second, once the model is selected, one defines the emergent phenomena of scientific interest. In the
105 STG example, we are concerned with neural spiking frequency, which emerges from the dynamics
106 of the circuit model 1B. An interesting emergent property of this stochastic model is when the hub
107 neuron fires at an intermediate frequency between the intrinsic spiking rates of the fast and slow
108 populations. This emergent property is shown in Figure 1C at an average frequency of 0.55Hz.

109 Third, parameter analyses ensue: brute-force parameter sweeps, ABC sampling [32], and sensitivity
110 analyses [96] are all routinely used to reason about what parameter configurations lead to an
111 emergent property. In this last step lies the opportunity for a precise quantification of the emergent
112 property as a statistical feature of the model. Once we have such a methodology, we can infer a

113 probability distribution over parameter configurations that produce this emergent property.
 114 Before presenting technical details (in the following section), let us understand emergent property
 115 inference schematically: EPI (Fig. 1D) takes, as input, the model and the specified emergent
 116 property, and as its output, produces the parameter distribution EPI (Fig. 1E). This distribution
 117 – represented for clarity as samples from the distribution – is then the most random parameter
 118 distribution producing the emergent property. In the STG model, this distribution can be specif-
 119 ically queried to reveal the prototypical parameter configuration for intermediate hub frequency
 120 (the mode; Figure 1E yellow star), and how it decays based on changes away from the mode. In-
 121 deed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity (v_1)
 122 and degeneracy (v_2) (Fig. 1E, arrows) agree with error contours (Fig. 1E contours) and have
 123 diminished or preserved hub frequency, respectively (Fig. 1F activity traces) (see Section 5.2.1).

124 3.2 A deep generative modeling approach to emergent property inference

125 Emergent property inference (EPI) formalizes the three-step procedure of the previous section with
 126 deep probability distributions. First, we consider the model as a coupled set of differential equations
 127 [57]. In the running STG example, the model activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$ is the membrane
 128 potential for each neuron, which evolves according to the biophysical conductance-based equation:

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

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

132 Second, we define the emergent property, which as above is “intermediate hub frequency” (Figure
 133 1C). Quantifying this phenomenon is straightforward: we stipulate that the hub neuron’s spiking
 134 frequency – denoted $\omega_{hub}(\mathbf{x})$ is close to a frequency of 0.55Hz. Mathematically, we achieve this
 135 with two constraints: by setting the mean hub frequency of model generated activity to 0.55Hz.

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

136 and requiring that the variance around this mean is small

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

137 This level of variance was chosen to be low enough to exclude the fast and slow frequencies of the
 138 two populations, but large enough to allow structural examination of the compatible parameter

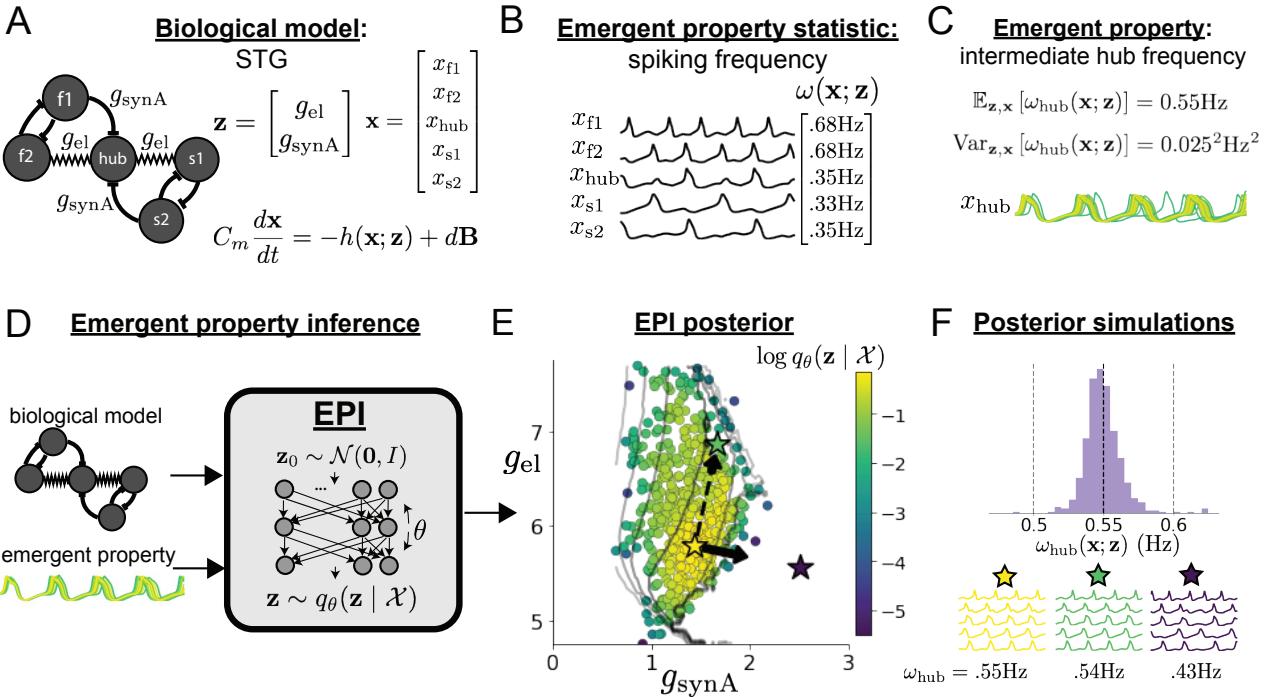


Figure 1: Emergent property inference (EPI) in the stomatogastric ganglion. **A.** Conductance-based biophysical model of the STG subcircuit. In the STG model, jagged connections indicate electrical coupling having electrical conductance g_{el} . Other connections in the diagram are inhibitory synaptic projections having strength g_{synA} onto the hub neuron, and $g_{synB} = 5\text{nS}$ 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 $g_{el} = 4.5\text{nS}$ and $g_{synA} = 3\text{nS}$. **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.

139 space. In general, an emergent property

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

140 defines a collection of datasets with a statistic $f(\mathbf{x}; \mathbf{z})$ (which may be comprised of multiple statistics) and the means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ of those statistics over the datasets. The choice of $\boldsymbol{\sigma}^2$ 141 142 predicates the degree of variability in around mean $\boldsymbol{\mu}$ that is consistent with the emergent property. 143

144 Third, we perform emergent property inference: we find a distribution over parameter configurations \mathbf{z} , and insist that samples from this distribution produce the emergent property; in other 145 words, they obey the constraints introduced in Equation 4. This distribution will be chosen from a 146 family of probability distributions $\mathcal{Q} = \{q_{\boldsymbol{\theta}}(\mathbf{z}) : \boldsymbol{\theta} \in \Theta\}$, defined by a deep generative distribution 147 148 of the normalizing flow class [47, 50, 51] – neural networks which transform a simple distribution 149 into a suitably complicated distribution (as is needed here). This deep distribution is represented 150 in Figure 1D (see Section 5.1). Many distributions in \mathcal{Q} will respect the emergent property 151 constraints, so we choose a normative selection principle imposing no additional structure beyond the 152 production of the emergent property [58, 59, 52, 60], which is the same normative principle of 153 Bayesian inference (see Section 5.1.6).

154 The probabilities of the distribution inferred from EPI are the densities of these parameters in the 155 most random distribution producing the emergent property. While existing approaches to proba- 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1062 1063 1064 1065 1066 1067 1068 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1079 1080 1081 1082 1083 1084 1085 1086 1087 1088 1089 1090 1091 1092 1093 1094 1095 1096 1097 1098 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1110 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155 1156 1157 1158 1159 1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185 1186 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1200 1201 1202 1203 1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1272 1273 1274 1275 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299

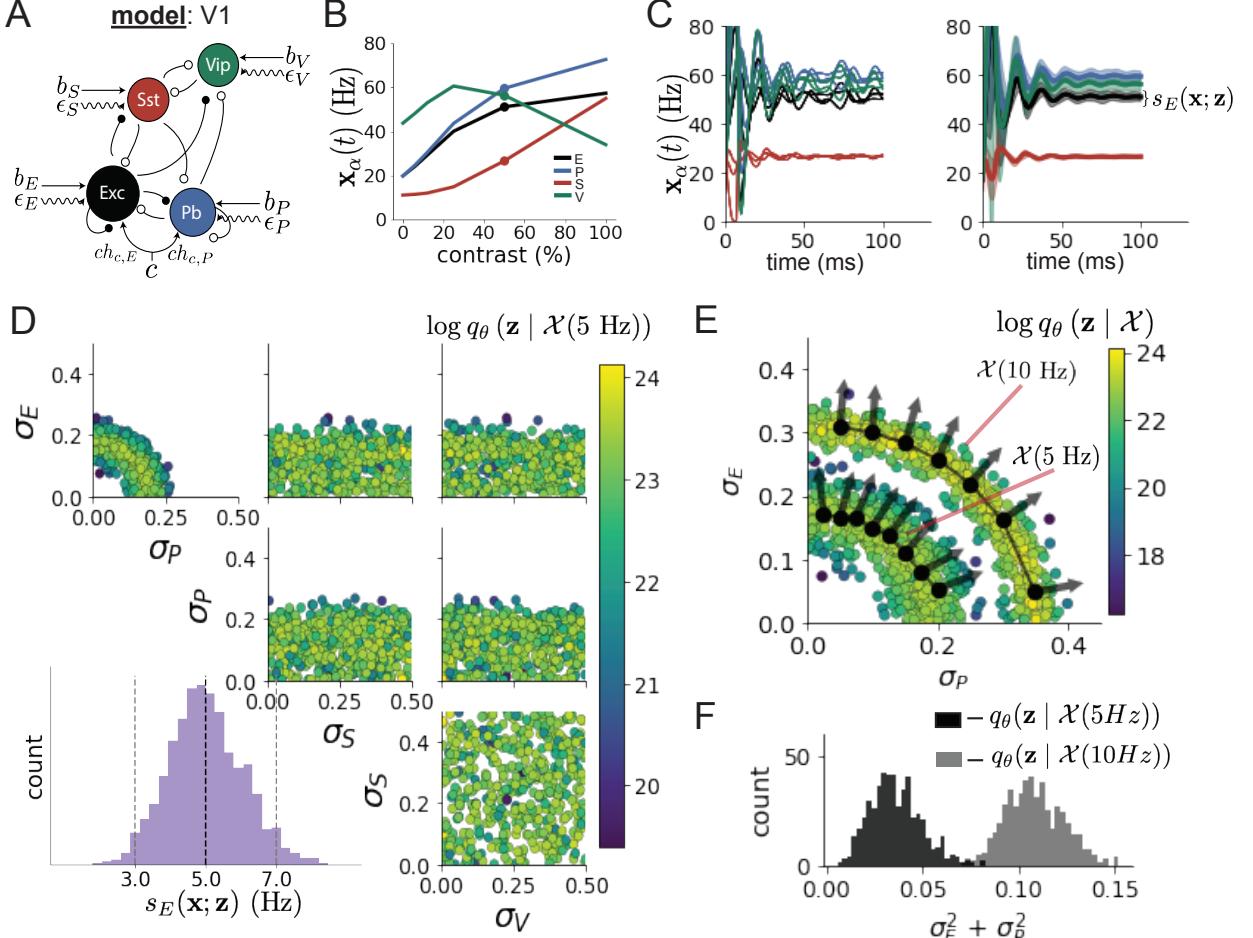


Figure 2: Emergent property inference in the stochastic stabilized supralinear network (SSSN) **A.** Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and VIP (green) neurons (excitatory and inhibitory projections filled and unfilled, respectively).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

166 **3.3 EPI reveals how neuron-type specific noise governs variability in the stochastic**
167 **stabilized supralinear network**

168 Dynamical models of excitatory (E) and inhibitory (I) populations with supralinear input-output
169 function have succeeded in explaining a host of experimentally documented phenomena. In a
170 regime characterized by inhibitory stabilization of strong recurrent excitation, these models give
171 rise to paradoxical responses [9], selective amplification [61, 62], surround suppression [63] and
172 normalization [64]. Despite their strong predictive power, E-I circuit models rely on the assumption
173 that inhibition can be studied as an indivisible unit. However, experimental evidence shows
174 that inhibition is composed of distinct elements – parvalbumin (P), somatostatin (S), VIP (V) –
175 composing 80% of GABAergic interneurons in V1 [65, 66, 67], and that these inhibitory cell types
176 follow specific connectivity patterns (Fig. 2A) [68]. Recent theoretical advances [54, 69, 70], have
177 only started to address the consequences of this multiplicity in the dynamics of V1, strongly relying
178 on linear theoretical tools. Here, we use EPI to analyze V1 models of greater complexity in order
179 to characterize properties of slow noise governing circuit variability.

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

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

186 This model is the stochastic stabilized supralinear network (SSSN) [71] generalized to have inhibitory
187 multiplicity, and introduces stochasticity to previous four neuron-type models of V1 [54]. Stochasticity and inhibitory multiplicity introduce substantial complexity to mathematical derivations
188 (see Section 5.2.3) motivating the treatment of this model with EPI. Here, we consider fixed
189 weights W and input \mathbf{h} according to a fit of the deterministic model to contrast responses [72] (Fig.
190 2B), and study the effect of noise parameterization $\mathbf{z} = [\sigma_E, \sigma_P, \sigma_S, \sigma_V]^\top$ on fluctuations at 50%
191 contrast.

193 For this SSSN, we are interested in how noise variability across neural populations governs stochastic
194 fluctuations in the E-population. Here, we quantify different levels y of E-population variability

195 with the emergent property

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

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

198 We ran EPI to obtain a posterior distribution $q_{\theta}(\mathbf{z} | \mathcal{X}(5 \text{ Hz})$ producing E-population variability
199 around 5 Hz (Fig. 2D). From the marginal distribution of σ_E and σ_P (Fig. 2D, top-left), we can
200 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
201 are degenerate with respect to $s_E(\mathbf{x}; \mathbf{z})$ evidenced by the high variability in those dimensions of the
202 posterior (Fig. 2D, bottom-right). Together, these observations imply a parametric manifold of
203 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
204 marginal (Fig. 2E). The dimensions of sensitivity conferred by EPI and this plain visual structure
205 suggest a quadratic relationship in the emergent property statistic $s_E(\mathbf{x}; \mathbf{z})$ and parameters \mathbf{z} , which
206 is preserved at a greater level of variability $\mathcal{X}(10 \text{ Hz})$ (Fig. 2E). Indeed, the sum of squares of σ_E
207 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
208 of σ_S and σ_V are not significantly different in the two posteriors (Fig. 11, $p = .402$).

