scientific reports

Check for updates

OPEN

DNN-assisted statistical analysis of a model of local cortical circuits

Yaoyu Zhang[1]✉ & Lai-Sang Young[2,3]✉

In neuroscience, computational modeling is an effective way to gain insight into cortical mechanisms, yet the construction and analysis of large-scale network models—not to mention the extraction of underlying principles—are themselves challenging tasks, due to the absence of suitable analytical tools and the prohibitive costs of systematic numerical exploration of high-dimensional parameter spaces. In this paper, we propose a data-driven approach assisted by deep neural networks (DNN). The idea is to first discover certain input-output relations, and then to leverage this information and the superior computation speeds of the well-trained DNN to guide parameter searches and to deduce theoretical understanding. To illustrate this novel approach, we used as a test case a medium-size network of integrate-and-fire neurons intended to model local cortical circuits. With the help of an accurate yet extremely efficient DNN surrogate, we revealed the statistics of model responses, providing a detailed picture of model behavior. The information obtained is both general and of a fundamental nature, with direct application to neuroscience. Our results suggest that the methodology proposed can be scaled up to larger and more complex biological networks when used in conjunction with other techniques of biological modeling.

One can distinguish between two types of mathematical models in the study of biological systems: *phenomenological models* that are intended to describe or summarize empirical observations, e.g. results of pyscho-physics experiments, and *biology-based models* that incorporate the underlying anatomy or physiology, e.g. neuronal interactions in the cerebral cortex. Both types of models are widely used, and they serve very different purposes. The work reported in this paper is motivated by the benefits and challenges of models of the second kind. The benefits are clear: by seeking to quantitatively reproduce a biological process, these models have the capability to capture emergent behaviors; they have the potential to offer insight into biological mechanisms, and to have predictive power. These benefits, however, come at considerable costs. Biology-based models are invariably highly complex, involving very large numbers of variables with complicated interactions. Gaps in one's knowledge of the system typically translate into unknown parameters in mathematical modeling, and in biological models, the number of such parameters tends to be large. As is well known to be the case, systematic exploration of high dimensional parameter spaces is computationally not feasible.

In this paper, we propose a strategy to assist in the construction and analysis of detailed biological models. The idea is as follows. Even though such models are high dimensional, complex dynamical systems, there tends to be a finite number of quantities or observations that are of special interest. Our proposal is to identify a finite number "inputs" and "outputs" of the model that are important to us—unknown parameters, for example, can be in the "inputs" category—and to first discover, without prejudice, an approximation of the *input-output mapping*. Such a task is well suited to deep neural nets (DNN). Once this mapping is constructed, we can use the information gained together with the vastly superior computing speeds of the DNN to assist in parameter tuning and model analysis.

That is to say, as a substitute for parameter exploration via direct simulation, our proposal is to train a DNN from limited mapping data obtained by simulation. After learning, a well-trained DNN can serve as a surrogate for the original model to inform on output values for given sets of parameters and inputs. Because the DNN can generate input-output pairs far more quickly than actual simulations of the network model, with speeds exceeding easily 10,000 times that of actual simulations (e.g. fractions of a millisecond *versus* minutes to hours per trial), it has the capability to provide large collections of data points, which can then be used for systematic statistical analyses leading to a better understanding of network behavior. On the practical level, such a surrogate model can be used for automated parameter tuning in model construction, and it can be used to inform on the limitations

[1]School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China. [2]School of Mathematics and School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA. [3]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA. ✉email: zhyy.sjtu@sjtu.edu.cn; lsy@cims.nyu.edu

of existing models, i.e., whether or not a model has the capability to produce certain outputs. Both model building and their statistical analysis are essential steps towards a better understanding of biological mechanisms.

A neuronal network of modest size and complexity will be used for demonstration. We view this model as a testbed to study the efficacy of the methodology, and to assess the feasibility of scaling up to models that are larger and more complex. In more detail, we consider in this paper a network of integrate-and-fire neurons intended to model a local circuit in the mammalian cerebral cortex; a mechanistic understanding of such circuits is instrumental to understanding cortical computation. In this model, the "inputs" include external drive to the local population and synaptic coupling weights within the population, and "outputs" are mean firing rates of excitatory and inhibitory neurons. The model has a $\sim$ 1000-dimensional phase space and 7 parameters. While it does not have the complexity of models such as[1–6], no systematic study of parameter dependence has been performed up until now; indeed exploration of a 7D parameter space by direct simulation is impossible. However, with the aid of a well-trained DNN, we were able to reveal the statistics of model responses and to provide a broad picture of model behavior.

