Response to reviewers

(original comments in bold)

Bittner and colleagues introduce a machine learning framework for maximum entropy inference of model parameter distributions that are consistent with certain emergent model properties, specified by the investigator. The approach is illustrated on several models of potential interest.

Reviewers were broadly enthusiastic about the potential usefulness of this methodology. However, the reviews and ensuing discussion revealed several points of concern about the manuscript and the approach. The full reviewer comments are included below.

We thank the reviewers for the excellent and constructive feedback. These valid criticisms of the original manuscript impressed upon us the importance of making key improvements to this research project. We spent more than half of a year advancing this work according to the reviewer requests and suggestions, and we are excited to present them to you in a dramatically improved revision.

We have made significant changes to our writing and explanation of emergent property inference (EPI), done extensive comparisons of EPI to other parameter inference methods, and have seriously increased the quality and depth of our scientific analyses, which now yield strong theoretical results. Our model analyses now focus on analyzing the rich structure of parameter distributions, which the deep probability distributions of EPI make possible. Here, we provide a list of key manuscript updates, and below we explain how these changes address each of the reviewer's concerns.

List of key manuscript improvements:

- We completely overhauled our presentation of the parameter inference technique emergent property inference (EPI). See response to main concern #1.
- We added an entirely new results section to compare EPI to alternative parameter inference techniques (SMC-ABC and SNPE) as parameter count increases, which demonstrates the advantages of EPI in high dimensional parameter spaces. See response to main concern #2.
- EPI is used to understand how noise across neuron type populations governs excitatory variability in a model of primary visual cortex. Here, EPI yields visually striking, novel insight where the conventional analytic approach became infeasible with increasing neuron types. See response to main concern #4.
- EPI is used to discover multiple parametric regimes of rapid task switching in a model of superior colliculus. These regimes are efficiently identified and mechanistically characterized using the probabilistic modeling tools afforded by EPI, and we relate the inferred connectivities to results from optogenetic silencing experiments in rats. See response to main concern #4.

The main concerns are summarized as follows:

1. The methodology is not adequately explained. Both the body text and methods section present a somewhat selective description that is very hard to follow in places and should be checked and rewritten for clarity, completeness, notational consistency and correctness.

We thank the reviewers for pointing out that our explanation of the method was too narrowly focused and hard to follow. In the introduction, we present our method by focusing on the key aspect of EPI that differentiates it from other approaches to inverse problems. Specifically, we explain how current methodology infers the parameters producing computation by conditioning on *exemplar datasets* (real or simulated), whereas in EPI we condition directly on the *emergent properties* that define the computation.

> Lines 45-56
> "Statistical inference, of course, requires quantification of the vague term *computation*. In neuroscience, two perspectives are dominant. First, often we directly use an *exemplar dataset*: a collection of samples that express the computation of interest, this data being gathered either experimentally in the lab or from a computer simulation. While in some sense the best choice given its connection to experiment [15] , some drawbacks exist: these data are well known to have features irrelevant to the computation of interest [16, 17, 18], confounding inferences made on such data. Related to this point, use of a conventional dataset encourages conventional data likelihoods or loss functions, which focus on some global metric like squared error or evidence, rather than the computation itself. Alternatively, researchers often quantify an *emergent property*: a statistic of data that directly quantifies the computation of interest, wherein the dataset is only implicit. While such a choice may seem esoteric, it is not: the above "gold standard" examples all quantify and focus on some derived feature of the data, rather than the data drawn from the model."

We now frame the method within the more general context of parameter inference techniques in neuroscience, rather than the context of recent advancements in machine learning. We have made concentrated efforts to simplify language and to relate to all relevant existing methodology.

> Lines 57-67
> "An emergent property (EP) is of course a dataset by another name, but it suggests different approach to solving the same inverse problem: here we directly specify the desired emergent property – a statistic of datasets drawn from the model – and the value we wish that property to have, and we set up an optimization program to find the distribution of parameters that produce this computation. This statistical framework is not new: it is intimately connected to the literature on approximate bayesian computation [19, 20, 21], parameter sensitivity analyses [22, 23], maximum entropy modeling [24, 25, 26], and approximate bayesian inference [27, 28]; we detail

these connections in Section 5.1.1. However, the adaptation of these techniques to the problem of theoretical circuit analysis requires recent developments in deep learning for constrained optimization [29], and architectural choices for scalable, flexible generative modeling [30, 31]. We detail our method, which we call emergent property inference (EPI) in Section 3.2."

In Section 3.2 and 5.1, we precisely explain the details of EPI, and in Section 5.1.4 we show how it is optimized. To improve clarity, we have changed the notation and presentation of emergent properties. Emergent properties are now denoted with $\mathcal{X}$ to signify an implicit dataset predicated by the emergent property. And rather than a vector of first- and second-moment constraints, we present the emergent property more readably as mean and variance constraints on emergent property statistics:

Line 705
“

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

"

In our revised writing, we deemphasize the role of maximum entropy in the EPI algorithm, because this has largely served as a distraction in our experience. We show in Section 5.1.6 that EPI is a special case of variational inference, since maximum entropy is the same normative selection principle as variational bayesian methods. Therefore, it makes sense to present EPI in the main text as a statistical inference technique that constrains the predictions of the inferred parameters to be an emergent property, and leave the details of the maximum entropy to the technically proficient in Section 5.1.