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

217 3.4 EPI identifies multiple regimes of rapid task switching

218 In a rapid task switching experiment [73], rats were explicitly cued on each trial to either orient
219 towards a visual stimulus in the Pro (P) task or orient away from a visual stimulus in the Anti
220 (A) task (Fig. 3A). Neural recordings in the midbrain superior colliculus (SC) exhibited two
221 populations of neurons that simultaneously represented both task context (Pro or Anti) and motor
222 response (contralateral or ipsilateral to the recorded side): the Pro/Contra and Anti/Ipsi neurons

[55]. Duan et al. proposed a model of SC that, like the V1 model analyzed in the previous section, is a four-population dynamical system. We analyzed this model, where the neuron-type populations are functionally-defined as the Pro- and Anti-populations in each hemisphere (left (L) and right (R)), their connectivity is parameterized geometrically (Fig. 3B). The input-output function of this model is chosen such that the population responses $\mathbf{x} = [x_{LP}, x_{LA}, x_{RP}, x_{RA}]^\top$ are bounded from 0 to 1 as a function ϕ of a dynamically evolving internal variable \mathbf{u} . The model responds to the side with greater Pro neuron activation; e.g. the response is left if $x_{LP} > x_{RP}$ at the end of the trial. The dynamics evolve with timescale $\tau = 90\text{ms}$ governed by connectivity weights W

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

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

We define rapid task switching behavior as accurate execution of each task. Inferred models should not exhibit fully random responses (50%), or perfect performance (100%), since perfection is never attained by even the best trained rats. We formulate rapid task switching as an emergent property 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% with variance $7.5\%^2$.

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

A variance of $7.5\%^2$ in each task will confer a posterior producing performances ranging from about 60% – 90%, allowing us to examine the properties of connectivity that yield better performance in each task. Notably, this is our first example using EPI to condition on multiple emergent property statistics ($|f(\mathbf{x}; \mathbf{z})| = 2$).

The EPI inferred parameters (Fig. 3C) generate a distribution of task accuracies (Fig. 3C, middle-left) according to our mathematical definition of rapid task switching (Equation 8). The nonlinear patterns of connectivity that govern each task accuracy (Fig. 12A-B) are not fully captured by linear prediction (Fig. 12C). For example, the patterns in connectivity increasing Pro accuracy

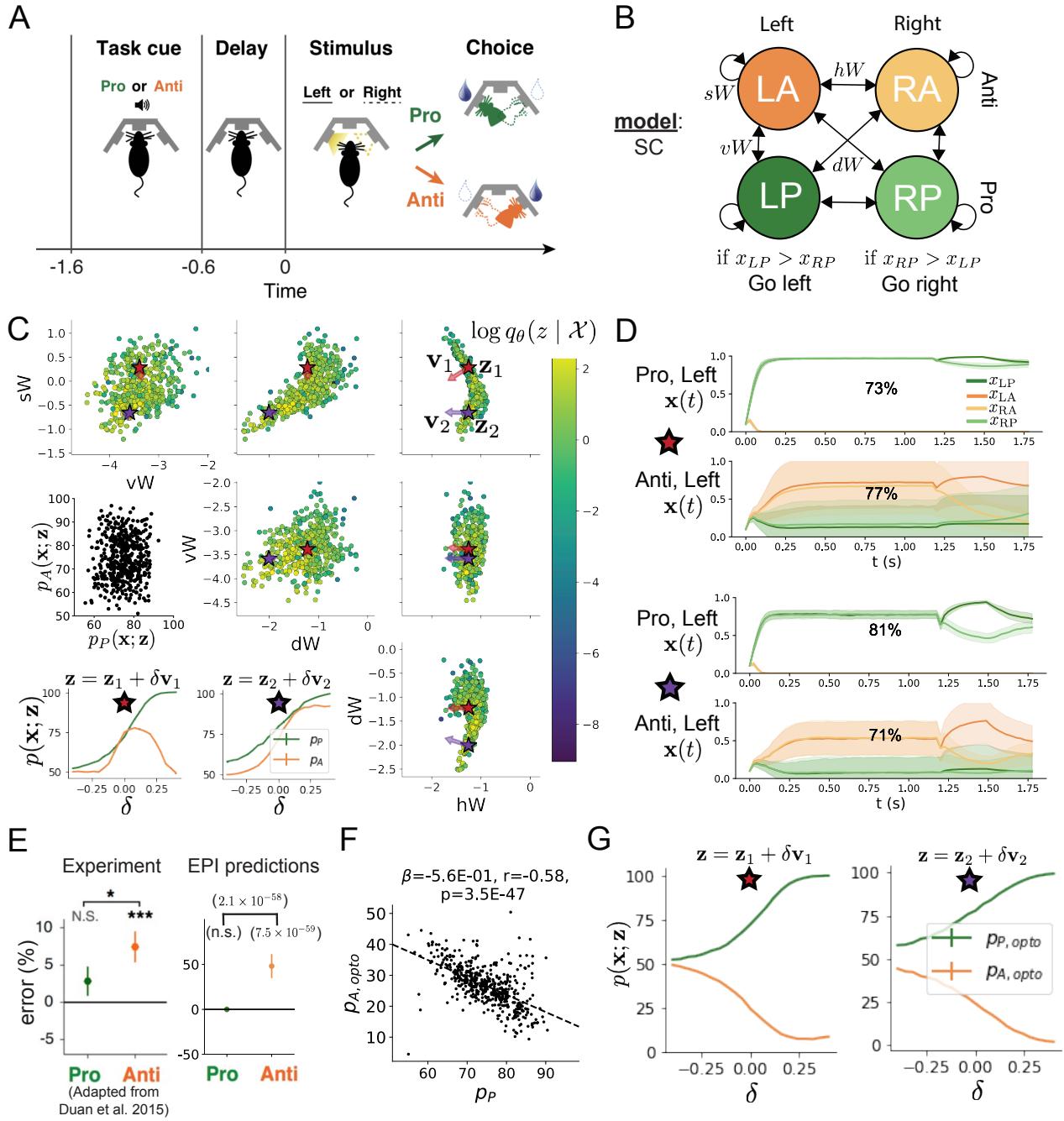


Figure 3: **A.** Rapid task switching behavioral paradigm (see text). **B.** Model of superior colliculus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. **C.** The EPI posterior distribution 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.

248 change dramatically after crossing a threshold of sW (Fig. 12A $sW-hW$ marginal). Not only has
249 EPI captured this complex nonlinear posterior, it offers probabilistic tools for understanding the
250 different regimes of model behavior.

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

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

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

²⁷⁹ **3.5 EPI scales well to high-dimensional parameter spaces**

²⁸⁰ Here, we study the scalability of EPI in number of parameters $|\mathbf{z}|$ by inferring the connectivities
²⁸¹ of recurrent neural networks (RNNs, Fig. 4A). We consider a rank-2 RNN with N neurons having
²⁸² connectivity

$$W = UV^\top + g\chi \quad (9)$$

²⁸³ and dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W\mathbf{x} \quad (10)$$

²⁸⁴ where $U = [\mathbf{u}_1 \ \mathbf{u}_2]$, $V = [\mathbf{v}_1 \ \mathbf{v}_2]$ and $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2 \in [-1, 1]^N$. The random component has
²⁸⁵ strength $g = 0.01$ and $\chi_{i,j} \sim \mathcal{N}(0, 1)$. We infer connectivity distributions $\mathbf{z} = [\mathbf{u}_1^\top, \mathbf{u}_2^\top, \mathbf{v}_1^\top, \mathbf{v}_2^\top]^\top$
²⁸⁶ producing stable amplification. RNN's exhibiting stable amplification amplify responses to input
²⁸⁷ along some dimensions, and are stable across all dimensions. Two conditions are both necessary
²⁸⁸ and sufficient for RNNs to exhibit stable amplification [74]: $\text{real}(\lambda_1) < 1$ and $\lambda_1^s > 1$, where λ_1 is
²⁸⁹ the eigenvalue of W with greatest real part and λ_1^s is the maximum eigenvalue of $W^s = \frac{W+W^\top}{2}$.

²⁹⁰ In our analysis, we seek to condition rank-2 networks of increasing size on a regime of stable ampli-
²⁹¹ fication. Networks with $\text{real}(\lambda_1) = 0.5 \pm 0.5$ and $\lambda_1^s = 1.5 \pm 0.5$ will yield moderate amplification.
²⁹² EPI can naturally condition on this emergent property

$$\begin{aligned} \mathcal{X} : \mathbb{E}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \\ \text{Var}_{\mathbf{z}, \mathbf{x}} \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} &= \begin{bmatrix} 0.25^2 \\ 0.25^2 \end{bmatrix}. \end{aligned} \quad (11)$$

²⁹³ For comparison, we infer rank-2 RNN connectivities with alternative approaches to likelihood free-
²⁹⁴ inference. ABC methods define a tolerance ϵ from observed data x_0 for which we keep sampled
²⁹⁵ parameters. To make this ABC approach as similar as possible to the EPI program defined by
²⁹⁶ Equation 11, we chose $\epsilon = 0.5$, an l_2 -distance metric, and

$$x_0 = \begin{bmatrix} \text{real}(\lambda_1) \\ \lambda_1^s \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \quad (12)$$

²⁹⁷ located at the mean of our desired emergent property. We use sequential Monte Carlo ABC (SMC-
²⁹⁸ ABC), to compare efficiency, since it is considered the most efficient ABC approach. SNPE [75] is
²⁹⁹ another deep likelihood-free inference method that emerged along with this work. In contrast to
³⁰⁰ EPI, SNPE cannot condition on the variance of the posterior predictive distribution. Also, there

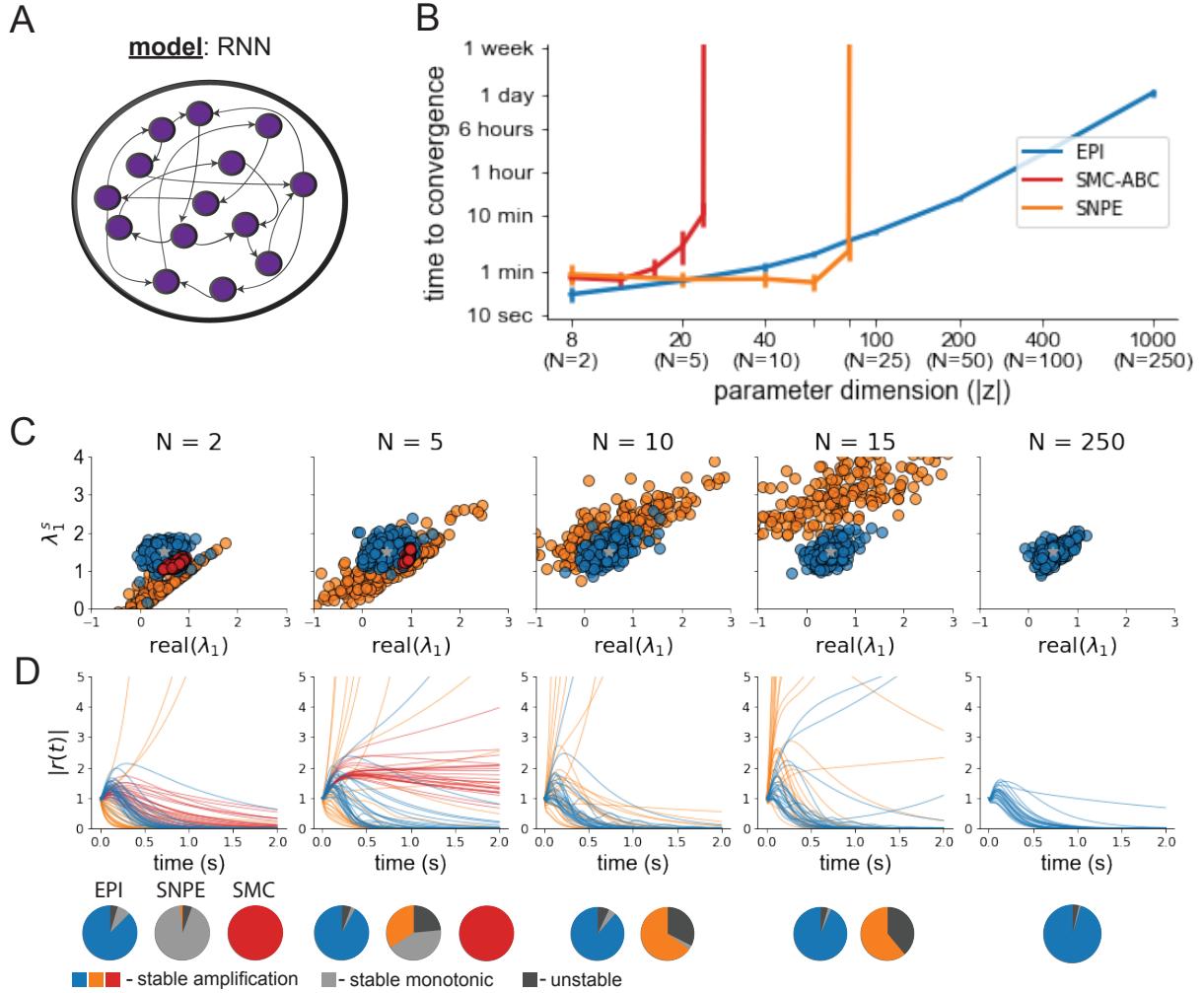


Figure 4: **A.** Recurrent neural network. **B.** EPI scales with z to high dimensions. Convergence definitions: EPI (blue) - satisfies all moment constraints, SNPE (orange)- produces at least $2/n_{\text{train}}$ parameter samples are in the bounds of emergent property (mean ± 0.5), and SMC-ABC (red) - 100 particles with $\epsilon < 0.5$ are produced. **C.** Posterior predictive distributions of EPI (blue), SNPE (orange), and SMC-ABC (red). Gray star indicates emergent property mean, and gray dashed lines indicate two standard deviations corresponding to the variance constraint. For $N \leq 6$ where SMC-ABC converges, samples are not diverse (path degeneracies). For $N \geq 25$, SNPE does not produce a posterior approximation yielding parameters with simulations near x_0 . **D.** Simulations of network parameters resulting from each method ($\tau = 100ms$). Each trace corresponds to simulation of one z . (Below) Ratio of obtained samples producing stable amplification.

301 is no tolerance parameter for SNPE like ϵ in ABC, so the comparative SNPE approach simply
302 conditions on observation x_0 .

303 As we scale the number of neurons N in the RNN, and thus the dimensionality of the parameter
304 space $\mathbf{z} \in [-1, 1]^{4N}$, we see that EPI has superior scaling properties (Fig. 4B). SMC-ABC and
305 SNPE become intractable around 25 and 90 dimensions respectively, while EPI can infer 1000-
306 dimensional distributions in about 1 day. No matter the number of neurons, EPI always produces
307 connectivity distributions with mean and variance of $\text{real}(\lambda_1)$ and λ_1^s of \mathcal{X} (Fig. 4C, blue), and high
308 variation in response profiles 4D, blue). For the dimensionalities in which SMC-ABC is tractable,
309 the inferred parameters always exhibit stable amplification, are less varied 4C, red) and largely
310 produce similar responses 4D, red). When using SNPE the inferred parameters are widely varied
311 4C, orange), but often produce non-amplified or unstable responses 4D, orange). In conclusion, we
312 found that deep likelihood-free inference techniques are capable of scaling to higher dimensional
313 inference than SMC-ABC. However, only EPI can scale to high dimensions while reproducing the
314 emergent property.

315 4 Discussion

316 In neuroscience, machine learning has primarily been used to reveal structure in neural datasets
317 [30]. Such careful inference procedures are developed for these statistical models allowing precise,
318 quantitative reasoning, which clarifies the way data informs beliefs about the model parameters.
319 However, these statistical models lack resemblance to the underlying biology, making it unclear
320 how to go from the structure revealed by these methods, to the neural mechanisms giving rise
321 to it. In contrast, theoretical neuroscience has focused on careful mechanistic modeling and the
322 production of emergent properties of computation. The careful steps of *i.*) model design and
323 *ii.*) emergent property definition, are followed by *iii.*) practical inference methods resulting in an
324 opaque characterization of the way model parameters govern computation. In this work, we improve
325 upon parameter inference techniques in theoretical neuroscience with emergent property inference,
326 harnessing deep learning towards careful inference in careful models of neural computation (see
327 Section 5.1.1).