We finish with the following remarks on the use of DNNs. That the subject has achieved huge success in many areas of applications[7,8] needs no elaboration. It has also firmly established its place in fundamental research[9]. In neuroscience, DNNs, specifically the hierarchical convolutional neural networks, have been used to model single-unit and population responses in higher visual cortical areas[10]. Our DNN-assisted approach falls into the general framework of surrogate-based modeling, a well established practice in engineering with wide applications to many problems that involve complex simulations or experiments (see[11–13] for reviews). In biology, the use of surrogate models has been more limited but there are precedents, e.g., support vector machines have been recently explored in hemorrhage and renal denervation[14] and yeast mating polarization[15]. A purpose of this paper is to further promote this approach in biological modeling, in the area of computational neuroscience in particular. Note that, there are other approaches proposed for parameter tuning of neuronal circuit models (e.g., Refs.[16,17]), and we believe a DNN-surrogate used in combination with these modeling techniques under experimental guidance can lead to substantial advances in the subject.

Results

This paper is about the use of a DNN-surrogate to assist in the analysis of model outputs for a neuronal network intended to model local circuits in the cerebral cortex. The model is a network of conductance-based integrate-and-fire neurons and is described in detail in "Materials and methods" ("I&F neuronal model"). The deep neural net that will serve as surrogate for this model is described in "Materials and methods" ("DNN surrogate"). We begin by framing the problem and outlining our approach, to give the reader a sense of our perspective. This is followed by preliminary information on the capabilities of the DNN. We then present our first key results, which consist of a statistical analysis of the derivatives of model responses and their interpretation. We will demonstrate that such analyses can have surprisingly rich implications. The last part of this paper discusses another use of surrogates in biological modeling, namely to assist in the evaluation of the capabilities and limitations of models.

DNN-assisted approach: setup and overview. We study a neuronal model of local cortical circuits with the goal of understanding its dependence on parameters and input values, and our approach is to first discover the mapping

$$\text{parameter space} \times \text{input space} \rightarrow \text{output space}.$$

This mapping is then used to assist in the analysis of model dynamics and cortical mechanisms. The proposed methodology avoids parameter tuning, and represents a different viewpoint than standard dynamical systems approaches. As we will show, it is well suited for data-driven inferences using neural networks, and provides useful statistical information that has the potential to help unravel what goes on in complex dynamical systems.

As illustration of this methodology, we consider a homogeneously connected network of integrate-and-fire (I&F) neurons that can be thought of as a generic model of a local neuronal population. This is a dynamical system of medium complexity, with $\mathcal{O}(10^3)$ state variables. The equations governing its dynamical evolution are given in "Materials and methods" ("I&F neuronal model"). The undetermined parameters of this model are the coupling weights between excitatory (E) and inhibitory (I) neurons. These synaptic coupling weights are denoted by S^{XY} where $X, Y \in \{E, I\}$; S^{EI}, for example, represents the amount of influence an I-spike has on a postsynaptic E-cell. The inputs to the model network are described by the following three numbers: $\eta^{\text{ext,E}}$ and $\eta^{\text{ext,I}}$ are the amounts of external drive supplied to the E and I-neurons in the model population, and η^{amb} is an "ambient" drive intended to depict modulatory influences from outside of the population.

The objects of our study are population mean firing rates, the most fundamental statistical quantities of a neuronal circuit. Specifically, we will focus on r^E and r^I, the mean firing rates of E and I-neurons in the model.

In the setup above, the mapping to be discovered and analyzed is

$$P \rightarrow O, \qquad \text{where} \qquad O = [r^E,\, r^I] \quad \text{and} \quad P = [P_{\text{S}},\, P_{\text{I}}],$$

and P_{S} and P_{I} are as follows: For reasons to become clear we have chosen to represent the parameters corresponding to synaptic coupling between E and I-cells as

$$P_{\text{S}} = [S^{EE},\, S^{EI}/S^{EE},\, S^{IE}/S^{EE},\, S^{II}/S^{EI}],$$

i.e., we scale the other three parameters to S^{EE} or S^{EI}, and to represent the input parameters as