Lines 152-154
“Many distributions in $\mathcal{Q}$ will respect the emergent property constraints, so we select the most random (highest entropy) distribution, which is the same choice made in variational bayesian methods (see Section 5.1.6)."

Figure 1 has been largely redone to reflect the modified presentation, and improve the pictorial representation of the method. In Section 5.3, we now show that EPI is the only simulation-based inference method that controls the predictions of its inferred distribution (Fig. 2C-D).

Finally, we emphasize the utility of this deep inference technique for scientific inquiry in a new paragraph at the end of Section 3.2.

Lines 159-174
“EPI produces parameter distributions that can be queried for scientific insight. The modes of $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ indicate parameter choices emblematic of the emergent property

(Fig. 1E yellow star). As probability in the EPI distribution decreases, the emergent property deteriorates. Perturbing $\mathbf{z}$ along a dimension in which $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ does not change will not disturb the emergent property, making this parameter combination *robust* with respect to the emergent property. In contrast, if $\mathbf{z}$ is perturbed along a dimension that strongly decreases $q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$, that parameter combination is deemed *sensitive*. By querying the second order derivative (Hessian) of $\log q_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathcal{X})$ at a mode, we can quantitatively identify how sensitive (or robust) each eigenvector is by its eigenvalue; the more negative, the more sensitive and the closer to zero, the more robust (see Section 5.2.4). Indeed, samples equidistant from the mode along these EPI-identified dimensions of sensitivity ($\mathbf{v}_1$, smaller eigenvalue) and robustness ($\mathbf{v}_2$, greater eigenvalue) (Fig. 1E, arrows) agree with error contours (Fig. 1E contours) and have diminished or preserved hub frequency, respectively (Fig. 1F activity traces). This suggests that changes in conductance along the parameter combination described by $\mathbf{v}_2$ will most preserve hub neuron firing between the intrinsic rates of the pyloric and gastric mill rhythms. Once an EPI distribution has been inferred, this Hessian calculation requires trivial computation (see Section 5.1.2)."

2. The computational resources required to use this method are not adequately benchmarked. For example, the cosubmission (Macke et al) reported wall clock time, required hardware and iterations required to produce results by directly comparing to existing methods (approximate bayesian computation, naive sampling, etc.) Without transparent benchmarks it is not possible to assess the advance offered by this method.

We thank the reviewers for emphasizing the importance of this methodological comparison. In Section 3.3, we provide a direct comparison of EPI to alternative simulation-based inference techniques SMC-ABC and SNPE by inferring RNN connectivities that exhibit stable amplification. These comparisons evaluate both wall time (Fig. 2A) and simulation count (Fig. 2B), and we explain how each algorithm was run in its preferred hardware setting in Section 5.3.4.

In this analysis, we demonstrate the improved scalability of deep inference techniques (EPI and SNPE) with respect to the state-of-the-art approximate bayesian computation technique (SMC-ABC). While controlling for architecture size (Fig. RNN1), we push the limits of SNPE through targeted hyperparameter modifications (Fig. RNN3), and show that EPI scales to higher dimensional RNN connectivities producing stable amplification than SNPE. Furthermore, we emphasize that SNPE does not constrain the properties of the inferred parameters; many connectivities inferred by SNPE result in unstable or nonamplified models (Fig. 2C, Fig. RNN2).

3. The extent to which this method is generally/straightforwardly applicable was in doubt. It seemed as though a significant amount of computation was required to do inference on one specified property and that the computation would need to be run afresh to query a new property. The methodology in the cosubmission (Macke) made clear that computation required for successive inferences is 'amortized' during training on random parameters. Moreover, EPI seemed less flexible than

the cosubmission's approach in that it required a differentiable loss function. The complementarity and advantages of this approach as opposed to the cosubmission's approach are therefore unclear.

The reviewers are right to point out these characteristics of EPI: it does not amortize across emergent properties, and it requires differentiability of the emergent properties of the model. Indeed, SNPE is more suitable for inference with nondifferentiable mechanistic models and scientific problems requiring many inferred parameter distributions. However, these relative drawbacks of EPI with respect to SNPE can be considered choices made in a trade-off between simulation-based inference approaches.

Unlike SNPE, EPI leverages gradients of the emergent property throughout optimization, which may lead to better efficiency and scalability (Section 3.3). The emergent properties of many models in neuroscience are tractably differentiable, four of which we analyze in this manuscript ranging across levels of biological realism, network size, and computational function. This trade-off is explained at the end of Section 3.3:

> Lines 240-251
> "Since gradients of the emergent property statistics are necessary in EPI optimization, gradient tractability is a key criteria when determining the suitability of a simulation-based inference technique. Evidenced by this analysis, EPI is a clear choice for inferring high dimensional parameter distributions when the emergent property gradient is efficiently calculated. This can be invaluable for understanding how RNNs produce complex computations. Even with a high degree of biophysical realism and expensive emergent property gradients, EPI was run successfully on intermediate hub frequency in a 5-neuron subcircuit model of the STG (Section 3.1). However, conditioning on the pyloric rhythm [50] in a model of the pyloric subnetwork model [12] proved to be prohibitive with EPI. The pyloric subnetwork requires many time steps for simulation and many key emergent property statistics (e.g. burst duration and phase gap) are not calculable or easily approximated with differentiable functions. In such cases, SNPE, which does not require differentiability of the emergent property has proved to be a powerful approach [28]."