328 Specifically, approximate Bayesian computation [42, 43, 31] has been the standard approach to
329 parameter inference in neural circuit models lacking tractable likelihoods. ABC methods do not
330 confer probabilities on accepted parameters, require an acceptance threshold chosen to trade-off

331 inference quality with tractability, do not scale efficiently to high-dimensional parameter spaces, and
332 require independent techniques to analyze sensitivity for local parameter choices [76]. In contrast,
333 EPI allows probability evaluations at any point in parameter space, conditions posteriors on the
334 natural quantification of emergent properties, scales to high dimensional parameter spaces, and
335 naturally admits sensitivity quantification via fast evaluations of the posterior Hessian.

336 Technically, EPI is a maximum entropy method, which learns parameter distributions that are
337 as random as possible given that they produce the emergent property. Conceptually, maximally
338 random distributions given some constraints are useful for understanding parametric sensitivity.
339 This is well understood in Bayesian inference, where maximum entropy is the chosen normative
340 principle. This is emphasized by an innovative formalism unifying top-down maximum entropy
341 normative models with bottom-up statistical models [77]. Indeed, EPI is an adaptive variational
342 inference program, and may be considered to have a Bayesian uniform prior (see Section 5.1.6).

343 Biologically realistic models of neural circuits often prove formidable to analyze for two main rea-
344 sons. A primary challenge is that the number of parameters scales dramatically with the number of
345 neurons, limiting analysis of its parameter space. We see in Section 3.5 that EPI scales seemlessly
346 to high dimensional parameter spaces of RNN connectivities, while maintaining the production
347 of the specified emergent property. EPI strongly outperforms the standard likelihood-free infer-
348 ence technique (SMC-ABC [31]), and a recently developed deep likelihood-free inference technique
349 (SNPE [75]), most likely because of it's ability to leverage the gradient information of the emer-
350 gent property statistics and to adapt it's parameter sampling distribution at every step of gradient
351 descent.

352 A secondary challenge is that the structure of the parametric regimes governing emergent properties
353 is intricate. For example, even in low dimensional circuits, models can support more than one steady
354 state [78] and non-trivial dynamics on strange attractors [79]. With EPI, we use deep probabillity
355 distributions to capture the complex nonlinear parameter distributions governing model behavior.
356 In Section 3.3, we used EPI to reveal a curved parametric manifolds governing circuit variability
357 in the stochastic stabilized supralinear network, and used hypothesis testing techniques to validate
358 our findings. In Section 3.4, we identified two regimes of SC connectivity resulting in rapid task
359 switching, and found that the full distribution of rapid task switching networks reproduced an
360 experimental result.

361 EPI leverages deep learning technology for neuroscientific inquiry in a categorically different way
362 than approaches focused on training neural networks to execute behavioral tasks [80]. These works

363 focus on examining optimized deep neural networks while considering the objective function, learn-
364 ing rule, and architecture used. This endeavor efficiently obtains sets of parameters that can be
365 reasoned about with respect to such considerations, but lacks the careful probabilistic treatment of
366 parameter inference in EPI. All of these approaches can be used complementarily to enhance the
367 practice of theoretical neuroscience.

368 **Acknowledgements:**

369 This work was funded by NSF Graduate Research Fellowship, DGE-1644869, McKnight Endow-
370 ment Fund, NIH NINDS 5R01NS100066, Simons Foundation 542963, NSF NeuroNex Award, DBI-
371 1707398, The Gatsby Charitable Foundation, Simons Collaboration on the Global Brain Postdoc-
372 toral Fellowship, Chinese Postdoctoral Science Foundation, and International Exchange Program
373 Fellowship. Helpful conversations were had with Francesca Mastrogiuseppe, Srdjan Ostojic, James
374 Fitzgerald, Stephen Baccus, Dhruva Raman, Liam Paninski, and Larry Abbott.

375 **Data availability statement:**

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

378 **Code availability statement:**

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

380 **References**

- 381 [1] Nancy Kopell and G Bard Ermentrout. Coupled oscillators and the design of central pattern
382 generators. *Mathematical biosciences*, 90(1-2):87–109, 1988.
- 383 [2] Eve Marder. From biophysics to models of network function. *Annual review of neuroscience*,
384 21(1):25–45, 1998.
- 385 [3] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.
- 386 [4] Xiao-Jing Wang. Neurophysiological and computational principles of cortical rhythms in
387 cognition. *Physiological reviews*, 90(3):1195–1268, 2010.
- 388 [5] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R
389 Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology
390 models. *PLoS Comput Biol*, 3(10):e189, 2007.

- 391 [6] Timothy O’Leary, Alex H Williams, Alessio Franci, and Eve Marder. Cell types, network
392 homeostasis, and pathological compensation from a biologically plausible ion channel expres-
393 sion model. *Neuron*, 82(4):809–821, 2014.
- 394 [7] John J Hopfield. Neural networks and physical systems with emergent collective computa-
395 tional abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- 396 [8] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural
397 networks. *Physical review letters*, 61(3):259, 1988.
- 398 [9] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Para-
399 doxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*,
400 17(11):4382–4388, 1997.
- 401 [10] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration
402 in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.
- 403 [11] WR Foster, LH Ungar, and JS Schwaber. Significance of conductances in hodgkin-huxley
404 models. *Journal of neurophysiology*, 70(6):2502–2518, 1993.
- 405 [12] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit
406 parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.
- 407 [13] Pablo Achard and Erik De Schutter. Complex parameter landscape for a complex neuron
408 model. *PLoS computational biology*, 2(7):e94, 2006.
- 409 [14] Leandro M Alonso and Eve Marder. Visualization of currents in neural models with similar
410 behavior and different conductance densities. *Elife*, 8:e42722, 2019.
- 411 [15] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*,
412 13(8):1713–1720, 2001.
- 413 [16] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson.
414 A statistical paradigm for neural spike train decoding applied to position prediction from
415 ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–
416 7425, 1998.
- 417 [17] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding
418 models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.

- 419 [18] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown.
420 A point process framework for relating neural spiking activity to spiking history, neural
421 ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.
- 422 [19] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek. Weak pairwise correla-
423 tions imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–
424 1012, 2006.
- 425 [20] Shaul Druckmann, Yoav Banitt, Albert A Gidon, Felix Schürmann, Henry Markram, and Idan
426 Segev. A novel multiple objective optimization framework for constraining conductance-based
427 neuron models by experimental data. *Frontiers in neuroscience*, 1:1, 2007.
- 428 [21] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature
429 analysis. *Neural computation*, 19(4):1022–1038, 2007.
- 430 [22] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and
431 Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of
432 neural population activity. In *Advances in neural information processing systems*, pages
433 1881–1888, 2009.
- 434 [23] Jakob H Macke, Lars Buesing, John P Cunningham, Byron M Yu, Krishna V Shenoy, and
435 Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural
436 information processing systems*, 24:1350–1358, 2011.
- 437 [24] Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. In
438 *Advances in neural information processing systems*, pages 1692–1700, 2011.
- 439 [25] Einat Granot-Atedgi, Gašper Tkačik, Ronen Segev, and Elad Schneidman. Stimulus-
440 dependent maximum entropy models of neural population codes. *PLoS Comput Biol*,
441 9(3):e1002922, 2013.
- 442 [26] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W
443 Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.
444 *Science*, 349(6244):184–187, 2015.
- 445 [27] Kaushik J Lakshminarasimhan, Marina Petsalis, Hyeshin Park, Gregory C DeAngelis, Xaq
446 Pitkow, and Dora E Angelaki. A dynamic bayesian observer model reveals origins of bias in
447 visual path integration. *Neuron*, 99(1):194–206, 2018.

- 448 [28] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable
449 continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- 451 [29] Josef Ladenbauer, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic.
452 Inferring and validating mechanistic models of neural microcircuits based on spike-train data.
453 *Nature Communications*, 10(4933), 2019.
- 454 [30] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-
455 analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241,
456 2018.
- 457 [31] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods.
458 *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- 459 [32] Juliane Liepe, Paul Kirk, Sarah Filippi, Tina Toni, Chris P Barnes, and Michael PH Stumpf.
460 A framework for parameter estimation and model selection from experimental data in systems
461 biology using approximate bayesian computation. *Nature protocols*, 9(2):439–456, 2014.
- 462 [33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.
- 464 [34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation
465 and variational inference in deep latent gaussian models. *International Conference on Machine Learning*, 2014.
- 467 [35] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical
468 neural population models through nonlinear embeddings. In *Advances in neural information processing systems*, pages 163–171, 2016.
- 470 [36] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for non-
471 linear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.
- 472 [37] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-
473 encoders. *bioRxiv*, page 399246, 2018.
- 474 [38] Chethan Pandarinath, Daniel J O’Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky,
475 Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R

- 476 Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-
477 encoders. *Nature methods*, page 1, 2018.
- 478 [39] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M
479 Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta.
480 Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.
- 481 [40] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R
482 Datta. Composing graphical models with neural networks for structured representations and
483 fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- 484 [41] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall,
485 Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet:
486 nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural
487 Information Processing Systems*, 2019.
- 488 [42] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computa-
489 tion in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- 490 [43] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo
491 without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328,
492 2003.
- 493 [44] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications.
494 1970.
- 495 [45] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and
496 Edward Teller. Equation of state calculations by fast computing machines. *The journal of
497 chemical physics*, 21(6):1087–1092, 1953.
- 498 [46] Lawrence Saul and Michael Jordan. A mean field learning algorithm for unsupervised neural
499 networks. In *Learning in graphical models*, pages 541–554. Springer, 1998.
- 500 [47] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows.
501 *International Conference on Machine Learning*, 2015.
- 502 [48] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R
503 Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics,
504 biology, and beyond. *The Journal of chemical physics*, 143(1):07B201_1, 2015.

- 505 [49] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-
506 free variational inference. In *Advances in Neural Information Processing Systems*, pages
507 5523–5533, 2017.
- 508 [50] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp.
509 *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- 510 [51] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for
511 density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347,
512 2017.
- 513 [52] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow
514 networks. *International Conference on Learning Representations*, 2017.
- 515 [53] Mark S Goldman, Jorge Golowasch, Eve Marder, and LF Abbott. Global structure, ro-
516 bustness, and modulation of neuronal models. *Journal of Neuroscience*, 21(14):5229–5238,
517 2001.
- 518 [54] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and
519 visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysi-
520 ology*, 115(3):1399–1409, 2016.
- 521 [55] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J
522 Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor
523 routing. *bioRxiv*, page 245613, 2018.
- 524 [56] Eve Marder and Vatsala Thirumalai. Cellular, synaptic and network effects of neuromodula-
525 tion. *Neural Networks*, 15(4-6):479–493, 2002.
- 526 [57] Gabrielle J Gutierrez, Timothy O’Leary, and Eve Marder. Multiple mechanisms switch an
527 electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators.
528 *Neuron*, 77(5):845–858, 2013.
- 529 [58] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
530 1957.
- 531 [59] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings:
532 an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.

- 533 [60] Cristina Savin and Gašper Tkačik. Maximum entropy models as a tool for building precise
534 neural controls. *Current opinion in neurobiology*, 46:120–126, 2017.
- 535 [61] Mark S Goldman. Memory without feedback in a neural network. *Neuron*, 61(4):621–634,
536 2009.
- 537 [62] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of
538 selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.
- 539 [63] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory
540 stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–
541 592, 2009.
- 542 [64] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
543 network: a unifying circuit motif underlying multi-input integration in sensory cortex.
544 *Neuron*, 85(2):402–417, 2015.
- 545 [65] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and
546 Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*,
547 5(10):793, 2004.
- 548 [66] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of
549 interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neu-*
550 *robiology*, 71(1):45–61, 2011.
- 551 [67] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocor-
552 tex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
- 553 [68] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhi-
554 bition of inhibition in visual cortex: the logic of connections between molecularly distinct
555 interneurons. *Nature Neuroscience*, 16(8):1068, 2013.
- 556 [69] Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing
557 Wang. Paradoxical response reversal of top- down modulation in cortical circuits with three
558 interneuron types. *Elife*, 6:1–15, 2017.
- 559 [70] Guang Chen, Carl Van Vreeswijk, David Hansel, and David Hansel. Mechanisms underlying
560 the response of mouse cortical networks to optogenetic manipulation. 2019.

- 561 [71] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D
562 Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-
563 tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
- 564 [72] Agostina Palmigiano, Francesco Fumarola, Daniel P Mossing, Nataliya Kraynyukova, Hillel
565 Adesnik, and Kenneth Miller. Structure and variability of optogenetic responses identify the
566 operating regime of cortex. *bioRxiv*, 2020.
- 567 [73] Chunyu A Duan, Jeffrey C Erlich, and Carlos D Brody. Requirement of prefrontal and
568 midbrain regions for rapid executive control of behavior in the rat. *Neuron*, 86(6):1491–1503,
569 2015.
- 570 [74] Giulio Bondanelli and Srdjan Ostojic. Coding with transient trajectories in recurrent neural
571 networks. *PLoS computational biology*, 16(2):e1007655, 2020.
- 572 [75] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan
573 Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P
574 Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural
575 dynamics. *bioRxiv*, page 838383, 2019.
- 576 [76] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian com-*
577 *putation*. CRC Press, 2018.
- 578 [77] Wiktor Mlynarski, Michal Hledík, Thomas R Sokolowski, and Gašper Tkačik. Statistical
579 analysis and optimality of neural systems. *bioRxiv*, page 848374, 2020.
- 580 [78] Nataliya Kraynyukova and Tatjana Tchumatchenko. Stabilized supralinear network can give
581 rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of*
582 *Sciences*, 115(13):3464–3469, 2018.
- 583 [79] Katherine Morrison, Anda Degeratu, Vladimir Itskov, and Carina Curto. Diversity of emergent
584 dynamics in competitive threshold-linear networks: a preliminary report. *arXiv preprint arXiv:1605.04463*, 2016.
- 585 [80] Blake A Richards and et al. A deep learning framework for neuroscience. *Nature Neuroscience*,
586 2019.

- 588 [81] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte
589 carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
590 73(2):123–214, 2011.
- 591 [82] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic bio-
592 chemical network models using particle markov chain monte carlo. *Interface focus*, 1(6):807–
593 820, 2011.
- 594 [83] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based infer-
595 ence. *Proceedings of the National Academy of Sciences*, 2020.
- 596 [84] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degener-
597 ate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience
598 Meeting (COSYNE), Lisbon, Portugal*, 2019.
- 599 [85] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller,
600 Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience
601 with degenerate solution networks. *Bernstein Conference 2019, Berlin, Germany*, 2019.
- 602 [86] Marcel Nonnenmacher, Pedro J Goncalves, Giacomo Bassetto, Jan-Matthis Lueckmann, and
603 Jakob H Macke. Robust statistical inference for simulation-based models in neuroscience. In
604 *Bernstein Conference 2018, Berlin, Germany*, 2018.
- 605 [87] Deistler Michael, , Pedro J Goncalves, Kaan Oecal, and Jakob H Macke. Statistical infer-
606 ence for analyzing sloppiness in neuroscience models. In *Bernstein Conference 2019, Berlin,
607 Germany*, 2019.
- 608 [88] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnen-
609 macher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural
610 dynamics. In *Advances in Neural Information Processing Systems*, pages 1289–1299, 2017.
- 611 [89] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast
612 likelihood-free inference with autoregressive flows. In *The 22nd International Conference on
613 Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.
- 614 [90] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized
615 approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–
616 4248. PMLR, 2020.