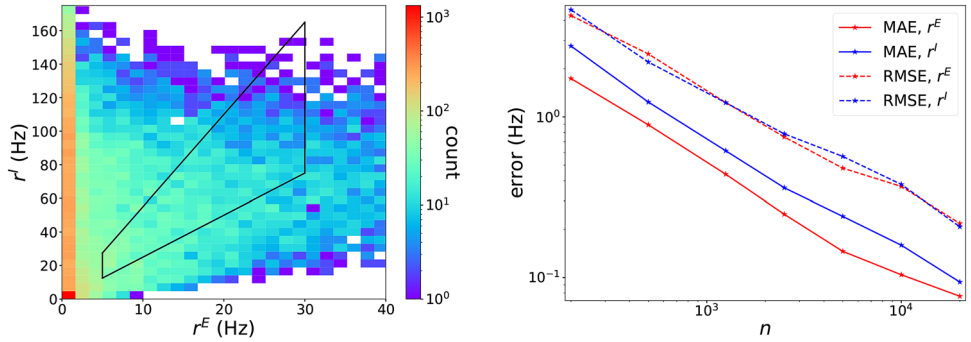


Figure 1. Viability of parameters and DNN performance. (**a**) Heat map of count of $O = [r^E, r^I]$ in $\mathscr{D}_{\text{train}}^{20,000}$ (white indicates zero counts). The trapezoid zone indicates the biological plausible range of E and I firing rate pairs in an active state of the circuit, i.e., 5Hz $\leq$ r$^{\text{E}}$ $\leq$ 30Hz and $2.5 \leq r^I/r^E \leq 5.5$. (**b**) Testing accuracy of DNN well-trained on $\mathscr{D}_{\text{train}}^{n}$ as a function of training sample size n. Mean-absolute errors (MAE) (solid line) and root-mean-square errors (RMSE) (dashed line) for r^E (red) and r^I (blue) are exhibited.

$$P_{\text{I}} = [\eta^{\text{amb}}/\eta_0,\ \eta^{\text{ext,E}},\ \eta^{\text{ext,I}}/\eta^{\text{ext,E}}]\,,$$

where η_0 is a kind of normalizing constant.

Additionally, we specify in advance a physiological domain $\mathscr{P}$ for P. These parameter ranges correspond to *a priori* biological constraints either deduced from indirect experimental measurements or gleaned from previous modeling results (such as[3]); they are effectively educated guesses. We also identify a physiological domain $\mathscr{O}$ for O consisting of firing rates observed in the laboratory under a variety of circumstances. We did not know in advance—and do not assume—that $P \in \mathscr{P}$ will produce $O \in \mathscr{O}$.

This completes a description of the setup for the rest of this paper. The mapping $P \to O$ is the mapping alluded to at the beginning of this section. We will train a DNN, details of which are given in "Materials and methods" ("DNN surrogate"), to learn this mapping from limited data obtained from simulation. Through the training, the DNN gradually interpolates the discrete data by a smooth function, allowing efficient evaluation and differentiation. Once we are satisfied that the DNN is performing satisfactorily, we will replace the original neuronal model by the DNN. The DNN surrogate is *a model of the original neuronal model*, one that is more limited in scope (it is focused solely on the mapping $P \to O$) but computes at vastly higher speeds and performs efficiently certain operations that are difficult or impossible via simulation of the original model. It serves as a compass, enabling us to explore more systematically model responses as parameters are varied in a high dimensional space.

In computational modeling, DNN surrogates can assist by offering baseline values to initialize searches and by proposing parameter corrections along the way. It provides a general description of input-output relations as well as statistical information on the effects of perturbations, tasks that are well suited to the DNN. This paper is not a modeling paper and we will not get into specific instances of parameter tuning, but as an example of the theoretical insight that DNN surrogates can offer, we will present a derivative analysis of the $P \to O$ mapping. To our knowledge such an analysis has not been done before for a large network of integrate-and-fire neurons.

Finally, there are two aspects of model analysis that we would like to illustrate in this paper. One is what the model can tell us about neural mechanisms, that is, having skipped over the dynamical process, how we can now use the $P \to O$ mapping to deduce what may be going on in the neuronal model, in the hope of shedding light on what goes on in real cortex. But there is another aspect to model analysis that is also very important: all models are limited in scope because they are orders or magnitudes simpler than the real brain, and it is important to understand the limitations of a model, whether it has the capability to reproduce specific types of neural phenomena. We will finish by presenting an example of that.