Furthermore, EPI focuses the entire expressivity of the approximating deep probability distribution on a single distribution, rather than spreading this expressivity to some uncharacterized degree across the chosen training distribution of amortized posteriors in SNPE (Section 5.1.1 Related Approaches).

> Lines 788-793
> "The approximating fidelity of the deep probability distribution in sequential approaches is optimized to generalize across the training distribution of the conditioning variable. This generalization property of the sequential methods can reduce the accuracy at the singular posterior of interest. Whereas in EPI, the entire expressivity

of the deep probability distribution is dedicated to learning a single distribution as well as possible. Amortization is not possible in EPI, since EPI learns an exponential family distribution parameterized by its mean (see Section 5.1.3)."

Finally, we emphasize that EPI does something fundamental that SNPE and other inference techniques cannot. EPI learns parameter distributions whose predictions are constrained to produce the emergent property. We show in Figure 2 and Supplementary Figure RNN2 that SNPE does not control the statistical properties of its predictions, resulting in the inference of many parameters that are not consistent with the desired emergent property.

Lines 229-236

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

4. Some examples lack depth in their treatment (see reviewer comments) and in some cases the presentation is somewhat misleading. The STG example is not in fact a model of the STG. The cosubmission (Macke) uses a model close to the original Prinz et al model, which is a model of the pyloric subnetwork. It would be instructive to benchmark against this same model, including computation time/resources required. Secondly, the subsequent example (input-responsivity in a nonlinear sensory system) appeared to imply that EPI permits 'generation' and testing of hypotheses in a way that other methods do not. All the method really does is estimate a joint distribution of feasible parameters in a specific model which is manually inspected to propose hypotheses. Any other method (including brute force sampling) could be used in a similar way, so any claim that this is an integral advantage of EPI would be spurious. Indeed, one reviewer was confused about the origin of these hypotheses. While it is helpful to illustrate how EPI (and other methods) could be used, the writing needs to be far clearer in general and should clarify that EPI does not offer any new specific means of generating or evaluating hypotheses.

We thank the reviewers for explaining how they found some of the presentation misleading. We have taken serious care in this manuscript to clarify a.) what is novel, appreciable scientific insight provided by EPI, as well as b.) which scientific analyses are made possible by EPI.

(a) In the revised manuscript we clarify that novel theoretical insights are not being made into the STG subcircuit model or the recurrent neural network models. The STG subcircuit serves as a motivational example to explain how EPI works, and we use RNNs exhibiting stable amplification as a substrate for scalability analyses.

> Lines 69-74
> "First, we show EPI's ability to handle biologically realistic circuit models using a five-neuron model of the stomatogastric ganglion [32]: a neural circuit whose parametric degeneracy is closely studied [33]. Then, we show EPI's scalability to high dimensional parameter distributions by inferring connectivities of recurrent neural networks (RNNs) that exhibit stable, yet amplified responses – a hallmark of neural responses throughout the brain [34, 35, 36]."

We do produce strong theoretical insights into a model of primary visual cortex (Section 3.4) and superior colliculus (Section 3.5). These analyses have substantially more depth than the previous manuscript.

> Lines 74-81
> "In a model of primary visual cortex [37, 38], EPI reveals the how recurrent processing across different neuron-type populations shapes excitatory variability: a finding that we show is analytically intractable. Finally, we investigated the possible connectivities of superior colliculus that allow execution of different tasks on interleaved trials [39]. EPI discovered a rich distribution containing two connectivity regimes with different solution classes. We queried the deep probability distribution learned by EPI to produce a mechanistic understanding of cortical responses in each regime. Intriguingly, the inferred connectivities of each regime reproduced results from optogenetic inactivation experiments in markedly different ways."

(b) The ability to infer a flexible approximation to a probability distribution constrained to produce an emergent property is novel in its own right (Figure 2). The deep probability distribution fit by EPI facilitates the mode identification (via gradient ascent of the parameter log probability) and sensitivity measurement (via the measurement of the eigenvector of the Hessian at a parameter value). These mode identifications and sensitivity measurements are done in Sections 3.1 (Fig. 1E), 3.4 (Fig. 3E), and 3.5 (Fig. 4C). By using this mode identification technique along the ridges of high parameter probability in the SC model, we identify the parameters transitioning between the two regimes. Finally, the sensitivity dimensions of the SC model identified by EPI facilitated regime characterization through perturbation analyses (Fig. 4D, Fig. 5C).

Importantly, we do not claim that these theoretical insights were necessarily dependent on using the techniques in b.). One could have come to these conclusions via various combinations of techniques mentioned in Section 5.1.1 Related Methods. In the case of the V1 model inference,

the main point is to indicate that such insight can be afforded by EPI and its related methods, in contrast to the analytic derivations emblematic of practice in theoretical neuroscience.

> Lines 302-310
> "EPI revealed the quadratic dependence of excitatory variability on input variability to the E- and P-populations, as well as its independence to input from the other two inhibitory populations. In a simplified model ($\tau = \tau_{\text{noise}}$), it can be shown that surfaces of equal variance are ellipsoids as a function of $\boldsymbol{\sigma}$. Nevertheless, the sensitive and degenerate parameters are challenging to predict mathematically, since the covariance matrix depends on the steady-state solution of the network [53, 61], and terms in the covariance expression increase quadratically with each additional neuron-type population (see also Section 5.4.4). This emphasizes the value of streamlined methods for gaining understanding about theoretical models when mathematical analysis becomes onerous or impractical."