- 617 [91] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and
618 variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- 619 [92] Sean R Bittner and John P Cunningham. Approximating exponential family models (not
620 single distributions) with a two-network architecture. *arXiv preprint arXiv:1903.07515*, 2019.
- 621 [93] Johan Karlsson, Milena Anguelova, and Mats Jirstrand. An efficient method for structural
622 identifiability analysis of large dynamic systems. *IFAC Proceedings Volumes*, 45(16):941–946,
623 2012.
- 624 [94] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary
625 differential equations. In *Advances in neural information processing systems*, pages 6571–6583,
626 2018.
- 627 [95] Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable
628 gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- 629 [96] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula
630 Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially
631 observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–
632 1929, 2009.
- 633 [97] Dhruva V Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter
634 unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, 2017.
- 635 [98] Maria Pia Saccomani, Stefania Audoly, and Leontina D’Angiò. Parameter identifiability of
636 nonlinear systems: the role of initial conditions. *Automatica*, 39(4):619–632, 2003.
- 637 [99] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Bal-
638 aji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv
639 preprint arXiv:1912.02762*, 2019.
- 640 [100] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolu-
641 tions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 642 [101] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
643 Improved variational inference with inverse autoregressive flow. *Advances in neural informa-
644 tion processing systems*, 29:4743–4751, 2016.

645 [102] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Internation-*
646 *al Conference on Learning Representations*, 2015.

647 [103] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for
648 statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

649 **5 Methods**

650 **5.1 Emergent property inference (EPI)**

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

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

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

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

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

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

677 Such deep probability distributions embed the posterior distribution in a deep network. Once
 678 optimized, this deep network representation has remarkably useful properties: immediate posterior
 679 sampling, and immediate probability, gradient, and Hessian evaluation at any parameter choice.
 680 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
 681 the neural network parameters θ to find a maximally entropic distribution q_{θ}^* within the deep
 682 variational family \mathcal{Q} producing the emergent property \mathcal{X} :

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

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

683 Entropy is chosen as the normative selection principle, since we want the posterior to only contain
 684 structure predicated by the emergent property [58, 59]. This choice of selection principle is also
 685 that of standard Bayesian inference, and we derive an exact relation between EPI and variational
 686 inference (see Section 5.1.5). However, a key difference is that variational inference and other
 687 Bayesian methods do not constrain the predictions of their inferred posteriors. This optimization
 688 is executed using the algorithm of Maximum Entropy Flow Networks (MEFNs) [52].

689 In the remainder of Section 5.1, we will explain the finer details and motivation of the EPI method.
 690 First, we explain related approaches and what EPI introduces to this domain (Section 5.1.1). Sec-
 691 ond, we describe the special class of deep probability distributions used in EPI called normalizing
 692 flows (Section 5.1.2). Next, we explain the constrained optimization technique used to solve Equa-
 693 tion 15 (Section 5.1.3). Then, we demonstrate the details of this optimization in a toy example
 694 (Section 5.1.4). Finally, we establish the known relationship between maximum entropy distribu-
 695 tions and exponential families (Section 5.1.5), which is used to explain the relation between EPI
 696 and variational inference (Section 5.1.6).

697 5.1.1 Related approaches

698 When Bayesian inference problems lack conjugacy, scientists use approximate inference methods
 699 like variational inference (VI) [46] and Markov chain Monte Carlo (MCMC) [45, 44]. After opti-
 700 mization, variational methods return a parameterized posterior distribution, which we can analyze.
 701 Also, the variational approximating distribution class is often chosen such that it permits fast
 702 sampling. In contrast MCMC methods only produce samples from the approximated posterior dis-
 703 tribution. No parameterized distribution is estimated, and additional samples are always generated
 704 with the same sampling complexity. Inference in models defined by systems of differential has been

705 demonstrated with MCMC [81], although this approach requires tractable likelihoods. Advances
706 have leveraged structure in stochastic differential equation models to improve likelihood
707 approximations, thus expanding the domain of applicable models [82].

708 Likelihood-free (or “simulation-based”) inference (LFI) [83] is model parameter inference in the
709 absence of a tractable likelihood function. The most prevalent approach to LFI is approximate
710 Bayesian computation [42], in which satisfactory parameter samples are kept from random prior
711 sampling according to a rejection heuristic. The obtained set of parameters do not have a prob-
712 abilities, and further insight about the model must be gained from examination of the parameter
713 set and their generated activity. Methodological advances to ABC methods have come through
714 the use of Markov chain Monte Carlo (MCMC-ABC) [43] and sequential Monte Carlo (SMC-ABC)
715 [31] sampling techniques. SMC-ABC is considered state-of-the-art ABC, yet this approach still
716 struggles to scale in dimensionality (cf. Fig. 4). Furthermore, once a parameter set has been
717 obtained by SMC-ABC from a finite set of particles, the SMC-ABC algorithm must be run again
718 with a new population of initialized particles to obtain additional samples.

719 For scientific model analysis, we seek a posterior distribution exhibiting the properties of a well-
720 chosen variational approximation: a parametric form conferring analytic calculations, and trivial
721 sampling time. For this reason, ABC and MCMC techniques are unattractive, since they only
722 produce a set of parameter samples and have unchanging sampling rate. EPI executes likelihood-
723 free inference using the MEFN [52] algorithm using a deep variational posterior approximation.
724 The deep neural network of EPI defines the parametric form of the posterior approximation. Fur-
725 thermore, the EPI distribution is constrained to produce an emergent property. In other words,
726 the summary statistics of the posterior predictive distribution are fixed to have certain first and
727 second moments. EPI optimization is enabled using stochastic gradient techniques in the spirit
728 of likelihood-free variational inference [49]. The analytic relationship between EPI and variational
729 inference is explained in Secton 5.1.6.

730 We note that, during our preparation and early presentation of this work [84, 85], another work
731 has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural
732 circuits ([86, 87, 75]). We are encouraged by this general problem being recognized by others in the
733 community, and we emphasize that these works offer complementary neuroscientific contributions
734 (different theoretical models of focus) and use different technical methodologies (ours is built on
735 our prior work [52], theirs similarly [88]).

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

737 recently developed LFI methods in which two neural networks are used for posterior inference.
738 This first neural network is a normalizing flow used to estimate the posterior $p(\mathbf{z} | \mathbf{x})$ (SNPE)
739 or the likelihood $p(\mathbf{x} | \mathbf{z})$ (sequential neural likelihood (SNL [89])). A recent advance uses an
740 unconstrained neural network to estimate the likelihood ratio (sequential neural ratio estimation
741 (SNRE [90])). In SNL and SNRE, MCMC sampling techniques are used to obtain samples from
742 the approximated posterior. This contrasts with EPI and SNPE, which afford a normalizing flow
743 approximation to the posterior, which facilitates immediate measurements of sample probability,
744 gradient, or Hessian for system analysis. The second neural network in this sequential class of
745 methods is the amortizer. This network maps data \mathbf{x} (or statistics $f(\mathbf{x}; \mathbf{z})$ or model parameters \mathbf{z})
746 to the weights and biases of the first neural network. These methods are optimized on a conditional
747 density (or ratio) estimation objective on a sequentially adapting finite sample-based approximation
748 to the posterior.

749 The approximating fidelity of the first neural network in sequential approaches is optimized to
750 generalize across the entire distribution it is conditioned upon. This optimization towards gen-
751 eralization of sequential methods can reduce the accuracy at the singular posterior of interest.
752 Whereas in EPI, the entire expressivity of the normalizing flow is dedicated to learning a single
753 distribution as well as possible. While amortization is not possible in EPI parameterized by the
754 mean parameter μ (due to the inverse mapping problem [91]), we have shown this two-network
755 amortization approach to be effective in exponential family distributions defined by their natural
756 parameterization [92].

757 Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in
758 natural models. Around a point, one can measure the Jacobian. One approach that scales well is
759 EAR [93]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [94] and
760 its stochastic adaptation [95]. Casting identifiability as a statistical estimation problem, the profile
761 likelihood can assess via iterated optimization while holding parameters fixed [96]. An exciting
762 recent method is capable of recovering the functional form of such unidentifiabilities away from a
763 point by following degenerate dimensions of the fisher information matrix [97]. Global structural
764 non-identifiabilities can be found for models with polynomial or rational dynamics equations using
765 DAISY [98]. With EPI, we have all the benefits given by a statistical inference method plus the
766 ability to query the gradient or Hessian of the inferred distribution at any chosen parameter value.

767 **5.1.2 Normalizing flows**

768 Deep probability distributions are comprised of multiple layers of fully connected neural networks
 769 (Equation). When each neural network layer is restricted to be a bijective function, the sample
 770 density can be calculated using the change of variables formula at each layer of the network. For
 771 $\mathbf{z}_i = g_i(\mathbf{z}_{i-1})$,

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

772 However, this computation has cubic complexity in dimensionality for fully connected layers. By
 773 restricting our layers to normalizing flows [47, 99] – bijective functions with fast log determinant
 774 Jacobian computations, which confer a fast calculation of the sample log probability. Fast log
 775 probability calculation confers efficient optimization of the maximum entropy objective (see Section
 776 5.1.3). We use the Real NVP [50] normalizing flow class, because its coupling architecture confers
 777 both fast sampling (forward) and fast log probability evaluation (backward). Fast probability
 778 evaluation in turn facilitates fast gradient and Hessian evaluation of log probability throughout
 779 parameter space. Glow permutations were used in between coupling stages [100]. This is in contrast
 780 to autoregressive architectures [51, 101], in which only forward or backward passes are efficient. In
 781 this work, normalizing flows are used as flexible posterior approximations $q_{\boldsymbol{\theta}}(\mathbf{z})$ having weights and
 782 biases $\boldsymbol{\theta}$. We specify the architecture used in each application by the number of Real-NVP affine
 783 coupling stages, and the number of neural network layers and units per layer of the conditioning
 784 functions.

785 **5.1.3 Augmented Lagrangian optimization**

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

$$L(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{opt}}, c) = -H(q_{\boldsymbol{\theta}}) + \boldsymbol{\eta}_{\text{opt}}^\top R(\boldsymbol{\theta}) + \frac{c}{2} \|R(\boldsymbol{\theta})\|^2 \quad (17)$$

788 where average constraint violations $R(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [T(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_{\text{opt}}]]$, $\boldsymbol{\eta}_{\text{opt}} \in \mathbb{R}^m$ are the
 789 Lagrange multipliers where $m = |\boldsymbol{\mu}_{\text{opt}}| = |T(\mathbf{x}; \mathbf{z})| = 2|f(\mathbf{x}; \mathbf{z})|$, and c is the penalty coefficient.
 790 The sufficient statistics $T(\mathbf{x}; \mathbf{z})$ and mean parameter $\boldsymbol{\mu}_{\text{opt}}$ are determined by the means $\boldsymbol{\mu}$ and
 791 variances $\boldsymbol{\sigma}^2$ of emergent property statistics $f(\mathbf{x}; \mathbf{z})$ defined in Equation 15. Specifically, $T(\mathbf{x}; \mathbf{z})$ is
 792 a concatenation of the first and second moments, $\boldsymbol{\mu}_{\text{opt}}$ is a concatenation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ (see section
 793 5.1.5), and the Lagrange multipliers are closely related to the natural parameters $\boldsymbol{\eta}$ of exponential

794 families (see Section 5.1.6). Weights and biases $\boldsymbol{\theta}$ of the deep probability distribution are optimized
795 according to Equation 17 using the Adam optimizer with learning rate 10^{-3} [102].

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

$$H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \int -q_{\boldsymbol{\theta}}(\mathbf{z}) \log(q_{\boldsymbol{\theta}}(\mathbf{z})) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [-\log(q_{\boldsymbol{\theta}}(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (18)$$

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

$$\nabla_{\boldsymbol{\theta}} H(q_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_{\mathbf{z}_0 \sim q_0} [-\nabla_{\boldsymbol{\theta}} \log(q_{\boldsymbol{\theta}}(g_{\boldsymbol{\theta}}(\mathbf{z}_0)))]. \quad (19)$$

801 The lagrangian parameters $\boldsymbol{\eta}_{\text{opt}}$ are initialized to zero and adapted following each augmented
802 Lagrangian epoch, which is a period of optimization with fixed $(\boldsymbol{\eta}_{\text{opt}}, c)$ for a given number of
803 stochastic optimization iterations. A low value of c is used initially, and conditionally increased
804 after each epoch based on constraint error reduction. The penalty coefficient is updated based
805 on the result of a hypothesis test regarding the reduction in constraint violation. The p-value of
806 $\mathbb{E}[|R(\boldsymbol{\theta}_{k+1})|] > \gamma \mathbb{E}[|R(\boldsymbol{\theta}_k)|]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. The
807 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
808 the study, $\gamma = 0.25$, while β was chosen to be either 2 or 4. The batch size of EPI also varied
809 according to application.

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

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

823 To assess whether the EPI distribution $q_{\theta}(\mathbf{z})$ produces the emergent property, we assess whether
 824 each individual constraint on the means and variances of $f(\mathbf{x}; \mathbf{z})$ is satisfied. We consider the EPI
 825 to have converged when a null hypothesis test of constraint violations $R(\boldsymbol{\theta})_i$ being zero is accepted
 826 for all constraints $i \in \{1, \dots, m\}$ at a significance threshold $\alpha = 0.05$. This significance threshold is
 827 adjusted through Bonferroni correction according to the number of constraints m . The p-values for
 828 each constraint are calculated according to a two-tailed nonparametric test, where 200 estimations
 829 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
 830 Lagrangian epoch.

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

843 5.1.4 Example: 2D LDS

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

$$846 \quad \tau \frac{d\mathbf{x}}{dt} = A\mathbf{x} \quad (20)$$

846 with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}. \quad (21)$$

847 To run EPI with the dynamics matrix elements as the free parameters $\mathbf{z} = [a_1, a_2, a_3, a_4]$ (fix-
 848 ing $\tau = 1$), the emergent property statistics $T(\mathbf{x})$ were chosen to contain the first and second
 849 moments of the oscillatory frequency, $\frac{\text{imag}(\lambda_1)}{2\pi}$, and the growth/decay factor, $\text{real}(\lambda_1)$, of the oscil-
 850 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 0.25^2 , and the oscillation frequency $2\pi\text{imag}(\lambda_1)$ having mean $\omega = 1$ Hz with variance $(0.1\text{Hz})^2$:

$$\mathbb{E}[T(\mathbf{x})] \triangleq \mathbb{E} \begin{bmatrix} \text{real}(\lambda_1) \\ \text{imag}(\lambda_1) \\ (\text{real}(\lambda_1) - 0)^2 \\ (\text{imag}(\lambda_1) - 2\pi\omega)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2\pi\omega \\ 0.25^2 \\ (2\pi 0.1)^2 \end{bmatrix} \triangleq \boldsymbol{\mu}. \quad (22)$$

855

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:

$$\lambda = \frac{\left(\frac{a_1+a_4}{\tau}\right) \pm \sqrt{\left(\frac{a_1+a_4}{\tau}\right)^2 + 4\left(\frac{a_2a_3-a_1a_4}{\tau}\right)}}{2}. \quad (23)$$

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 [91]. 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 .

