Thank you for submitting your Technical Report, "Interrogating theoretical models of neural computation with deep inference." Thank you also for your patience during the rather long review process, which was probably delayed due to the holidays.  
  
It has now been seen by 3 referees, whose comments are below. After such a delay, I am very sorry to say that they have raised serious concerns and do not find the paper suitable for Nature Neuroscience. We must therefore reject the paper and we will not be able to consider a revised version of the manuscript or an appeal on this decision.  
  
While the referees find your work of some interest, they raise concerns about the overall framing of the work, the overall impact and accessibility for a broad non-expert audience, and the lack of sufficient validation against other approaches, parameter choices, or enough experimental data. We feel that these reservations are sufficiently important as to preclude publication of this study in Nature Neuroscience.  
  
  
\*\*\*Although we cannot offer to publish your paper in Nature Neuroscience, I have discussed your manuscript and the reviewers’ comments with our colleagues at Nature Communications. They are willing to send a suitably revised version back to the reviewers if you transfer the revised manuscript to Nature Communications. Should you wish to have your revised paper considered by Nature Communications, please use the link to the Springer Nature manuscript transfer service in the footnote once the revision is ready, and include a point-by-point response to the reviewers’ concerns.  
  
Please note that the editor will require all reviewer concerns raised by the referees related to the lack of benchmarking against existing methods and more complex examples to be addressed in full before sending the revised paper back to the original referees.  
  
Your handling editor at Nature Communications would be Sachin Ranade ([sachin.ranade@us.nature.com](mailto:sachin.ranade@us.nature.com)). If there is anything you would like to discuss before transferring the paper and its reviews, please don’t hesitate to contact him by e-mail.\*\*\*  
  
  
Please note that Nature Communications is a fully open access journal. For information about article processing charges, open access funding, and advice and support from Springer Nature, please consult the Nature Communications Open Access page ([www.nature.com/ncomms/open\_access/index.html](http://www.nature.com/ncomms/open_access/index.html)).  
  
We are very sorry that we cannot be more positive on this occasion, and we thank you for the opportunity to consider this manuscript. We hope that you will find our referees' comments helpful and that you will soon receive a more encouraging response elsewhere.  
  
  
Sincerely,  
Jean  
  
Jean Mary Zarate, PhD  
Senior Editor  
Nature Neuroscience, New York  
ORCiD: 0000-0002-7418-2199  
  
  
Referee expertise:  
  
Referee #1: computational neuroscience  
  
Referee #2: theoretical neuroscience  
  
Referee #3: computational neuroscience, machine learning  
  
  
Reviewers' Comments:  
  
**Reviewer #1:**  
Remarks to the Author:  
In this paper the authors propose statistical methods called deep inference, based on machine learning tools like autoencoders, for determining parameters of neural network models. The setup is a seeming dichotomy: simplicity of models that neglects salient biological properties versus complexity of models that defies analytical calculations. The new approach is based on the idea of Bayesian inference in terms of probability distributions of model parameters rather than a specific parameter values. Examples are illustrated using central pattern generator in the stomatogastric ganglion, primary visual cortex, superior colliculus and a model for a simple mathematical task.  
  
Unfortunately, the described dichotomy is a false one. It is the task of theoretical neuroscience to identify general principles from mathematical models that do capture the essence of the biological brain, and there are a number of good examples including the proposal of the excitation-inhibition balance, theoretical prediction of a disinhibitory motif, etc. This work is mathematically sophisticated but the proposed method does not yield much that we already know, it is fair to say. Note that concerning parameter identifications, the real challenge for understanding complex models is to identify multiple parameter sets that are often not connected with each other in a high dimensional parameter space, not a continuous/smooth probability distribution of their values which is the focus of this paper. The paper is well written; yet with jargons and 87 equations it is doubtful that it is accessible to the majority of experimentalists. It seems that the paper is more appropriate for a  
specialized journal.  
  
  
  
**Reviewer #2:**  
Remarks to the Author:  
This manuscript presents a novel method for studying models of neural computations. Starting from the fact that most models are difficult to handle analytically, this manuscript proposes to focus on a pre-determined output of a model, and then use statistical techniques to determine the set of parameter values consistent with that output. To this end, the manuscript introduces a new set of methods based on deep networks that perform statistical inference. These methods are show-cased on a series of previously published models, specifically the stomatogastric ganglion model, an excitatory-inhibitory rate network, a model of flexible choice, and a randomly connected recurrent network.  
  