In the case of the SC model inference, random sampling would have taken prohibitively long, and it is unclear how the continuum between the two connectivity regimes would have been identified with alternative techniques:

> Lines 414-419
> "The probabilistic tools afforded by EPI yielded this insight: we identified two regimes and the continuum of connectivities between them by taking gradients of parameter probabilities in the EPI distribution, we identified sensitivity dimensions by measuring the Hessian of the EPI distribution, and we obtained many parameter samples at each step along the continuum (in 7.36 seconds with the EPI distribution rather than 4.2 days with brute force methods, see Section 5.5)."

As the reviewers indicate, the STG model analyzed in our manuscript is not that of Prinz et al. 2004, and thus not the model analyzed by the cosubmission. We found the Prinz et al. model prohibitive to infer with EPI, since the gradients of spiking frequency from its simulation are quite computationally intensive. However, we found it critical to show that it's very possible to do inference in such biophysically realistic Morris-Lecar models when gradients are tractable, which is the case of the 5-neuron STG model we analyzed from Gutierrez et al. 2013. This 5-neuron model represents the IC neuron (hub) and its coupling to the pyloric (fast) or gastric mill (slow) subcircuit rhythms. In the introductory text, we refer to this model as the "STG subcircuit" model (rather than "STG model"), and we better clarify what aspect of the STG is being modeled in Results Section 3.1.

> Lines 92-97
> "A subcircuit model of the STG [32] is shown schematically in Figure 1A. The fast population (f1 and f2) represents the subnetwork generating the pyloric rhythm and the slow population (s1 and s2) represents the subnetwork of the gastric mill rhythm.

The two fast neurons mutually inhibit one another, and spike at a greater frequency than the mutually inhibiting slow neurons. The hub neuron couples with either the fast or slow population, or both depending on modulatory conditions."

The difference between this model and the STG model of the pyloric subnetwork is emphasized in Section 3.3:

Lines 246-250
"However, conditioning on the pyloric rhythm [50] in a model of the pyloric subnetwork model [12] proved to be prohibitive with EPI. The pyloric subnetwork requires many time steps for simulation and many key emergent property statistics (e.g. burst duration and phase gap) are not calculable or easily approximated with differentiable functions."

5. There is a substantial literature on parameter sensitivity analysis and inference in systems biology, applied dynamical systems and control that has been neglected in this manuscript. The manuscript needs to acknowledge, draw parallels and explain distinctions from current methods (ABC, profile likelihood, deep learning approaches, gaussian processes, etc). The under-referencing of this literature deepened concerns about whether this approach represented an advance. DOIs for a small subset of potentially relevant papers include:

https://doi.org/10.1038/nprot.2014.025
http://doi.org/10.1085/jgp.201311116
http://doi.org/10.1016/j.tcs.2008.07.005
http://doi.org/10.3182/20120711-3-BE-2027.00381
http://doi.org/10.1093/bioinformatics/btm382
http://doi.org/10.1111/j.1467-9868.2010.00765.x
http://doi.org/10.1098/rsfs.2011.0051
https://doi.org/10.1098/rsfs.2011.0047
http://doi.org/10.1214/16-BA1017
https://doi.org/10.1039/C0MB00107D

Thank you for pointing us to these references on sensitivity analyses and applied dynamical systems. We have incorporated many of them into the current manuscript and explain EPI's relation to each class of techniques in Section 5.1.1 Related approaches.

Lines 809-822
"Structural identifiability analysis involves the measurement of sensitivity and unidentifiabilities in scientific models. Around a single parameter choice, one can measure the Jacobian. One approach for this calculation that scales well is EAR [64]. A popular efficient approach for systems of ODEs has been neural ODE adjoint [79] and its

stochastic adaptation [80]. Casting identifiability as a statistical estimation problem, the profile likelihood works via iterated optimization while holding parameters fixed [22]. An exciting recent method is capable of recovering the functional form of such unidentifiabilities away from a point by following degenerate dimensions of the fisher information matrix [23]. Global structural non-identifiabilities can be found for models with polynomial or rational dynamics equations using DAISY [63]. With EPI, we have all the benefits given by a statistical inference method plus the ability to query the first- or second-order gradient of the probability of the inferred distribution at any chosen parameter value. The second-order gradient of the log probability (the Hessian), which is directly afforded by EPI distributions, produces quantified information about parametric sensitivity of the emergent property in parameter space (see Section 3.2)."

6. One of the reviewers expressed concern that the work might have had significant input from a senior colleague during its early stages, and that that it might be worth discussing with the senior colleague whether their contribution was worthy of authorship. The authors may take this concern into account in revising the manuscript.

We have reached out to Woods Hole Course Project mentors where this work was discussed and explored in its early stages. James Fitzgerald and Dhruva Raman are happy with their acknowledgement, and do not insist on deeper involvement and authorship on this paper. Stephen Baccus has requested that we acknowledge the summer course, which we have agreed to do.