Performance of DNN surrogate. Firing rates can be measured experimentally using electrophysiology, or estimated using various kinds of optimal imaging techniques. On the theoretical level, however, how firing rates depend on network properties and inputs is not well understood, as firing rates cannot be computed analytically in semi-realistic network models such as the one described in "Materials and methods" ("I&F neuronal model"). In this paper we will use the DNN surrogate as an investigative tool to study these questions, but before we do that, we need to first confirm the viability of our physiological range $\mathscr{P}$ (see "Materials and methods", "I&F neuronal model" for details) and document the performance of the DNN surrogate. With regard to the latter, we will examine the accuracy of the DNN surrogate as a function of the size of its training set, and we will investigate its performance in parameter tuning, i.e. to solve the inverse problem of locating parameters to produce target outputs.

Viability of parameters and DNN performance. To confirm the viability of our *a priori* choice of physiological domain $\mathscr{P}$, we randomly selected 20, 000 sets of P from this domain and computed from simulations their mean

population firing rates O, which forms a training dataset $\mathscr{D}_{\text{train}}^{20,000}$. The results are presented in Fig. 1a. The physiological domain $\mathscr{O}$ consists of values in the region bounded by the trapezoid. Fig. 1a confirms that parameters from $\mathscr{P}$ produce firing rates in a broad region containing $\mathscr{O}$, justifying our choice of $\mathscr{P}$. It also shows that only about 10% of the outputs O actually fall into the trapezoidal zone, underscoring the challenges in prescribing P for desired firing rates.

We then investigated the accuracy of DNN surrogates trained on datasets of various sizes from 200 to 20, 000, The mean-absolute error (MAE) and root-mean-square error (RMSE) of well-trained DNNs on the testing dataset are presented in Fig. 1b. The error follows approximately a power law decay of $\sim n^{-2/3}$ where n is the size of the training set, much faster than the $\sim n^{-1/7}$ law implied by the curse of dimensionality. This curse-of-dimensionality free convergence behavior of DNN is supported by theoretical studies[18,19]; it is one of the reasons why DNNs are widely used for high dimensional problems.

Note also that with a surprisingly small size of 500 training data points, a small error (MAE) of ~ 1 Hz was obtained. In these experiments, errors are roughly independent of firing rate, resulting in smaller relative errors at high firing rates and larger relative errors at low firing rates. For predictions that result in a target E-firing rate of ~ 10Hz, the relative prediction errors of our DNNs typically are $\sim 10\%$ and $\sim 1\%$ with 500 and 20, 000 training data, respectively. By theoretical studies of DNN[20,21], such a good performance suggests a low complexity/frequency nature of the $P \to O$ mapping, i.e., its power is mainly concentrated at low frequencies in the Fourier domain.

The sigmoid function $1/(1+\mathrm{e}^{-x})$, which is used as the activation function of our DNN, yields far lower testing errors than the popular choice of ReLU. A key difference between ReLU and sigmoid activation is their smoothness, a property more important for regression problems as considered in this paper than for classification problems which are commonly considered by the AI community. As suggested in Ref.[21], when the smoothness of activation matches the smoothness of the target function, an optimal error bound can be achieved. Thus the better empirical performance of sigmoid compared to ReLU activation suggests a smooth nature of the $P \to O$ mapping, a point we will revisit later on in our analysis. We remark also that smooth activation functions like sigmoid or tanh (hyperbolic tangent, a rescaled sigmoid function) have been shown to be better choices for other regression problems, e.g., in molecular dynamics simulation[22,23]. In practice, apart from the sigmoid or tanh activation, elu, selu, gelu are also suitable for fitting smooth target functions. Their subtle differences are a subject of study in its own right; this is out of scope of the present work.

In the rest of this paper, we will use the most accurate DNN well-trained on $\mathscr{D}_{\text{train}}^{20,000}$ as a surrogate to investigate the statistical properties of the $P \to O$ mapping.

Performance of DNN surrogate for parameter tuning. Realistic models of neuronal circuits typically involve large numbers of parameters corresponding to quantities not directly measurable in the laboratory. Fitting these parameters to experimental observations is an essential task. Up until now, this task has often been done "by hand", relying on the experience of the modeler. As such, it is both laborious and time-consuming if it can be successfully carried out at all. Because of the high dimensionality of the parameter space, and the difficulty in directly computing the derivatives $\nabla_P O$ from discrete data points, automated gradient-based approaches widely used in many applications are not viable in this kind of parameter tuning.