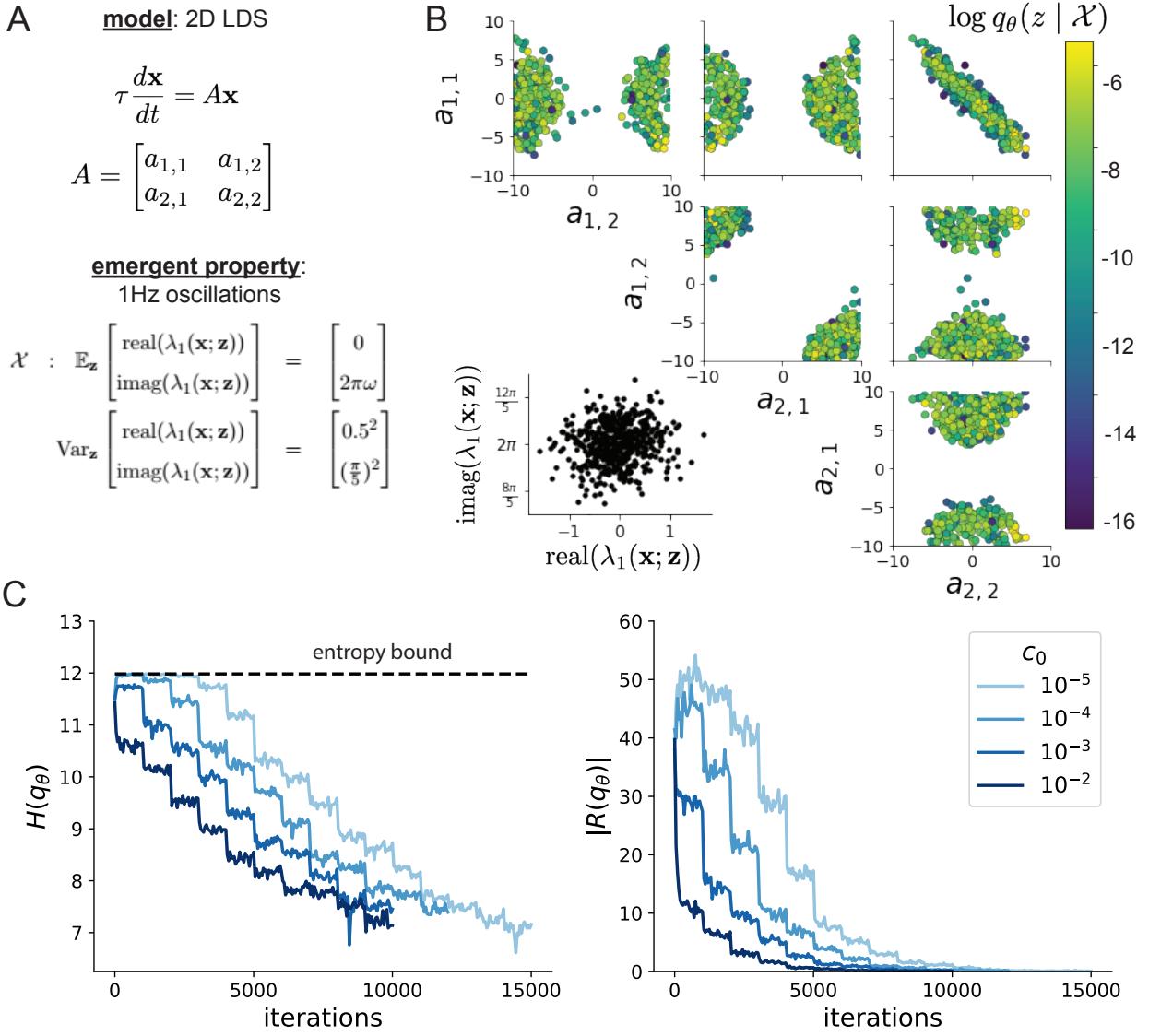


Figure 5: (LDS1): A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The EPI distribution for a two-dimensional linear dynamical system with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. Dashed 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.

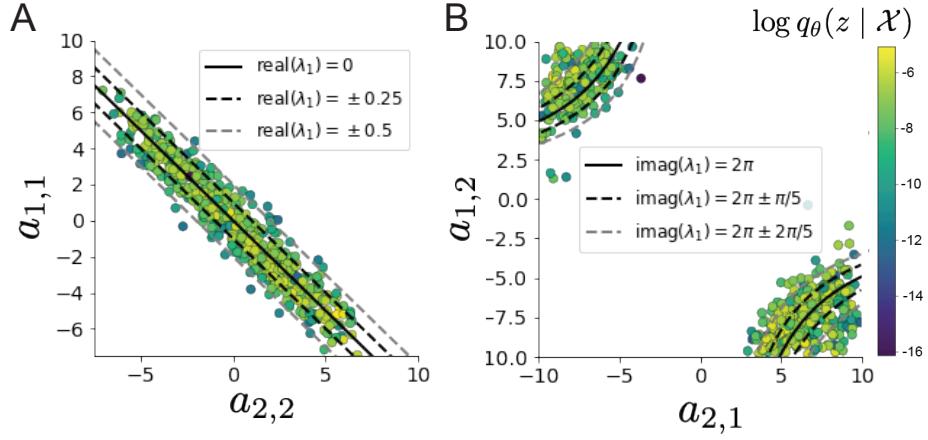


Figure 6: (LDS2): A. Probability contours in the a_1-a_4 plane were derived from the relationship to emergent property statistic of growth/decay factor $\text{real}(\lambda_1)$. B. Probability contours in the a_2-a_3 plane were derived from the emergent property statistic of oscillation frequency $2\pi\text{imag}(\lambda_1)$.

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

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4 - a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases}. \quad (24)$$

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

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \quad (25)$$

878 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
879 $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S2B). With $\sigma_{1,4} = \mathbb{E}_{\mathbf{z} \sim q_\theta} [|a_1a_4 - E_{q_\theta}[a_1a_4]|]$,
880 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
881 dotted). This validates the curved structure of the inferred distribution learned through EPI. We
882 took steps in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive
883 values a_1a_4 in the learned distribution. Subtler combinations of model and emergent property will
884 have more complexity, further motivating the use of EPI for understanding these systems. As we
885 expect, the distribution results in samples of two-dimensional linear systems oscillating near 1Hz
886 (Fig. S3).

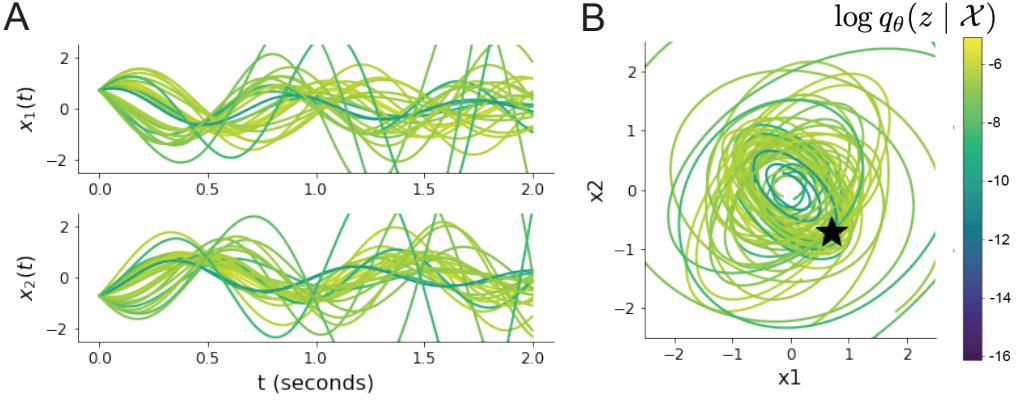


Figure 7: (LDS3): Sampled dynamical systems $\mathbf{z} \sim q_{\theta}(\mathbf{z})$ and their simulated activity from $\mathbf{x}(0) = [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]$ colored by log probability. A. Each dimension of the simulated trajectories throughout time. B The simulated trajectories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

887 5.1.5 Maximum entropy distributions and exponential families

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

$$p^*(\mathbf{z}) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} H(p(\mathbf{z})) \quad (26)$$

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

890 will have probability density in the exponential family:

$$p^*(\mathbf{z}) \propto \exp(\boldsymbol{\eta}^\top T(\mathbf{z})). \quad (27)$$

891 The mappings between the mean parameterization $\boldsymbol{\mu}_{\text{opt}}$ and the natural parameterization $\boldsymbol{\eta}$ are
 892 formally hard to identify [91].

893 In EPI, emergent properties are defined as statistics having a fixed mean and variance as in Equation
 894 4

$$\mathbb{E}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\mu}, \operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \boldsymbol{\sigma}^2. \quad (28)$$

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

$$\operatorname{Var}_{\mathbf{z}, \mathbf{x}}[f(\mathbf{x}; \mathbf{z})] = \mathbb{E}_{\mathbf{z}, \mathbf{x}}[(f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2] \quad (29)$$

896 As a general maximum entropy distribution (Equation 26), the sufficient statistics vector contains

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

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu})^2 \end{bmatrix}, \quad (30)$$

898 which are constrained to the chosen means and variances

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} \boldsymbol{\mu} \\ \sigma^2 \end{bmatrix}. \quad (31)$$

899 5.1.6 EPI as variational inference

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

$$p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \quad (32)$$

904 The posterior distribution is analytically available when the prior is conjugate with the likelihood.
 905 However, conjugacy is rare in practice, and alternative methods, such as variational inference [103],
 906 are utilized.

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

$$q_{\boldsymbol{\theta}}^*(\mathbf{z}) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}}(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})). \quad (33)$$

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

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

$$= -H(q_{\boldsymbol{\theta}}) - \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z})) - \log(p(\mathbf{x}))] \quad (35)$$

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

$$\underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmin}} KL(q_{\boldsymbol{\theta}} || p(\mathbf{z} | \mathbf{x})) = \underset{q_{\boldsymbol{\theta}} \in \mathcal{Q}}{\operatorname{argmax}} H(q_{\boldsymbol{\theta}}) + \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\theta}}} [\log(p(\mathbf{x} | \mathbf{z})) + \log(p(\mathbf{z}))]. \quad (36)$$

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

$$p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(f(\mathbf{x}; \mathbf{z}) | \boldsymbol{\mu}_f, \Sigma_f), \quad (37)$$

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

$$\log p(\mathbf{x} | \mathbf{z}) \propto \boldsymbol{\eta}_f^\top T(\mathbf{x}, \mathbf{z}), \quad (38)$$

918 where

$$\boldsymbol{\eta}_f = \begin{bmatrix} \frac{\boldsymbol{\mu}_f}{\sigma_f^2} \\ \frac{-1}{2\sigma_f^2} \end{bmatrix}, \text{ and} \quad (39)$$

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} f(\mathbf{x}; \mathbf{z}) \\ (f(\mathbf{x}; \mathbf{z}) - \boldsymbol{\mu}_f)^2 \end{bmatrix}. \quad (40)$$

920 The variational objective is then

$$\underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta) + \boldsymbol{\eta}_f^\top \mathbb{E}_{\mathbf{z} \sim q_\theta} [T(\mathbf{x}; \mathbf{z})] \quad (41)$$

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

$$\begin{aligned} q_\theta^*(\mathbf{z}) &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top (\mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})] - \boldsymbol{\mu}_{\text{opt}}) \\ &= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \boldsymbol{\eta}_{\text{opt}}^\top \mathbb{E}_{\mathbf{z}, \mathbf{x}} [T(\mathbf{x}; \mathbf{z})]. \end{aligned} \quad (42)$$

923 where $T(\mathbf{x}; \mathbf{z})$ consists of the first and second moments of the emergent property statistic $f(\mathbf{x}; \mathbf{z})$
 924 (Equation 30). Thus, EPI is implicitly executing variational inference with a uniform prior and a
 925 mean-field gaussian likelihood on the emergent property statistics. The data \mathbf{x} used by this implicit
 926 variational inference program would be that generated by the adapting variational approximation
 927 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{z})q_\theta(\mathbf{z})$, and the likelihood parameters $\boldsymbol{\eta}_f$ of EPI optimization epoch k are predicated
 928 by $\boldsymbol{\eta}_{\text{opt},k}$. However, in EPI we have not specified a prior distribution, or collected data, which can
 929 inform us about model parameters. Instead we have a mathematical specification of an emergent
 930 property, which the model must produce, and a maximum entropy selection principle. Accordingly,
 931 we replace the notation of $p(\mathbf{z} | \mathbf{x})$ with $p(\mathbf{z} | \mathcal{X})$ conceptualizing an inferred distribution that obeys
 932 emergent property \mathcal{X} (see Section 5.1).

933 5.2 Theoretical models

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

936 **5.2.1 Stomatogastric ganglion**

937 We analyze how the parameters $\mathbf{z} = [g_{el}, g_{synA}]$ govern the emergent phenomena of intermediate
 938 hub frequency in a model of the stomatogastric ganglion (STG) [57] shown in Figure 1A with
 939 activity $\mathbf{x} = [x_{f1}, x_{f2}, x_{hub}, x_{s1}, x_{s2}]$, using the same hyperparameter choices as Gutierrez et al.
 940 Each neuron's membrane potential $x_\alpha(t)$ for $\alpha \in \{f1, f2, hub, s1, s2\}$ is the solution of the following
 941 stochastic differential equation:

$$C_m \frac{dx_\alpha}{dt} = -[h_{leak}(\mathbf{x}; \mathbf{z}) + h_{Ca}(\mathbf{x}; \mathbf{z}) + h_K(\mathbf{x}; \mathbf{z}) + h_{hyp}(\mathbf{x}; \mathbf{z}) + h_{elec}(\mathbf{x}; \mathbf{z}) + h_{syn}(\mathbf{x}; \mathbf{z})] + dB. \quad (43)$$

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

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

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

$$h_{elec}(\mathbf{x}; \mathbf{z}) = g_{el}(x_\alpha^{post} - x_\alpha^{pre}) \quad (45)$$

$$h_{syn}(\mathbf{x}; \mathbf{z}) = g_{syn}S_\infty^{pre}(x_\alpha^{post} - V_{syn}) \quad (46)$$

$$h_{Ca}(\mathbf{x}; \mathbf{z}) = g_{Ca}M_\infty(x_\alpha - V_{Ca}) \quad (47)$$

$$h_K(\mathbf{x}; \mathbf{z}) = g_KN(x_\alpha - V_K) \quad (48)$$

$$h_{hyp}(\mathbf{x}; \mathbf{z}) = g_hH(x_\alpha - V_{hyp}). \quad (49)$$

953 The reversal potentials were set to $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$,
 954 and $V_{syn} = -75mV$. The other conductance parameters were fixed to $g_{leak} = 1 \times 10^{-4}\mu S$. g_{Ca} ,
 955 g_K , and g_{hyp} had different values based on fast, intermediate (hub) or slow neuron. The fast
 956 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
 957 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
 958 slow conductances had values $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

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

$$M_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_1}{v_2} \right) \right) \quad (50)$$

