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 very constructive feedback. To address these points of concern, we have made extensive improvements to the manuscript. We have made significant changes to our writing and explanation of the inference technique, made an extensive comparison to other parameter inference methods, and increased the quality and depth of our scientific analyses, which 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 significantly changed how we explain the parameter inference technique (EPI) to clarify the methodological contribution of this research. We now emphasize the novelty and scientific value of conditioning probability distributions on theoretical criteria of interest (emergent properties), in contrast to noisy observed neural datasets.
- We show EPI's superior scalability in parameter dimension to alternative likelihood-free inference techniques SMC-ABC and SNPE when inferring RNN connectivities exhibiting stable amplification.
- We show how derivations of excitatory variability in a model of primary visual cortex become analytically intractable as additional neuron types are added. We then use EPI to produce a clear insight that mathematical analysis did not produce.
- We discover multiple parametric regimes of computation in a model of superior colliculus from the fine distributional structure captured by EPI. We then use the deep probability distribution to mechanistically characterize each regime, and also find that both of the discovered regimes reproduce an important result from inactivation experiments.

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 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, relate to all relevant existing methodology (Section 5.1.1 Related approaches), and precisely explain the novelty and utility of this deep inference technique.

Primarily, we explain the incongruity between parameter inference techniques and the practice of theoretical neuroscience that motivates EPI: modern parameter inference techniques are designed to condition on noisy datasets rather than the mathematically defined criteria. This mismatch has consequences – primarily that data predicted from the inferred distribution are left unconstrained and often violate the desired conditioning criteria. We then explain how formalizing the mathematical criteria of interest into a statistical feature of the model – the emergent property – we enable deep inference of mechanistic model parameters with respect to this property.

In Section 3.1 and 3.2, we present EPI through a motivational example of the STG subcircuit. The figure has been updated to reflect the modified presentation, and we more clearly emphasize the scientific tools afforded once a distribution is fit with EPI.

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 likelihood-free inference techniques SMC-ABC and SNPE by inferring RNN connectivities that exhibit stable amplification. These comparisons evaluate both wall time and simulation count, and we explain how each algorithm was run in its ideal hardware setting.

In this analysis, we demonstrate the superior scalability of deep inference techniques (EPI and SNPE) to the state-of-the-art random sampling technique (SMC-ABC) . We then show that SNPE does not constrain the properties of inferred parameters, resulting in many

unstable and nonamplified models. Furthermore, we push the limits of SNPE through targeted hyperparameter modifications, and show that for this model EPI scales to a higher dimensional parameter space than SNPE.

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 likelihood-free inference approaches. Unlike SNPE, EPI leverages gradients of the model to improve efficiency (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. Furthermore, EPI focuses the entire expressivity of the approximating deep probability distribution on a single distribution, rather than spreading this expressivity to some degree across the chosen training distribution of amortized posteriors in SNPE (Section 5.1.1 Related Approaches).

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 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.

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. 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.

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, 3.4, and 3.5. 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 identified by the SC model facilitated regime characterization through perturbation analyses.

Importantly, we do not claim that these theoretical insights were necessary with the techniques in b.). One could have come to these conclusions via brute force techniques, although less efficiently. The point is that we have brought deep inference to bear on theoretical neuroscience in a way that is principled, efficient, and scientifically compelling.

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 clarify what is being modeled in Results Section 3.1.

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

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. We are confident that there are no authorship issues.

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 intere-

sting 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.

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.

In our examples, we only infer distributions of parameters with the same units. 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 a standard range of values, and that sensitivity measurements are interpretable.

(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.

(*Do this?) We have edited the introduction to more appropriately characterize the practice of theoretical neuroscience.

(3) I am somewhat familiar with the stomatognathic 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?

asdfasdfa

(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?

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?

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?

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?

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.

Things I will say

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.

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).

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).

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.

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).

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

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.

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

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).

Additional data files and statistical comments:

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