Our first demonstration of the usefulness of an accurate DNN surrogate is to apply it to the problem of automated parameter tuning. This is an inverse problem, requiring that we find parameter P given target output O_{target}. Assisted by the DNN surrogate $\hat{O}(P)$ well trained on $\mathscr{D}_{\text{train}}^{20,000}$, whose derivatives can be easily computed by back-propagation, a gradient-based approach can be efficiently applied as follows. In each iteration step t, P^t is updated as

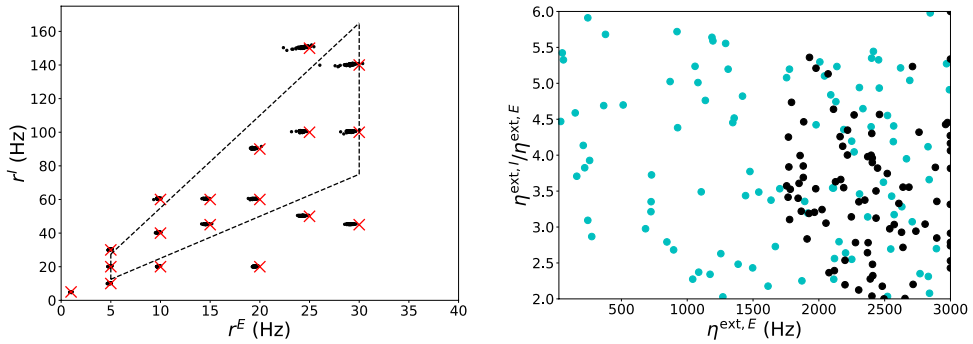
$$P^{t+1} = P^t - \alpha \nabla \hat{O}(P^t)\Big(\hat{O}(P^t) - O_{\text{target}}\Big),$$

where α is the learning rate.

Figure 2 shows the results of a numerical experiment we performed. In this experiment, the initial parameter P^0 was randomly sampled from $\mathscr{P}$, and if P^t fell outside of the domain, it was projected back to $\mathscr{P}$. For each O_{target}, we selected 100 random initial parameters. After 10000 steps of iteration, all final P's with predicted output sufficiently close to the target, e.g., with $||\hat{O}(P) - O_{\text{target}}||_1 < 0.2$ Hz, formed the candidate set of parameters for O_{target}. For acceleration, we incorporated the scheme of Adam[24].

In general, given O_{target}, a randomly chosen parameter in $\mathscr{P}$ has probability $< 0.01\%$ to be a candidate parameter. The iterative scheme above was intended to autonomously steer it towards a candidate. An example is depicted in Fig. 2b: parameters at initialization (projected from 7D) are represented by cyan dots; they are steered to black dots through the tuning for a given target. For each target, we found that of the 100 initial parameters picked, on average over 90% successfully yielded a candidate after 10000 steps. The accuracy of the candidate parameters were then evaluated by comparing their simulated outputs (black points) with the corresponding targets (crosses). The error was larger than the mean testing error of ~ 0.1 Hz for the DNN, as can be expected for an inverse problem. However, except from parameters in the periphery of $\mathscr{P}$, most tuning results were faithful to the target. Note that, the accuracy of the above parameter tuning approach can be further improved by incorporating a few trials of simulation online to fix the local prediction error of the DNN surrogate.

Figure 2b illustrates another important point, namely that for a given target O_{target}, the parameters obtained by the above tuning process are far from unique. In Fig. 2b, different pairs of input strengths $\eta^{\text{ext,E}}$ and $\eta^{\text{ext,I}}$ indicated by black dots (each with its own accompanying parameters in the other 5 dimensions) give rise to the same E and I firing rates of 25 Hz and 100 Hz, respectively. Indeed, if the $P \to O$ mapping is smooth, one would expect, for



(a) (b)

Figure 2. Visualization of DNN-assisted parameter tuning. (**a**) Visualization of accuracy of parameter tuning. Crosses are target firing rates, black dots are the simulated output of the candidate parameters found through gradient-based approach assisted by the DNN surrogate. The trapezoid constrained by the dashed lines indicates the physiological domain of output. (**b**) Parameters at initialization (cyan dots) and after tuning (black dots) for a target $O_{\text{target}} = [25\,\text{Hz}, 100\,\text{Hz}]$ projected to the 2D plane of $\eta^{\text{ext,E}}$ and $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$.

each given O_{target}, the set $\{P : \hat{O}(P) = O_{\text{target}}\}$ to be a 5D submanifold in our 7D parameter space. In modeling, additional physiological phenomena will likely place further constraints on the set of viable parameters.

Statistical analysis of parameter dependence: first derivatives. Crucial for understanding cortical mechanisms is a quantitative description of how the firing rates of a brain region depend on its structural and input parameters. Yet except for extremely idealized models with few state variables, there is no explicit relation between these parameters and firing rates, and exploration of parameter space via simulations is not feasible as we have explained earlier. In this paper, we propose a statistical approach to this problem via the use of DNN surrogates.