The manuscript addresses an important issue that is central to understanding neural computations. The deep inference methods developed in the manuscript seem highly sophisticated, and potentially very promising. Unfortunately the results obtained so far using these methods do not quite live up to the expectations raised in the introduction. In general, the results seem hard to interpret, and there appears to be a large mismatch between the complexity of the approach, and the insight it generates. Technically this is a very interesting paper, but it is not appropriate for the audience of Nature Neuroscience.  
  
Specific comments:  
  
1. The use of the expression "emergent property" throughout the text seems inappropriate. By "emergent property", the authors simply refer to moments of the model variables. These moments are defined in the full parameter space. In contrast, in standard usage, "emergent property" refers to a collective property that typically appears only in a subset of parameter space, often through a bifurcation or phase transition. An "emergent property" often does correspond to a non-vanishing moment of a distribution, but the converse is not true, any non-vanishing moment cannot be called an "emergent property". In this regard, the choice of the statistics used in different examples seems questionable (see below).  
  
  
2. Beyond semantics, understanding emergent properties usually means understanding qualitative behaviours of the system. In contrast, the approach developed here focuses on fixing specific values for various statistics of the system. How are these values to be chosen in general? This seems to be a key issue. To get an insight into emergent properties, one would need to vary the values of the statistics, e.g. understand in which part of the parameter space a specific moment is non-zero. This is not done in the manuscript, and it is not clear how the proposed approach really confers a qualitative understanding of model dynamics, let alone bifurcations/phase-transitions.  
  
3. It would be important to explicitly demonstrate that the proposed approach is in some sense more efficient than brute force grid search. In particular, some quantification of the computational cost as function of number of parameters should be provided.  
  
4. STG model:  
- network syncing is quantified via the spiking frequency of the neurons. This does not seem appropriate, the neurons can have a non-zero spiking frequency in absence of syncing in the network.  
- it it surprising that there is no mention of extensive parameter redundancy in this context. It seems this should be a major output of the method. Instead, Fig 1B displays a very clear mode in the distribution.  
  
5. Excitatory-inhibitory network:  
- l.210: "we are interested in the differential responses of each neuron-type population to changes in input dh": why is this the main quantity of interest  
- Eq.5: I was confused here, are the firing rates fixed for all populations except one?  
- l.254 "These hypotheses were in stark contrast to what was available to us via traditional analytical linear prediction (Fig. 2C, magenta)": the magenta points seems to be always at zero, I do not understand what it represents. While the linear approximation is not expected to be quantitatively accurate, it should capture the qualitative behavior, eg whether the firing rate increases or decreases when one of the inputs is varied. More specifically, the hypotheses H1-H3 should be captured by a linear approximation. Moreover in Fig 3, the output seems essentially threshold-linear. So altogether this statement seems puzzling. A better case needs to be made for the insight gained from the new method.  
- l.206 "The effective connectivity weights W were obtained from experimental recordings": It is misleading to suggest that all the parameters of the connectivity can be determined from experimental data. Some of the conclusions, in particular H4, presumably depend on the choice of the connectivity matrix. That point needs clarification.  
  
  
6. Model of flexible choice:  
- the choice of statistic seems also questionable here. To focus on the behavioral performance of the model, it seems that activity should be conditioned on the inputs. Is this possible within the developed approach?  
- I have a hard time understanding the significance of the obtained result.  
  
  
  
**Reviewer #3:**  
Remarks to the Author:  
The authors address a very important question for those of us building computational models of brain function, that is, how to construct and interpret models that contain multiple parameters under conditions where it is not possible to derive analytical solutions. The authors refer to three steps: model design, choosing a property of interest, and finally assessing how model parameters relate to the property of interest. It is in this last step that current approaches struggle with, because many efforts involve major analytical simplifications like linearization or ad hoc ways of characterizing how model parameters relate to the model output. Here the authors propose a deep generative modeling approach for emergent property inference to systematically find good parameters or parameter distributions constrained by the target model properties. What is particularly exciting and useful about this work is that the proposed method finds a distribution over the model parameters such  
that the model produces a certain emergent phenomenon, which is defined by some moments of the model output. This is done by searching for a distribution of parameters that will maximize entropy under a constraint forcing the average (over the model realizations and over the parameters distribution) of a function of the model output to be equal to some “ground truth” value.  
  