961

$$\frac{dN}{dt} = \lambda_N(N_\infty - N) \quad (51)$$

962

$$N_\infty = 0.5 \left(1 + \tanh \left(\frac{x_\alpha - v_3}{v_4} \right) \right) \quad (52)$$

963

$$\lambda_N = \phi_N \cosh \left(\frac{x_\alpha - v_3}{2v_4} \right) \quad (53)$$

964

$$\frac{dH}{dt} = \frac{(H_\infty - H)}{\tau_h} \quad (54)$$

965

$$H_\infty = \frac{1}{1 + \exp \left(\frac{x_\alpha + v_5}{v_6} \right)} \quad (55)$$

966

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp \left(\frac{-x_\alpha + v_7}{v_8} \right)} \right). \quad (56)$$

967 where we set $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$,

968 $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

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

$$S_\infty = \frac{1}{1 + \exp \left(\frac{v_{th} - x_\alpha}{v_9} \right)}. \quad (57)$$

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

973 In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for
 974 $T = 300$ time steps of $dt = 25\text{ms}$. The chosen dt and T were the most computationally convenient
 975 choices yielding accurate frequency measurement. We used a basis of complex exponentials with
 976 frequencies from 0.0-1.0 Hz at 0.01Hz resolution to measure frequency from simulated time series

$$\Phi = [0.0, 0.01, \dots, 1.0]^\top .. \quad (58)$$

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

981 To differentiate through the maximum frequency identification, we used a soft-argmax Let $X_\alpha \in$
 982 $\mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_\alpha \in \mathbb{R}^N$, where $\alpha \in$

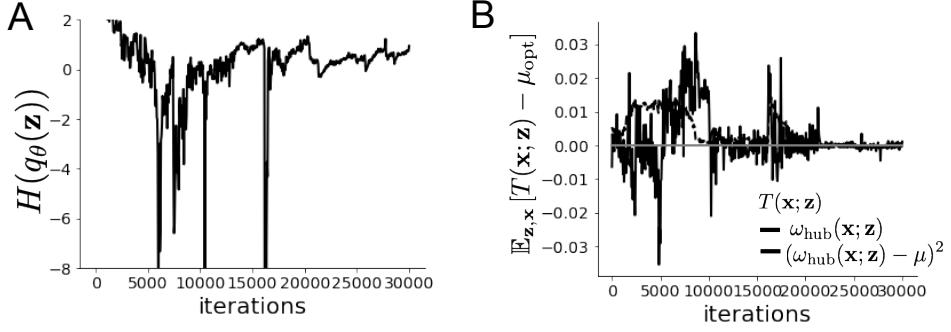


Figure 8: (STG1): EPI optimization of the STG model producing network syncing. A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 25,000 iterations following the fifth augmented Lagrangian epoch.

983 $\{f_1, f_2, \text{hub}, s_1, s_2\}$. The soft-argmax is then calculated using temperature parameter $\beta = 100$

$$\psi_\alpha = \text{softmax}(\beta |X_\alpha| \odot i), \quad (59)$$

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

$$\omega_\alpha = 0.01\psi_\alpha \text{Hz}. \quad (60)$$

985 Intermediate hub frequency, like all other emergent properties in this work, is defined by the mean
986 and variance of the emergent property statistics. In this case, we have one statistic, hub neuron
987 frequency, where the mean was chosen to be 0.55Hz, and variance was chosen to be $(0.025\text{Hz})^2$ to
988 capture variation in frequency between 0.5Hz and 0.6Hz (Equation 4). As a maximum entropy dis-
989 tribution, $T(\mathbf{x}, \mathbf{z})$ is comprised of both these first and second moments of the hub neuron frequency
990 (as in Equations 30 and 31)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} \omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) \\ (\omega_{\text{hub}}(\mathbf{x}; \mathbf{z}) - 0.55)^2 \end{bmatrix}, \quad (61)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 0.55 \\ 0.025^2 \end{bmatrix}. \quad (62)$$

991
992 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
993 epoch of 5,000 iterations(see Section 5.1.3). The optimization converged after five epochs (Fig. S4).

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

996 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
 997 returned by a preliminary ABC search. We did not include $g_{\text{synA}} < 0.01$, for numerical stability.
 998 EPI optimization was run using 5 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an
 999 augmented Lagrangian coefficient of $c_0 = 10^5$, a batch size $n = 400$, and $\beta = 2$. The distribution
 1000 shown is that of the architecture converging with criteria $N_{\text{test}} = 100$ at greatest entropy across
 1001 random seeds.

1002 We calculated the Hessian at the mode of the inferred EPI distribution. The Hessian of a probability
 1003 model is the second order gradient of the log probability density $\log q_{\boldsymbol{\theta}}(\mathbf{z})$ with respect to the
 1004 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
 1005 throughout distribution, to indicate the dimensions of parameter space that are sensitive (strongly
 1006 negative eigenvalue), and which are degenerate (low magnitude eigenvalue) with respect to the
 1007 emergent property produced. In Figure 1D, the eigenvectors of the Hessian v_1 (solid) and v_2
 1008 (dashed) are shown evaluated at the mode of the distribution. The length of the arrows is inversely
 1009 proportional to the square root of absolute value of their eigenvalues $\lambda_1 = -10.7$ and $\lambda_2 = -3.22$.
 1010 Since the Hessian eigenvectors have sign degeneracy, the visualized directions in 2-D parameter
 1011 space are chosen arbitrarily.

1012 5.2.2 Primary visual cortex

1013 In the stochastic stabilized supralinear network [71], population rate responses \mathbf{x} to input \mathbf{h} , recur-
 1014 rent input $W\mathbf{x}$ and slow noise $\boldsymbol{\epsilon}$ are governed by

$$\tau \frac{d\mathbf{x}}{dt} = -\mathbf{x} + \phi(W\mathbf{x} + \mathbf{h} + \boldsymbol{\epsilon}), \quad (63)$$

1015 where the noise is an Ornstein-Uhlenbeck process $\boldsymbol{\epsilon} \sim OU(\tau_{\text{noise}}, \boldsymbol{\sigma})$

$$\tau_{\text{noise}} d\epsilon_\alpha = -\epsilon_\alpha dt + \sqrt{2\tau_{\text{noise}}} \tilde{\sigma}_\alpha dB \quad (64)$$

1016 with $\tau_{\text{noise}} = 5\text{ms} > \tau = 1\text{ms}$. The noisy process is parameterized as

$$\tilde{\sigma}_\alpha = \sigma_\alpha \sqrt{1 + \frac{\tau}{\tau_{\text{noise}}}}, \quad (65)$$

1017 so that $\boldsymbol{\sigma}$ parameterizes the variance of the noisy input in the absence of recurrent connectivity
 1018 ($W = \mathbf{0}$). As contrast increases, input to the E- and P-populations increases relative to a baseline
 1019 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
 1020 deterministic V1 circuit model [72]

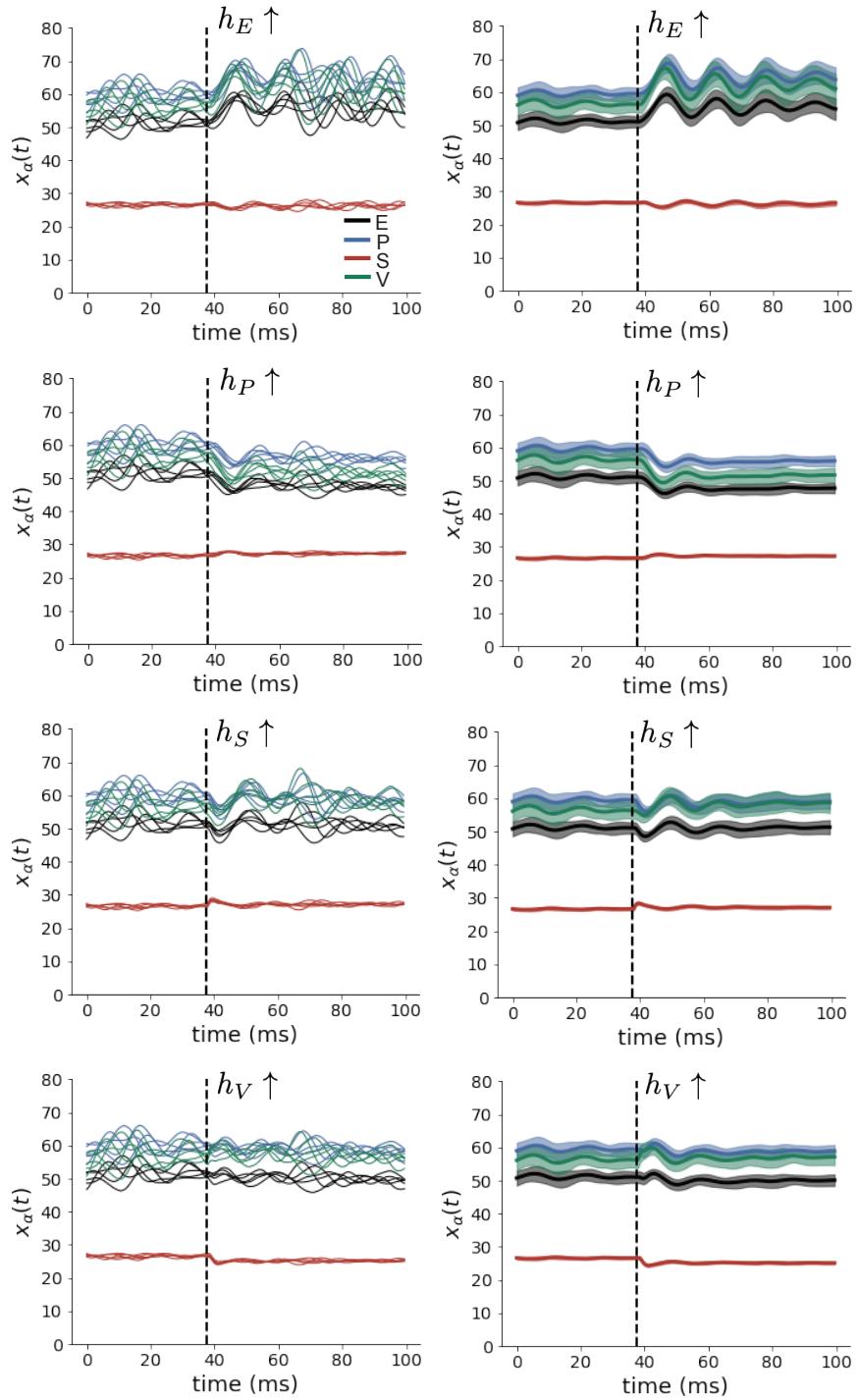


Figure 9: (V1 1) (Left) Simulations for small increases in neuron-type population input. Input magnitudes are chosen so that effect is salient (0.002 for E and P, but 0.02 for S and V). (Right) Average (solid) and standard deviation (shaded) of stochastic fluctuations of responses.

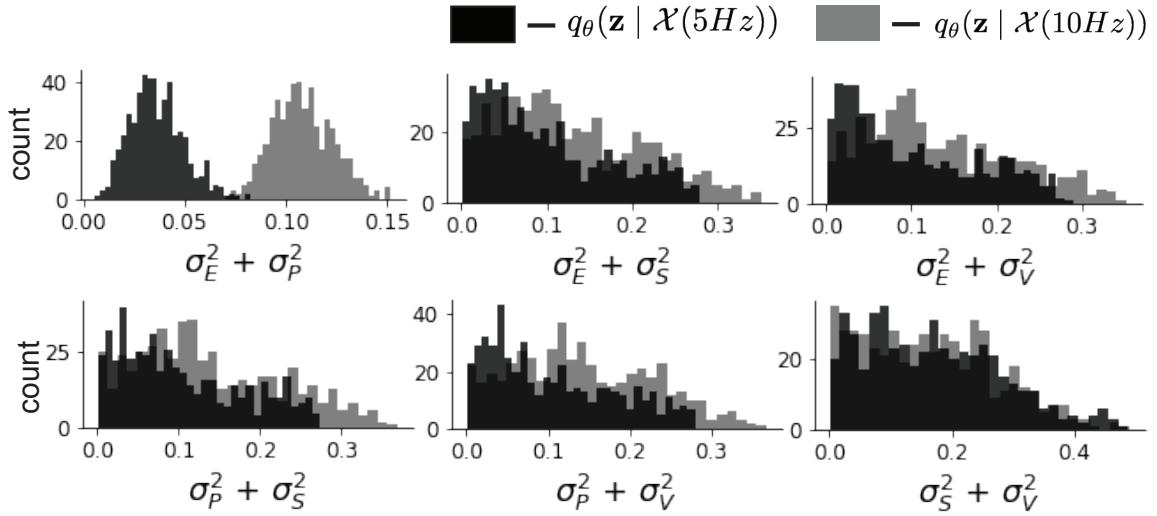


Figure 10: (V1 2) Posterior predictive distributions of the sum of squares of each pair of noise parameters.

$$W_{\text{fit}} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} 2.18 & -1.19 & -.594 & -.229 \\ 1.66 & -.651 & -.680 & -.242 \\ .895 & -5.22 \times 10^{-3} & -1.51 \times 10^{-4} & -.761 \\ 3.34 & -2.31 & -.254 & -2.52 \times 10^{-4} \end{bmatrix}, \quad (66)$$

$$\mathbf{h}_{b,\text{fit}} = \begin{bmatrix} .416 \\ .429 \\ .491 \\ .486 \end{bmatrix}, \quad (67)$$

1021 and

$$\mathbf{h}_{c,\text{fit}} = \begin{bmatrix} .359 \\ .403 \\ 0 \\ 0 \end{bmatrix}. \quad (68)$$

1022 To obtain rates on a realistic scale (100-fold greater), we map these fitted parameters to an equivalence class
1023

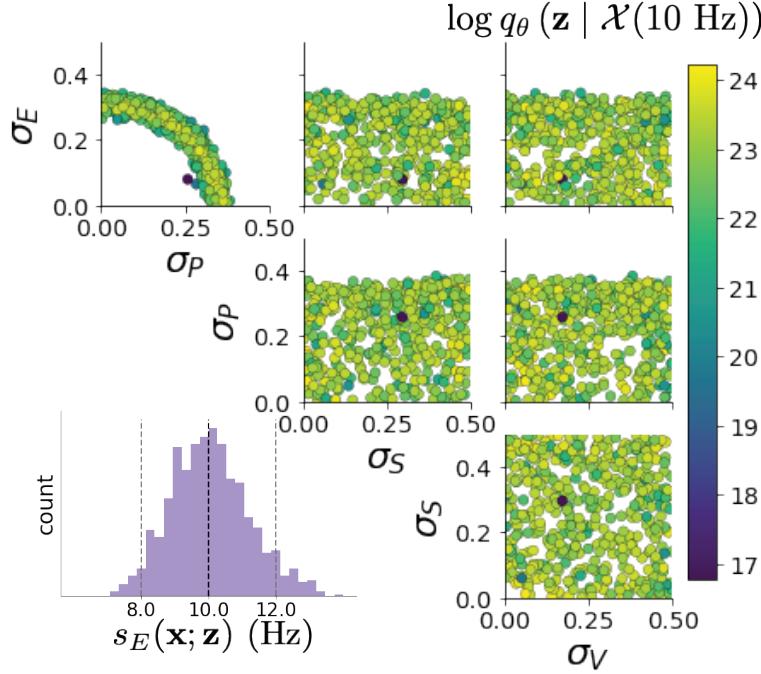


Figure 11: (V1 3) EPI posterior for $\mathcal{X}(10 \text{ Hz})$.