In Fig. 1a, we presented the statistics of firing rate responses for parameters in $\mathscr{P}$. This section focuses on statistics on the *derivatives* of output responses. Our study is assisted by the well-trained DNN surrogate $\hat{O}(P)$, which allows very efficient evaluation and differentiation. To our knowledge, this is the first time that parameter dependence of firing rates in integrate-and-fire models are systematically investigated through a statistical analysis.

Quantitative information on $\nabla_P O$ will shed light on a number of questions. Of particular interest is a system's response to changes in its input. As we will show, our statistical analysis points to a dichotomy in the response behavior of neuronal populations. It supports a novel interpretation of "high gain" that may have implications in cortical phenomena such as surround suppression.

Derivative analysis. Recall that in our model, input parameters are

$$P_I = [\eta^{\text{amb}}/\eta_0,\ \eta^{\text{ext,E}},\ \eta^{\text{ext,I}}/\eta^{\text{ext,E}}] \quad \text{and} \quad P_S = [S^{EE},\ S^{EI}/S^{EE},\ S^{IE}/S^{EE},\ S^{II}/S^{EI}]\,,$$

and output parameters are

$$O = [r^E,\ r^I]\,.$$

Using the DNN surrogate $\hat{O}(P)$ trained on $\mathscr{D}_{\text{train}}^{20{,}000}$ (see "Materials and methods", "DNN surrogate"), one can easily compute $\nabla_P \hat{O}$, which approximates $\nabla_P O$, over a very large number of input parameters. Figure 3a,b show the distributions of partial derivatives $\nabla_P \hat{r}^E$ and $\nabla_P \hat{r}^I$, respectively, with respect to each of the seven parameters in $\{P_S, P_I\}$. (We write $\hat{r}^E, \hat{r}^I$ to stress that these results are computed from the DNN surrogate $\hat{O}(P)$.) The histograms in Fig. 3 were computed from 5×10^5 randomly selected $P \in \mathscr{P}$, keeping only the $\sim 10\%$ of P for which $\hat{O}(P) \in \mathscr{O}$ and discarding the rest.

To familiarize the reader with the meaning of the plots in Fig. 3, consider, for example, differentiating with respect to S^{EI}/S^{EE} keeping the other 6 parameters fixed. The second columns of the two panels show that both $\partial_{S^{EI}/S^{EE}}\,\hat{r}^E$ and $\partial_{S^{EI}/S^{EE}}\,\hat{r}^I$ are almost always negative indicating that increase of strength from I to E consistently decreases the firing rate of both the E and the I-population. In addition, the magnitude of $\partial_{S^{EI}/S^{EE}}\,\hat{r}^I$ is in general larger than $\partial_{S^{EI}/S^{EE}}\,\hat{r}^E$, indicating that changes in S^{EI}/S^{EE} have a larger effect on I-firing rate, not surprisingly since I-firing rates are generally 3 to 4 times larger than E-firing rates[3,25,26].

Differentiating with respect to S^{IE}/S^{EE} yields rather curious results: while $\partial_{S^{IE}/S^{EE}}\,\hat{r}^E$ is always strongly negative, $\partial_{S^{IE}/S^{EE}}\,\hat{r}^I$ can be positive or negative with a relatively small magnitude. This statistical result suggests the existence of an interesting regime where increasing the synaptic strength of E to I (while keeping that from E to E fixed) decreases the firing of the I-population (even though the strength of E to I is increased) and it suppresses the firing of the E-population (even though I-firing is lowered). This model behavior is reminiscent of the "paradoxical effect" identified earlier in[27–30]. We will revisit this point in the next subsection.