7. Finally, please also address specific points raised by the reviewers, included below.

Reviewer #1:

General Assessment: The authors introduce a machine learning framework for maximum entropy inference of model parameters which are consistent with certain emergent properties, which they call 'emergent property inference'. I think this is an interesting and direction, and this looks like a useful step towards this program. I think the paper could be improved with a more thorough discussion both of the broad principles their black box approach seeks to optimize, as well as the details of its implementation. I also think the detailed examples should be more self-contained. Finally I find this work to be somewhat misrepresented as a key to all of theoretical neuroscience. This approach may have some things to offer to the interesting problem of finding parameter regions of models, but this is not the entirety of, nor really a major part of theoretical neuroscience as I see it.

We thank the reviewer for their positive comments and thoughtful feedback. We have made serious effort to improve our presentation and explanation of EPI (see response to main concern #1). Furthermore, we have focused on clearly motivating and describing each neural circuit

model studied in this manuscript. Sufficient mathematical detail is written in each results section, while full details are presented in Methods. All code for training EPI on these models and their analysis are available in well-documented scripts and notebooks in our github repository.

We modify our writing to clarify that EPI is not a key to all of theoretical neuroscience, but rather a powerful solution to inverse problems in neural circuit modeling. Inverse problems are indeed a major part of theoretical neuroscience. Their solution is necessary for the scientific evaluation of mathematical representations of neural circuits, and ultimately for progress in the field. We clarify this point in the introduction:

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

Other concerns:

(1) Maximizing the entropy of the distribution is not a reparameterization invariant exercise. That is, results depend on whether the model parameters contains rates or time constants, for example. I wonder if this approach is attempting to use a 'flat prior' in some sense which has the same reparameterization issue? Can the authors comment?

The reviewer is correct to point out that maximum entropy solutions are not reparameterization invariant, and indeed the units matter. The reviewer's suggestion that the method is in some sense using a flat prior is also correct. To clarify, EPI does not execute posterior inference, because there is no empirical dataset or specified prior belief in EPI framework. However, we derive the relation of EPI to bayesian inference in Section 5.1.6, which shows EPI uses a uniform prior when framed as variational inference.

> Lines 986-989
> "Thus, EPI is implicitly executing variational inference with a uniform prior and a mean-field gaussian likelihood on the emergent property statistics. The mean and variances of the mean-field gaussian likelihood are predicated by $\boldsymbol{\eta}_{\text{opt}}$ (Equations 36 and 38), which is adapted after each EPI optimization epoch based on $\mathcal{X}$ (see Section 5.1.4)."

In our examples, we only infer distributions of parameters with the same units, so issue should not draw concern over the validity of our model analyses. As suggested, the EPI solution will differ according to relative scaling of parameter values under the maximum entropy selection principle. Thus, an important clarification is that sensitivity quantifications are made in the context of the chosen parameter scalings. A final concern here is numerical: if the inferred distribution is in a thin region of parameter space, that is not well represented by the precision of the numerical format, there can be issues in optimization. It makes most sense to make sure EPI is exploring through parameter space through an industry-recommended range of values, and that sensitivity measurements are interpretable.

SB: Should we add text about this in the paper?

(2) I don't think this is a criticism of the work, but instead of the writing about it: I find the introductory paragraphs to give a rather limited overview of theory as finding parameters of models which contain the right phenomenology.

We appreciate this feedback. We have adapted the introduction to clarify that we are focusing on solving inverse problems in theoretical neuroscience.

(3) I am somewhat familiar with the stomatogastric circuit model, and so that is where I think I understand what they have done best. I don't understand what I should take away from their paper with regards to this model. Are there any findings that hadn't been appreciated before? What does this method tell us about the system and or its model?

We clarify in point 4 above that we are not claiming to produce novel, appreciable scientific insight about the STG subcircuit model, which is used as a motivation example. The takeaway is that the conductance parameters producing intermediate hub frequency belong to a complex 2-D distribution, which EPI captures accurately, and that EPI can tell us the parameter changes away from the prototypical configuration that change hub frequency the most or least. For example, for increases in g_{el} and g_{synA} according to the proportions of the degenerate dimension of parameter space, intermediate hub neuron frequency will be preserved in this model. This suggests that in the real STG neural circuit, the IC neuron will remain at an intermediate frequency between the pyloric and gastric mill rhythms if parameter changes are made along such a dimension.

> Lines 171 -173
> "This suggests that changes in conductance along the parameter combination described by $\mathbf{v}_2$ will most preserve hub neuron firing between the intrinsic rates of the pyloric and gastric mill rhythms."

(4) I don't follow the other examples. Ideally more details should be given so that readers like myself who don't already know these systems can understand what's been done.

Thank you for the feedback. We have taken care to give more general context, and motivation for each neural circuit model.

(5) In figure 2C, the difference between the confidence interval between linear and nonlinear predictions is huge! How much of this is due to nonlinearities, and how much is due to differences in the way these models are being evaluated?

In the current manuscript, we do not examine the difference between linear and nonlinear predictions of the V1 model.

Reviewer #2:

General assessment

This is a very interesting approach to an extremely important question in theoretical neuroscience, and the mathematics and algorithms appear to be very rigorous. The complexities in using this in practice make me wonder if it will find wide application though: setting up the objective to be differentiable, tweaking hyperparameters for training, and interpreting the results; all seem to require a lot from the user. On the other hand, the authors are to be congratulated on providing high quality open source code including clear tutorials on how to use it.