The paper is very interesting but it was difficult to read. The authors should do a much better job explaining what they are doing. There is a lot of nomenclature that is not really explained or defined. For example, what is theta in the definition of Q (line 162)? This is probably some sort of parameter of the deep network performing the probability transformation (a lot of people in deep learning use this type of nomenclature), but this is not clear enough. How are the authors choosing the family of probability distributions? In general, none of the figure legends define the variables shown there. Oftentimes variables are not defined in the text next to the place where they are used. Many steps and assumptions in the method are not justified or explained.  
  
One important example is H. What is H in equation 3? The text eventually indicates that this is a primal objective, and next suggests that this is the Shannon entropy. The entropy of what? From Equation 3 (but not in the text), they are trying to maximize the entropy of the distribution of q over parameters. It is not very clear why this is a sensible or useful thing to do. One could imagine different possible aspects of q to maximize and it is not clear how any of these multiple possible definitions relate to the problems under study. For example, why would maximizing Shannon entropy of q (assuming that this is what they are doing), lead to better results in terms of understanding parameters for a given computational model?  
The authors then state, but do not show, that the methodology is unaffected by this choice. What they probably mean is that in the EPI methodology, they could maximize other properties of q in equation 3. But the results probably depend a lot on their choice here. And it seems that this is the case given the parenthetical comment from the authors that the results of course depend on this definition. But isn’t this the whole point of the paper, to come up with useful ways of studying the distribution of parameters? It seems that the choice of H is critical to the whole enterprise, and there is essentially nothing said about this. One could choose lots of different ways of generating different parameters z (while satisfying the constraints in \mu), and it would seem that those choices could have a large impact in the results, or at least it is not intuitive why this would not be the case.  
  
The ideas discussed here could potentially be useful to others in the computational neuroscience community. For people to embrace this, there are several important comparisons to alternative methods that would be extremely useful. The authors compellingly show that under certain conditions (e.g. large inputs in Figure 2), linearization is not a good idea. However, what about other forms of choosing the parameters. For example, and as the simplest version, what about a brute force sweep of parameter space? Particularly for the type of toy problems with few parameters illustrated here, it seems that this would work equally well, perhaps even better because it is not tied to arbitrary definitions such as maximizing Shannon entropy. Comparison to other Bayesian inference methods would also be useful to help readers assess whether it is worth trying this method on their problems.  
  
Another consideration is the feasibility / efficiency of the method. There are no specs here about computational costs and efficiency considerations. The number of parameters and system dimensionalities studied here are pretty small. What about problems that have large numbers of parameters? Or cases which require simulating thousands of neurons? Perhaps this is a case where brute force just does not work or is highly inefficient. Would this method work better? Would it be more efficient? How feasible is the method with large number of parameters? The authors approach the issue, but the answer is not clear. One answer to this question would be to illustrate the method with a heavier example (e.g., 20 parameters, conditions with different time scales, …). Details on learning times and hardware used can also help here.  
  
The abstract nicely states that “... such models are valuable when they give rise to an experimentally observed phenomenon -- whether behavioral or in terms of neural activity…”.  
In this spirit, it was generally difficult to connect the elegant formalism presented here with actual experimental observations. Take Figure 2 as an example. The authors claim that this is a model of V1, but it is not clear at all what aspect of V1 function this model aims to capture, if any at all. This seems to be a model of 4 different types of interconnected units, not a model of V1, at least there does not seem to be anything specific to V1 (other than the fact that these types of neurons may be present and interconnected in V1, as well as in many other places in cortex, in all likelihood). To the best of our understanding, there does not seem to be any experimental observation in this figure. As an “emergent property”, here the authors seem to use a certain differential steady state in a given model unit activity.

Fundamental lack of clarity in what we’re suggesting to do with the method. Naming semantics. Ambiguity and softness on entropy lead to confusion, when ultimately we have a clear and compelling answer.

Scientific findings should be fundamentally general, and demonstrate that generality.

It would be interesting to compare the similarities and differences with the work of Conglaves et al BiorXiv 2019 (the authors mention they are aware of that work, but a short comparison could be beneficial):  
<https://www.biorxiv.org/content/10.1101/838383v2>

R2 does not get that we condition on multiple levels of the EP.

MUST compare to other methods and demonstrate improvement in some sense.