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & W_{EV} \\ W_{PE} & W_{PP} & W_{PS} & W_{PV} \\ W_{SE} & W_{SP} & W_{SS} & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & W_{VV} \end{bmatrix} = \begin{bmatrix} .218 & -.119 & -.0594 & -.0229 \\ .166 & -.0651 & -.068 & -.0242 \\ .0895 & -5.22 \times 10^{-4} & -1.51 \times 10^{-5} & -.0761 \\ .334 & -.231 & -.0254 & -2.52 \times 10^{-5} \end{bmatrix}, \quad (69)$$

$$\mathbf{h}_b = \begin{bmatrix} h_{b,E} \\ h_{b,P} \\ h_{b,S} \\ h_{b,V} \end{bmatrix} = \begin{bmatrix} 4.16 \\ 4.29 \\ 4.91 \\ 4.86 \end{bmatrix}, \quad (70)$$

¹⁰²⁴ and

$$\mathbf{h}_c = \begin{bmatrix} h_{c,E} \\ h_{c,P} \\ h_{c,S} \\ h_{c,V} \end{bmatrix} = \begin{bmatrix} 3.59 \\ 4.03 \\ 0 \\ 0 \end{bmatrix}. \quad (71)$$

¹⁰²⁵ Circuit responses are simulated using $T = 200$ time steps at $dt = 0.5\text{ms}$ from an initial condition

1026 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
 1027 as the square root of the temporal variance from $t_{ss} = 75\text{ms}$ to $Tdt = 100\text{ms}$ averaged over 100
 1028 independent trials.

$$s_E(\mathbf{x}; \mathbf{z}) = \mathbb{E}_x \left[\sqrt{\mathbb{E}_{t > t_{ss}} [(x_E(t) - \mathbb{E}_{t > t_{ss}} [x_E(t)])^2]} \right] \quad (72)$$

1029 For EPI in Fig 2D-E, we used a real NVP architecture with three Real NVP coupling layers
 1030 and two-layer neural networks of 50 units per layer. The normalizing flow architecture mapped
 1031 $z_0 \sim \mathcal{N}(\mathbf{0}, I)$ 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
 1032 different random seeds for architecture initialization $\boldsymbol{\theta}$ with an augmented Lagrangian coefficient of
 1033 $c_0 = 10^{-1}$, a batch size $n = 100$, and $\beta = 2$. The distributions shown are those of the architectures
 1034 converging with criteria $N_{\text{test}} = 100$ at greatest entropy across random seeds.

1035 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
 1036 calculated

$$\begin{aligned} \mathbf{z}^*(\sigma_{P,\text{fixed}}) &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } \sigma_P &= \sigma_{P,\text{fixed}} \end{aligned} \quad (73)$$

1037 At each mode \mathbf{z}^* , we calculated the Hessian and visualized the sensitivity dimension in the direction
 1038 of positive σ_E .

1039 5.2.3 Primary visual cortex: challenges to analysis

1040 TODO Agostina and I are putting this together now.

1041 5.2.4 Superior colliculus

1042 In the model of Duan et al [55], there are four total units: two in each hemisphere corresponding to
 1043 the Pro/Contra and Anti/Ipsi populations. They are denoted as left Pro (LP), left Anti (LA), right
 1044 Pro (RP) and right Anti (RA). Each unit has an activity (x_α) and internal variable (u_α) related
 1045 by

$$x_\alpha = \phi(u_\alpha) = \left(\frac{1}{2} \tanh \left(\frac{u_\alpha - a}{b} \right) + \frac{1}{2} \right) \quad (74)$$

1046 where $\alpha \in \{LP, LA, RA, RP\}$, $a = 0.05$ and $b = 0.5$ control the position and shape of the nonlin-
 1047 earity, respectively. During periods of optogenetic inactivation, activity was decreased proportional
 1048 to the optogenetic strength γ

$$x_\alpha = (1 - \gamma)\phi(u_\alpha). \quad (75)$$

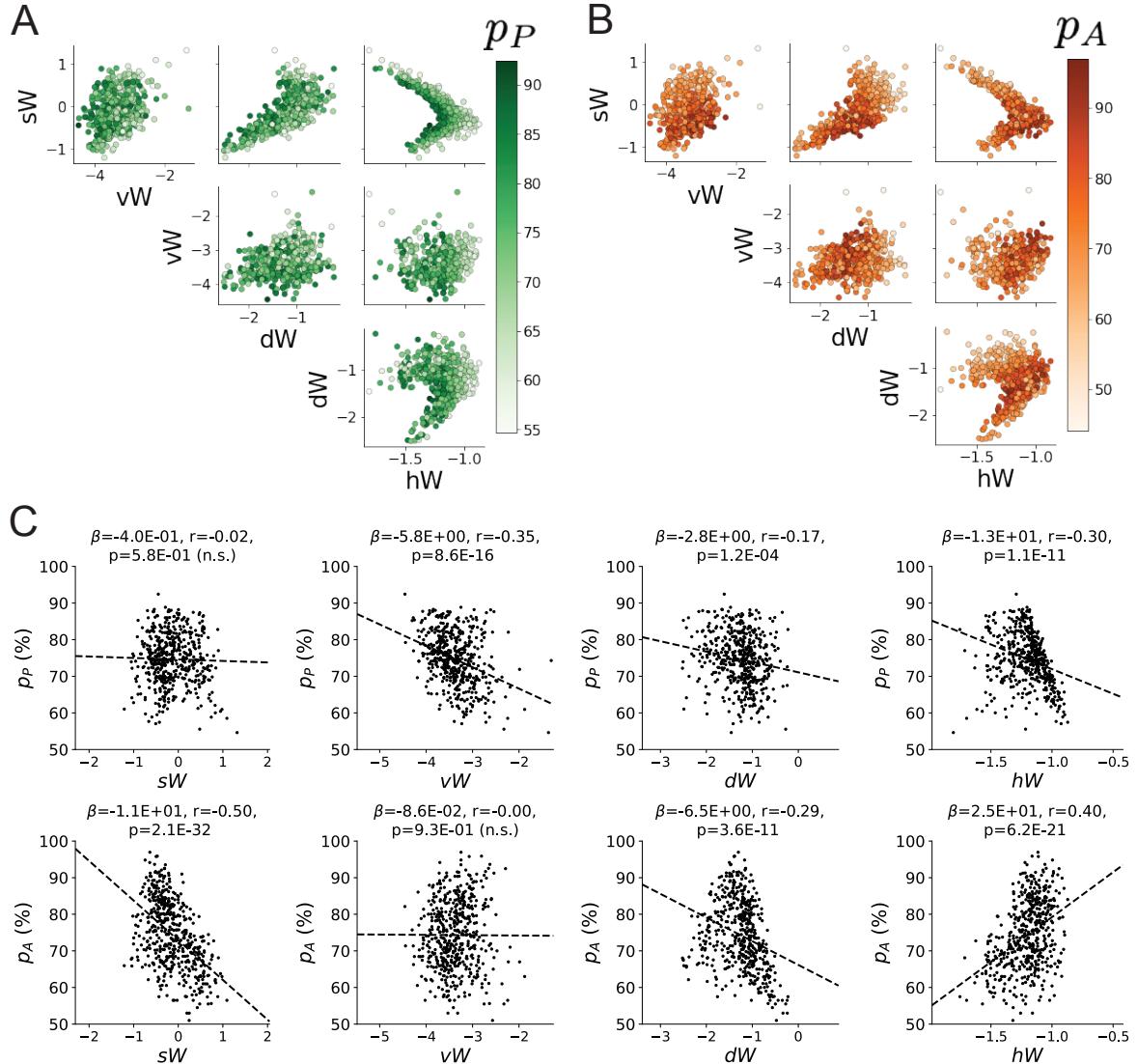


Figure 12: (SC1): **A.** Same pairplot as Fig. 3C colored by Pro task accuracy. **B.** Same as A colored by Anti task accuracy. **C.** Connectivity parameters of EPI distributions versus task accuracies. β is slope coefficient of linear regression, r is correlation, and p is the two-tailed p-value.

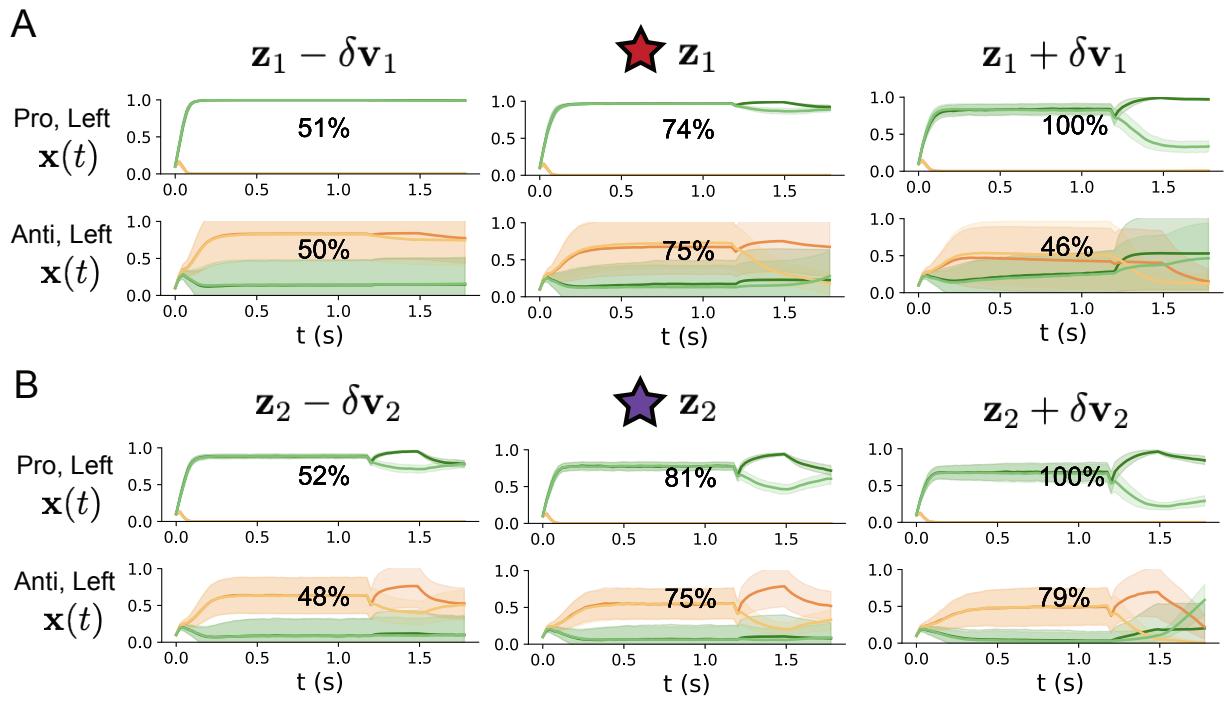


Figure 13: (SC2): **A.** Simulations in network regime \mathbf{z}_1 (center) with simulations given connectivity perturbations in the negative direction of the sensitivity vector \mathbf{v}_1 (left) and positive direction (right). **B.** Same as A for network regime \mathbf{z}_2 .

1049 We order the neural populations of x and u in the following manner

$$\mathbf{x} = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix}, \quad (76)$$

1050 which evolve according to

$$\tau \frac{d\mathbf{u}}{dt} = -\mathbf{u} + W\mathbf{x} + \mathbf{h} + d\mathbf{B}. \quad (77)$$

1051 with time constant $\tau = 0.09s$, step size 24ms and Gaussian noise $d\mathbf{B}$ of variance 0.2^2 . The weight
1052 matrix has 4 parameters sW , vW , hW , and dW :

$$W = \begin{bmatrix} sW & vW & hW & dW \\ vW & sW & dW & hW \\ hW & dW & sW & vW \\ dW & hW & vW & sW \end{bmatrix}. \quad (78)$$

1053 The circuit receives four different inputs throughout each trial, which has a total length of 1.8s.

$$\mathbf{h} = \mathbf{h}_{\text{constant}} + \mathbf{h}_{\text{P,bias}} + \mathbf{h}_{\text{rule}} + \mathbf{h}_{\text{choice-period}} + \mathbf{h}_{\text{light}}. \quad (79)$$

1054 There is a constant input to every population,

$$\mathbf{h}_{\text{constant}} = I_{\text{constant}}[1, 1, 1, 1]^\top, \quad (80)$$

1055 a bias to the Pro populations

$$\mathbf{h}_{\text{P,bias}} = I_{\text{P,bias}}[1, 0, 1, 0]^\top, \quad (81)$$

1056 rule-based input depending on the condition

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

1057

$$\mathbf{h}_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}}[0, 1, 0, 1]^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases}, \quad (83)$$

1058 a choice-period input

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

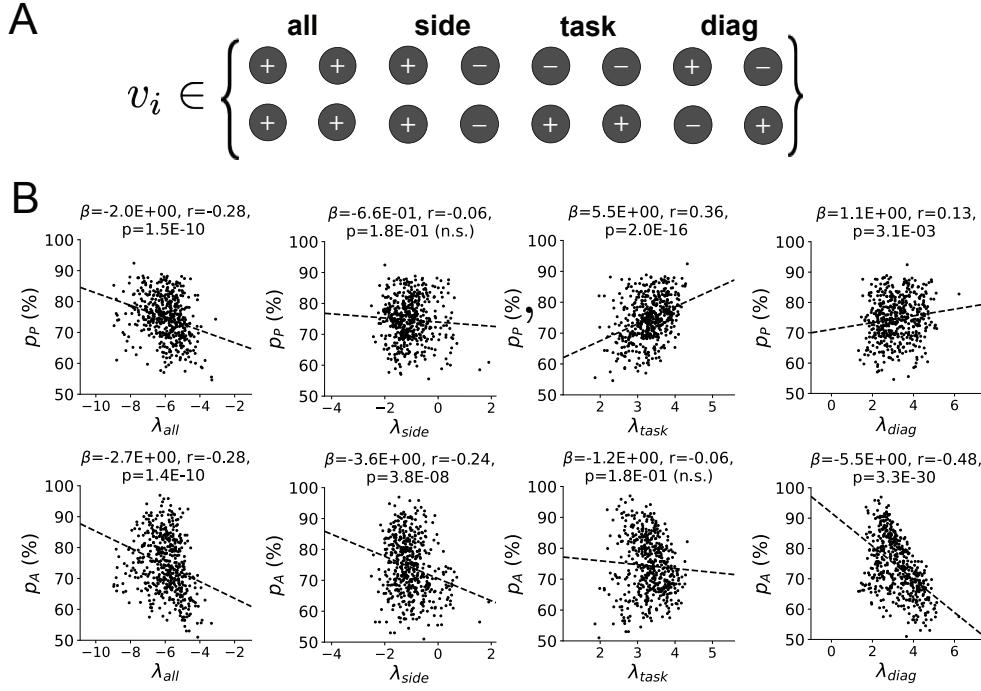


Figure 14: (SC3): **A.** Invariant eigenvectors of connectivity matrix W . **B.** Eigenvalues of connectivities of EPI distribution versus task accuracies.