Major points

1. Training deep networks is hard. Indeed the authors devote a substantial amount of the manuscript to techniques for training them, and note that different hyperparameters were necessary for each of the different studies. Can the authors be confident that they have found the network which gives maximum entropy or close to it? If so, how. If not, how does that affect the conclusions?

We thank the reviewer for their positive comments. To draw a parallel, training deep networks for visual processing used to be considered infeasible, but became easier through iterative improvements in architectural and hyperparameter choices that spread across the field. Similarly, training deep networks via EPI to learn parameter distributions became easier throughout this research project as we learned through trial and error what works well. In fact, there has been extraordinary progress in the field of deep probability distributions (specifically normalizing flows), that have allowed EPI to converge while capturing complex structure much more regularly (e.g. Dinh et al. 2017 and Kingma et al. 2018). This manuscript has much value in its explanation of hyperparameter choices, and the extensive set of examples in the online code github. Every figure of this paper is reproducible with the jupyter notebooks, and there are several tutorials for understanding the most consequential hyperparameters: augmented Lagrangian constant, normalizing flow architecture.

In general, we cannot know if we have arrived at the global maximum entropy distribution for a given emergent property. The reviewer is correct to point out that multiple distributions may

satisfy the emergent property and have different levels of entropy. In the new manuscript, we present the method as an inference technique without focusing very greatly on maximum entropy, since it tends to distract and confuse the reader. We derive an analytic equivalence to variational inference (Section 5.1.6) showing that a.) EPI is a valid inference method, and b.) to emphasize that maximum entropy is the normative selection principle of bayesian inference methods in general. Thus, the concern of not having the globally optimal inferred distribution is the same that applies to all other inference techniques.

Practically, this has scientific implications. It means that we may be missing important structure in the inferred distribution, or we may be missing additional modes in parameter space. To handle this methodologically, we run EPI with multiple random seeds, and select the distribution that has converged with the greatest entropy for scientific analysis. Throughout the manuscript, we compare to analytic, error contour, and brute-force ground truth to ensure we are capturing the correct distribution with EPI (see response to R3 concern 1).

2. Interpreting the results still seems to require quite a lot of work. For example, from inspecting Fig 2 the authors extract four hypotheses. Why these four? Are there other hypotheses that could be extracted and if not how do we know there aren't? Could something systematic be said here?

This analysis is no longer in the manuscript.

3. Scalability. The authors state that the method should in principle be scalable, but does that apply to interpreting the results? For example, for the V1 model it seems that you need to look at 48 figures for 4 variables, and I believe this would scale as $O(n^2)$ with n variables. This seems to require an unsustainable amount of manual work?

We refer the reviewer to Figure 2 and Section 3.3 for scalability analysis. The scaling analysis addresses the question of the issue of parameter discovery with EPI in high-dimensional parameter spaces.

Another important question the reviewer brings up is how well one can analyze the high-dimensional parameter distributions that EPI produces? Indeed, these distributions become more challenging to understand and visualize in high dimensions. This is where the sensitivity measurements appearing in sections 3.1, 3.4, and 3.5 can be particularly useful. Even in high dimensions, trained deep probability distributions offer tractable quantitative assessments of how parametric combinations affect the emergent property that was conditioned upon.

4. There are some very particular choices made in the applications and I wonder how general the conclusions are as a consequence. For example, in equation (5) the authors choose an arbitrary amount of variance 0.01^2 - why? In the same example, why look at y=0.1 and 0.5?

In the current manuscript, we make sure to explain all choices of the emergent property constraints Here, we show the description of each emergent property with equations omitted.

> Section 3.2, Lines 136-140
> "We stipulate that the hub neuron's spiking frequency – denoted by statistic $\omega_{\text{hub}}(\mathbf{x})$ – is close to a frequency of 0.55Hz, between that of the slow and fast frequencies. Mathematically, we define this emergent property with two constraints: that the mean hub frequency is 0.55Hz, and that the variance of the hub frequency is moderate."

> Section 3.3, Lines 200-207
> "Two conditions are necessary and sufficient for RNNs to exhibit stable amplification [44]: real$(\lambda_1) < 1$ and $\lambda_1^s > 1$ ... EPI can naturally condition on this emergent property under the notion that variance constraints with standard deviation 0.25 predicate that the vast majority of samples (those within two standard deviations) are within the specified ranges."

> Section 3.4, Lines 281-284
> "We quantify levels y of E-population variability with the emergent property where $s_E(\mathbf{x}; \mathbf{z})$ is the standard deviation of the stochastic E-population response about its steady state (Fig. 3C). In the following analyses, we compare levels of 5Hz and 10Hz, and select 1 Hz^2 variance such that the two emergent properties do not overlap in $s_E(\mathbf{z}; \mathbf{x})$."

> Section 3.5, Lines 336-339
> "We stipulate that accuracy be on average .75 in each task with variance $.075^2$. 75% accuracy is a realistic level of performance in each task, and with the chosen variance, inferred models will not exhibit fully random responses (50%), nor perfect performance (100%)."