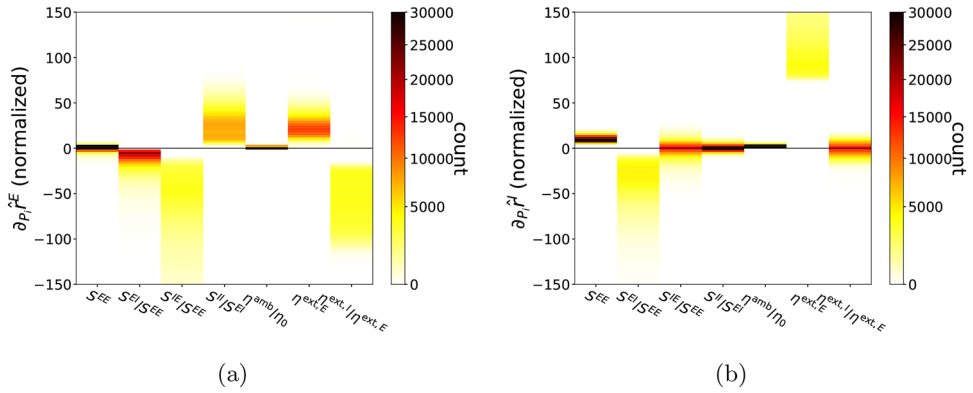


Figure 3. First derivative statistics. 2D Histograms of $\nabla_P \hat{O}$ for P's satisfying $P \in \mathscr{P}$ and $\hat{O}(P) \in \mathscr{O}$. The partial derivatives are normalized as (**a**) $\nabla_P \hat{r}^E \times \Delta$, (**b**) $\nabla_P \hat{r}^I \times \Delta$ where Δ is a vector consisting of the length of physiological domain for each parameter in P. Δ is used to fix $\nabla_P \hat{r}^E$ to a dimensionless unit with $\mathscr{P}$ scaled as a unit box. P_i refers to the i-th element in P indicated sequentially in the x-axis.

The following information on the dependence of response properties on parameters can be gleaned from Fig. 3:

(1) *Parameter dependences are nonlinear.* Figure 3a,b ruled out the possibility that r^E and r^I are as simple as a linear function of P because most of the partial derivatives are clearly nonconstant; some in fact have quite a large spread.
(2) *Dependence on η^{amb}/η_0 is insignificant and dependence on S^{EE} is weak.* As the other three synaptic weights are indexed to S^{EE} in our bookkeeping, the relatively weak dependence on S^{EE} when the other parameters are fixed confirms our conjecture (see "Materials and methods", "I&F neuronal model") that not a great deal changes when the four synaptic weights S^{EE}, S^{EI}, S^{IE} and S^{II} are scaled up and down together as long as they maintain the same relationship.
(3) *Near-monotonicity of the function $P \mapsto \hat{r}^E$.* Differentiating $\hat{r}^E$, one sees that 5 out of the 7 partial derivatives have a single sign, i.e., they are either positive or negative for all the parameters tested, and the remaining two are relatively small. All this points to a simple structure for the mapping $P \mapsto \hat{r}^E$. One notes also that the signs of the 5 all go in directions expected: increasing I to E and E to I lowers $\hat{r}^E$ as one would expect as that increases the power of the inhibition, increasing I to I increases $\hat{r}^E$, and increasing external drive to E increases $\hat{r}^E$ while increasing external drive to I lowers $\hat{r}^E$—all are as expected.
(4) *The mapping $P \mapsto \hat{r}^I$ is more complex.* Our statistics show that the I-responses are not as clean as E-responses, in that changes in $\hat{r}^I$ in response to increases in S^{IE}/S^{EE}, S^{II}/S^{EI} and $\eta^{ext,I}/\eta^{ext,E}$ can be positive or negative. As noted earlier, the idea that increasing drive to I-neurons could decrease $\hat{r}^I$ is somewhat counter-intuitive. With the help of the DNN surrogate, we examine next in more detail the circumstances surrounding this response reversal of I-neurons.

Cortical mechanisms via DNN-assisted derivative analysis: an illustrative example. The phenomenon that stimulation of an inhibitory population not only decreases the activity of the excitatory population but that it can also decrease the activity of the stimulated population is known to the neuroscience community. The intuition is that the excitatory population is sufficiently suppressed that the total excitation received by the inhibitory population is reduced[27–32]. In rate models, it has been demonstrated mathematically that this occurs in inhibition stabilized networks (ISN), where recurrent excitation is strong and the regime is stabilized by inhibition[27–29]. Models with multiple inhibitory populations have also been investigated recently[30,33–35]. For network models of integrate-and-fire neurons such as the one studied here, analytical approaches are not viable, and conditions for the reversal of I-response have not been investigated. This is what we would like to do using a DNN-assisted statistical analysis.

Response of I-neurons: a dichotomy. Following up on the observation in Item (4) above, namely that $\hat{r}^I$ may increase or decrease in response to changes in S^{IE}/S^{EE}, S^{II}/S^{EI} and $\eta^{ext,I}/\eta^{ext,E}$, we looked into potential correlations between the signs of these partial derivatives. The results are shown in Fig. 4a, and they show that the signs of these partial derivatives are highly correlated to one another, with correlations very close to ± 1 (see "Materials and methods", "Correlation analysis and logistic regression" for details). This suggests the existence of two distinct regimes: one in which an increase in S^{IE}/S^{EE} or $\eta^{ext,I}/\eta^{ext,E}$, or a decrease in S^{II}/S^{EI} causes r^I to increase, and another in which the same changes cause r^I to decrease.