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

$$\mathbf{h}_{\text{light}}(t) = \begin{cases} I_{\text{light}}[1, 1, 0, 0]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Left} \\ I_{\text{light}}[0, 0, 1, 1]^\top, & \text{if } 1.2s < t < 1.5s \text{ and Right} \\ 0, & \text{otherwise} \end{cases} \quad (85)$$

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

$$p_P(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{LP}(t = 1.8s) - x_{RP}(t = 1.8s)]] \quad (86)$$

and

$$p_A(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})} [\Theta[x_{RP}(t = 1.8s) - x_{LP}(t = 1.8s)]] \quad (87)$$

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

$$\Theta(\mathbf{x}) = \text{sigmoid}(\beta \mathbf{x}), \quad (88)$$

where $\beta = 100$.

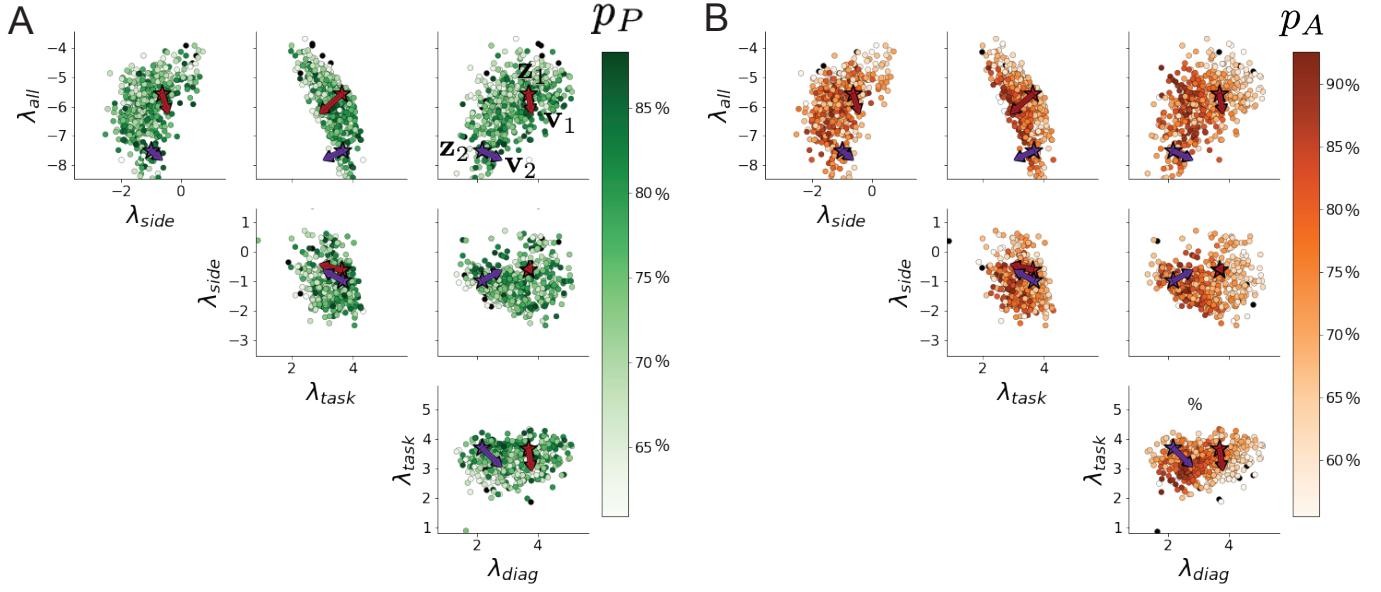


Figure 15: (SC4): **A.** Pairplots of eigenvalues of connectivity matrices in EPI distribution colored by Pro task accuracy. Red and purple stars and arrows correspond to eigenvalues and sensitivity directions \mathbf{z}_1 , \mathbf{z}_2 , \mathbf{v}_1 , and \mathbf{v}_2 . **B.** Same colored by Anti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

1067 Writing the EPI posterior as a maximum entropy distribution, $T(\mathbf{x}; \mathbf{z})$ is comprised of both these
 1068 first and second moments of the accuracy in each task (as in Equations 30 and 31)

$$T(\mathbf{x}; \mathbf{z}) = \begin{bmatrix} p(\mathbf{x}; \mathbf{z})_P \\ p(\mathbf{x}; \mathbf{z})_A \\ (p(\mathbf{x}; \mathbf{z})_P - 75\%)^2 \\ (p(\mathbf{x}; \mathbf{z})_A - 75\%)^2 \end{bmatrix}, \quad (89)$$

$$\boldsymbol{\mu}_{\text{opt}} = \begin{bmatrix} 75\% \\ 75\% \\ 7.5\%^2 \\ 7.5\%^2 \end{bmatrix}. \quad (90)$$

1070 Throughout optimization, the augmented Lagrangian parameters η and c , were updated after each
 1071 epoch of 2,000 iterations(see Section 5.1.3). The optimization converged after six epochs (Fig. 17).

1072 For EPI in Fig. 3C, we used a real NVP architecture with three coupling layers of affine transfor-
 1073 mations parameterized by two-layer neural networks of 50 units per layer. The initial distribution
 1074 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
 1075 augmented Lagrangian coefficient of $c_0 = 10^2$, a batch size $n = 100$, and $\beta = 2$. The distribution

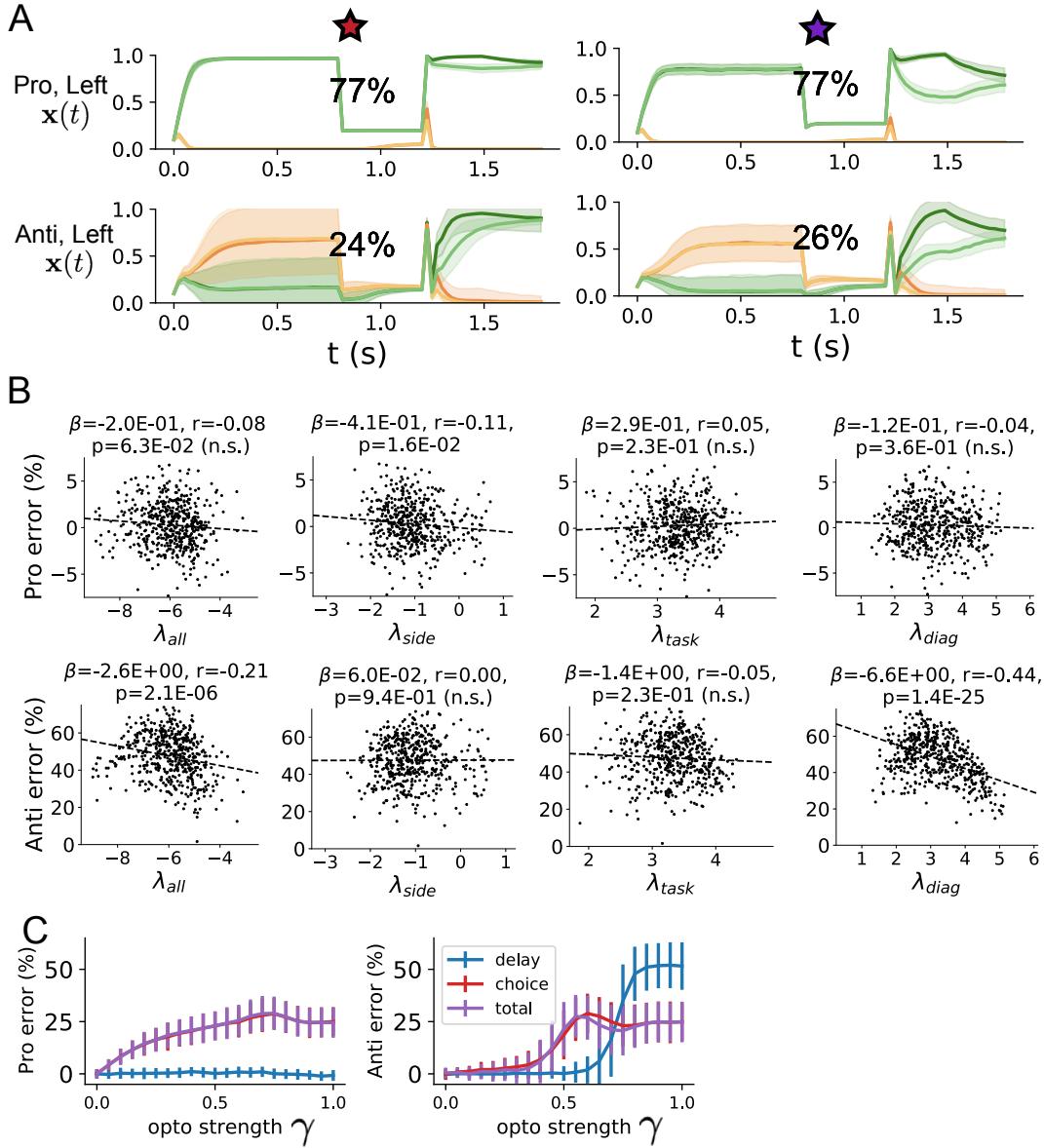


Figure 16: (SC5): **A.** Response of each parameter regime to optogenetic silencing during the delay period. **B.** Connectivity eigenvalues versus the task error induced by delay period inactivation. **C.** Error induced by delay period inactivation with increasing optogenetic strength. Means and standard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

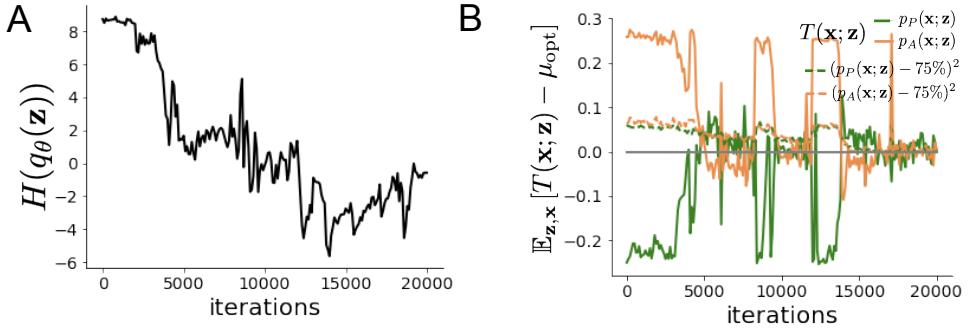


Figure 17: (SC6): A. Entropy throughout optimization. B. The emergent property statistic means and variances converge to their constraints at 20,000 iterations following the tenth augmented Lagrangian epoch.

1076 shown is that of the architecture converging with criteria $N_{\text{test}} = 25$ at greatest entropy across
1077 random seeds.

1078 To make sense of this inferred distribution, we identified two modes used to represent the two
1079 regimes of connectivity in this posterior:

$$\begin{aligned} \mathbf{z}_1 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW > 0 \end{aligned} \tag{91}$$

1080 and

$$\begin{aligned} \mathbf{z}_2 &= \underset{\mathbf{z}}{\operatorname{argmax}} \log q_\theta(\mathbf{z} \mid \mathcal{X}) \\ \text{s.t. } hw &= -1.25, sW < 0 \end{aligned} \tag{92}$$

1081 To understand the connectivity mechanisms governing task accuracy, we took the eigendecomposi-
1082 tion of the symmetric connectivity matrices $W = V\Lambda V^{-1}$, which results in the same basis vectors
1083 \mathbf{v}_i for all W parameterized by \mathbf{z} (Fig. 14A). These basis vectors have intuitive roles in processing
1084 for this task, and are accordingly named the *all* mode - all neurons co-fluctuate, *side* mode - one
1085 side dominates the other, *task* mode - the Pro or Anti populations dominate the other, and *diag*
1086 mode - Pro- and Anti-populations of opposite hemispheres dominate the opposite pair. We found
1087 significant trends across the EPI posterior connectivities: the eigenvalues λ_{task} and λ_{diag} were cor-
1088 related with p_P , while λ_{all} was anticorrelated with p_P . λ_{all} , λ_{side} , and λ_{diag} were all significantly
1089 anticorrelated with p_A .

1090 Under this decomposition, we can re-visualize the posterior in eigenvalue space (Fig. 15). Fur-
1091 thermore, we can project the dimensions of sensitivity into eigenvalue space as well, giving us a
1092 more intuitive sense of how connectivity affects computation in each regime. We see that sensitivity

1093 dimensions \mathbf{v}_1 and \mathbf{v}_2 , which cause p_P to increase and a regime dependent change in p_A , both point
1094 in the direction of increasing λ_{side} and decreasing λ_{task} . These eigenvalue changes are evident in
1095 the simulations of connectivity perturbations away from the modes (Fig. 13). As the component
1096 of connectivity along \mathbf{v}_1 and \mathbf{v}_2 becomes stronger (left-to-right), there is less separation between
1097 Pro an Anti populations (lower λ_{task}) and greater separation between Left and Right populations
1098 following stimulus presentation (greater λ_{side}). A key differentiating factor is that \mathbf{v}_1 substantially
1099 increases λ_{diag} , while \mathbf{v}_2 does not.

1100 During optogenetic silencing simulations, activations $x_\alpha(t)$ were set to a fraction of their values ($1 -$
1101 γ), where γ is the optogenetic perturbation strength. We found that λ_{all} and λ_{diag} were significantly
1102 anticorrelated with Anti error during delay period inactivation. Delay period inactivation was from
1103 $0.8 < t < 1.2$, choice period inactivation was for $t > 1.2$ and total inactivation was for the entire
1104 trial.

1105 5.2.5 Rank-2 RNN

1106 Traditional approaches to likelihood-free inference – approximate Bayesian computation (ABC)
1107 methods – randomly sample parameters \mathbf{z} until a suitable set is obtained. State-of-the-art ABC
1108 methods leverage sequential Monte Carlo (SMC) sampling techniques to obtain parameter sets more
1109 efficiently. To obtain more parameter samples, SMC-ABC must be run from scratch again. ABC
1110 methods do not confer log probabilities of samples. Like EPI, sequential neural posterior estimation
1111 (SNPE) uses deep learning to produce flexible posterior approximations. Like traditional Bayesian
1112 inference methods, SNPE conditions directly on the statistics of data. This differs from EPI, where
1113 posteriors are conditioned on emergent properties (moment constraints on the posterior predictive
1114 distribution). Peculiarities of SNPE (density estimation approach, two deep networks) make scaling
1115 in \mathbf{z} prohibitive.

1116 SMC-ABC has many hyperparameters, of which pyABC selects automatically by running some ini-
1117 tial diagnostics upon initialization. In concurrence with the literature, SMC-ABC fails to converge
1118 around 25-30 dimensions, since it's proposal samples never get close enough to the target statis-
1119 tics. We searched over many SNPE hyperparameter choices: $n_{\text{train}} \in [2,000, 10,000, 100,000]$ is the
1120 number of simulations run per training epoch, and $n_{\text{mades}} \in [2, 3]$ is the number of masked autore-
1121 gressive density estimators in the deep parameter distribution architecture. The greater n_{train} , the
1122 longer each epoch will take, but the more likely SNPE may converge during that epoch. Greater
1123 n_{mades} increases the flexibility of the deep parameter distribution of SNPE, but slows optimization.

₁₁₂₄ For the timing plot, we show the fastest among all of these choices, and for the convergence plot,
₁₁₂₅ we show the best convergence among all of these choices. During optimization, we used $n_{\text{atom}}=100$
₁₁₂₆ atomic proposals as is recommended.