Minor points

The introduction and discussion are very clearly written but the results section is hard going. Partly this is unavoidable given the subject matter, but a few sentences here and there might help the reader along. Things like "x is the internal state of the model, z are the parameters we will change, ...". When introducing entropy in equation (3), H isn't previously defined, and again it might help to give the reader a hand here, e.g. max entropy means the distribution is as spread out as possible"(you can surely find a better thing to say than this, but just to give an idea). The other point which is quite hard to follow is interpreting e.g. Fig 2C. Perhaps for Hypothesis 1 you could write a couple of sentences explaining slightly more clearly why seeing small blobs or horizontal/vertical lines in these distribution plots means that it's mainly determined by the direct input.

Thank you for the detailed suggestions on how to improve writing in the results sections. We have taken the specific suggestion of explicitly calling out **x** as network state, and we have moved complicated discussion of the role of entropy to the Methods section.

Reviewer #3:

This paper addresses a major issue in fitting neuroscience models: how to identify the often degenerate, or nearly degenerate, set of parameters that can underlie a set of experimental observations. Whereas previous techniques often depended upon brute force explorations or special parametric forms or local linearizations to generate sets of parameters consistent with measured properties, the authors take advantage of deep generative networks to address this problem. Overall, I think this paper and the complementary submission have the potential to be transformative contributions to model fitting efforts in neuroscience. That being said, since the primary contribution is the methodology, I think the paper requires more systematic comparisons to ground truth examples to demonstrate potential strengths and weaknesses, and more focus on methodology rather than applications.

We thank the reviewer for their positive comments and thoughtful feedback. We have made great efforts to provide several clear comparisons to ground truth (see response to concern #1), and now provide an extensive methodological comparison to SMC-ABC and SNPE (see response to Main concern #2).

Substantive concerns:
1) The authors only have a single ground-truth example where they compare to a known result (a 2x2 linear dynamical system). It would be good to show how well this method compares to results from, for example, a direct brute force grid search of a system with a strongly non-elliptical (e.g. sharply bent) shaped parameter regime and a reasonably large (e.g. 5?) number of parameters corresponding to a particular property, to see how well the derived probability distribution overlaps the brute force grid search parameters (perhaps shown via several 2-D projections).

We thank the reviewer for pointing out the importance of ground truth comparisons in this manuscript. In this revision, we make ground truth comparisons via analytic derivations, empirical error contours, and brute-force sampling.

Analytic comparisons: The 2x2 linear dynamical system is chosen as a worked example because it has multi-modal non-elliptical structure (Fig. 6), and its contours can be derived analytically (Fig. 7). Similarly, in Section 5.4.4, we derive the quadratic relationship between excitatory variability and input noise variability (in a simplified model) suggesting that the quadratic relationship uncovered by EPI (see Section 3.4) is correct.

Error contours: In the motivation example, we compare the EPI inferred distribution of STG conductances to hub frequency contours (Figure 1E), which show that the non-elliptical parametric structure captured by EPI is in agreement with these contours. This general region of parameter

space was labeled following grid search analyses in a previous study (Gutierrez et al. 2013, Figure 2, parameter regime G).

Brute-force: The EPI inferred distribution for rapid task switching in the SC model is sharply bent (Fig. 4), and matches the parameter set returned from random sampling (Figure 24A). We note that the brute-force parameter set is actually not the ground-truth solution, because it does not obey the constraints of the emergent property as the EPI distribution does (Figure 24B). This can explain the spurious samples in the brute-force set that are not in the EPI inferred distribution.

All EPI distributions shown in this manuscript are "validated" in the sense that they pass a hypothesis testing criteria for emergent property convergence; all EPI distributions produce their emergent properties. Finally, the underlying maximum entropy flow network (MEFN) algorithm is compared to a ground truth solution (Loaiza Ganem et al. 2017, Figure 2) by deriving ground truth from the duality of maximum entropy distributions and exponential families (see Section 5.1.3).

2) It was not obvious whether EPI actually scales well to higher dimensions and how much computation it would take (there is one claim that it 'should scale reasonably'). While I agree that examples with a small number of parameters is nice for illustration, a major issue is how to develop techniques that can handle large numbers of parameters (brute force, while inelegant, inefficient, and not producing an explicit probability distribution can do a reasonable job for small #'s of parameters). The authors should show some example of extending to larger number of parameters and do some checks to show that it appears to work. As a methodological contribution, the authors should also give some sense of how computationally intensive the method is and some sense of how it scales with size. This seems particularly relevant to, for example, trying to infer uncertainties in a large weight matrix or a non-parametric description of spatial or temporal responses or a sensory neuron (which I'm assuming this technique is not appropriate for? See point#4 below).

The reviewer is right to point out the importance of a scaling analysis. Please see response to Main concern #2.

3) For the STG-like example, this was done for a very simple model that was motivated by the STG but isn't based on experimental recordings. Most of the brute force models of the STG seek to fit various waveform properties of neurons and relative phases. Could the model handle these types of analyses, or would it run into problems due to either needing to specify too many properties or because properties like number of spikes per burstäre discrete rather than continuous? This isn't fatal, but would be good to consider and/or note explicitly.

The STG subcircuit model of Gutierrez et al. 2013 is certainly less complex than other models of the STG, yet 5 connected Morris-Lecar neurons is certainly a nontrivial system. We clarify why

this model is analyzed in Section 3.1 instead of more complex STG models when discussing the differences between EPI and SNPE:

> Lines 240-251
> "Since gradients of the emergent property statistics are necessary in EPI optimization, gradient tractability is a key criteria when determining the suitability of a simulation-based inference technique. Evidenced by this analysis, EPI is a clear choice for inferring high dimensional parameter distributions when the emergent property gradient is efficiently calculated. This can be invaluable for understanding how RNNs produce complex computations. Even with a high degree of biophysical realism and expensive emergent property gradients, EPI was run successfully on intermediate hub frequency in a 5-neuron subcircuit model of the STG (Section 3.1). However, conditioning on the pyloric rhythm [50] in a model of the pyloric subnetwork model [12] proved to be prohibitive with EPI. The pyloric subnetwork requires many time steps for simulation and many key emergent property statistics (e.g. burst duration and phase gap) are not calculable or easily approximated with differentiable functions. In such cases, SNPE, which does not require differentiability of the emergent property has proved to be a powerful approach [28]."

4) The discussion should be expanded to be more specific about what problems the authors think the model is, or is not, appropriate for. Comparisons to the Goncalves article would also be helpful since users will want to know the comparative advantages/disadvantages of each method. (if the authors could coordinate running their methods on a common illustrative example, that would be cool, but not required).

Thank you for this recommendation. We now include a substantial text in discussion devoted to this topic.

> Lines 431-454
> "Methodology for statistical inference in circuit models has evolved considerably in recent years. Early work used rejection sampling techniques [19, 20, 21], but more recently developed methodology employs deep learning to improve efficiency or provide deep, flexible distribution approximations. SNPE [28], developed along with EPI (see Section 5.1.1), has been used for posterior inference of parameters in circuit models conditioned upon exemplar data used to represent computation. Like SNPE, EPI is a deep inference technique, but it infers parameter distributions that only produce the computation of interest (see Section 3.3).
>
> Exemplary data versus emergent properties aside, EPI has better scaling properties than SNPE when emergent property gradients are tractable (Section 3.3). However, SNPE has its own relative advantages. SNPE is effective when circuit model simulations are lengthy or nondifferentiable. For example, SNPE can infer the STG parameters that produce the pyloric rhythm [28], while EPI cannot. Thus, while it is

nice to infer parameter distributions with constrained emergent properties with EPI, SNPE is most appropriate when emergent property gradients are intractable.

The scientific analyses of Sections 3.4 and 3.5 derived theoretical findings by querying the structure of the inferred distribution of EPI. By measuring the dimensions in which probability decreases fastest, the dimensions of *sensitivity*, we gain valuable understanding of how model parameters govern the emergent property. A rich literature on parameter sensitivity analyses in biological models presents several methods towards this scientific approach [63, 22, 64, 23]. The value offered by EPI (and other deep inference methods like SNPE), is that the once the flexible deep probability distributions are fit to the parameter distribution, this distribution's structure can be quantified at any parameter choice, offering instantly available sensitivity measurements. Together, the ability to condition upon emergent properties, the efficient inference algorithm, and the capacity for parameter sensitivity analyses make EPI a powerful new method for addressing inverse problems in theoretical neuroscience."

5) Given that the paper is heavily a (very valuable!) methods paper for a general audience, the method should be better explained both in the main text and the supplement. Some specific ones are below, but the authors should more generally send the paper to naïve readers to check what is/is not well explained. -Figure 1 is somewhat opaque and also has notational issues (e.g. omega is the frequency but also appears to be the random input sample). -For the general audience of eLife, panels C and D are not well described individually or well connected to each other and don't illustrate or describe all of the relevant variables (including what q0 is and what x is). -In equation 2 (and also in the same equation in the supplement), it was not immediately obvious what the expectation was taken over. -The authors don't specific the distribution of w (it's referred to only as 'a simple random variable', which is not clear). -It was also sometimes hard to quickly find in the text basic, important quantities like what z was for a given simulation. -The augmented Lagrangian optimization was not well explained or motivated. There is a reference to m=absolute value(mu) but I didn't see m in the above equation. -Using mu to describe a vector that includes means and variances is confusing notation since mu often denotes means -It would be helpful to have a pseudo-code 'Algorithm' figure or section of the text

Thank you for this detailed list of improvements to be made when describing the method. We have taken care to address each item in this list.

Minor Comments:

1) I'm not sure if the authors are referring to a particular constrained form of the Schur decomposition, but the general statement in the Figure caption that the Schur decomposition is unique is not true. Also, one does not need to refer to Schur

eigenvaluessince the diagonal elements of the Schur decomposition are the (usual) eigenvalues.

We do not use a Schur decomposition to analyze the SC model in the current manuscript.

2) p. 31: usually one reserves the variable omega for angular frequencies: omega = 2*pi*f where f is frequency.

We have removed the variable ω as it was unnecessary, and avoided introducing another variable f to be conflated with f's use as emergent property statistic.

3) Some references for other approaches and work that might be worth listing for scholarship: Sloppy models and information geometry (including MCMC approaches, e.g. Mannakkee, Ragsdale, Transtrum, Gutenkunst); higher dimensional sloppy models in neuroscience (O'Leary, Sutton, & Marder 2015, Fisher, Olasagasti, et al., 2013); Compensatory parameter combinations through the implicit function theorem (Olypher and Calabrese, J. Neurophys. 2007).

Thank you for these references, we have incorporated some where appropriate.

Additional data files and statistical comments:

Code should be made available in well-documented form if it isn't already.