While S^{IE}, $\eta^{ext,I}$, and S^{II} directly contribute to the input received by the I-population as illustrated in Fig. 4b, the positivity of correlations with respect to changes in S^{IE} and $\eta^{ext,I}$ is not clear *a priori*, because these changes also affect the firing rates of E-neurons, and the synaptic excitatory input from within the population to an I-neuron is determined not just by S^{IE} but also by r^E, the firing rate of the E-population. The same is true for the

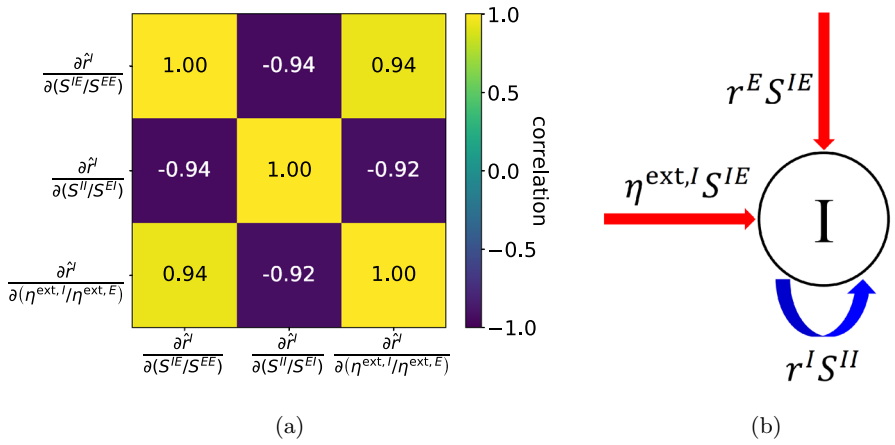


Figure 4. Illustration and analysis of I-population response. (**a**) Correlation matrix of the sign of $\frac{\partial r^I}{\partial (S^{IE}/S^{EE})}$, $\frac{\partial r^I}{\partial (S^{II}/S^{EI})}$ and $\frac{\partial r^I}{\partial (\eta^{\text{ext,I}}/\eta^{\text{ext,E}})}$. (**b**) Illustration for different input sources received by the I population. Red indicates excitation and blue indicates inhibition.

effect of S^{II}: increasing that does not necessarily mean that an I-neuron will receive greater suppression, because the amount of inhibitory synaptic input it receives depends also on r^I.

To summarize, our results as shown in Figs. 3 and 4a show that the set of parameters

$$\mathscr{P}^* := \{P \in \mathscr{P} \text{ such that } \hat{O}(P) \in \mathscr{O}\}$$

can be divided into two distinct groups according to the sign of $\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}} \hat{r}^I$, equivalently the sign of either one of the other two partial derivatives. This means that the mapping $P \mapsto \hat{r}^I$, which we had noted earlier might be considerably more complex than $P \mapsto \hat{r}^E$, has a fairly simple structure after all. The simplicity of the mapping $P \mapsto O$ may be the reason why DNNs achieve very good accuracy even for training datasets of small sizes.

Below we will refer to the phenomenon of $\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}} \hat{r}^I < 0$ as "inhibitory response reversal".

Correlating network properties to inhibitory response reversal. We first used the seven quantities in P to predict the sign of $\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}} \hat{r}^I$ by logistic regression, i.e., we used the logistic function $1/(1+e^{-\mathbf{a}\cdot P+b})$ to fit the probability of $\text{sign}\left(\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}} \hat{r}^I(P)\right) = 1$ (it is either equal to 0 or to 1 in this problem). Similar to linear regression for real-valued output, in machine learning, logistic regression is often a first try for fitting binary output with real-valued input. After regression, the accuracy of prediction using the sign of $1/(1+e^{-\mathbf{a}\cdot P+b}) - 0.5$ is $\sim 83\%$ over $P \in \mathscr{P}^*$. This accuracy indicates that signs of the target can roughly be separated by a hyperplane in the space of P (100% indicates perfectly linearly separable while chance rate 50% indicates complex behavior far from linearly separable). Relative importance of each parameter P_j is evaluated by $\frac{a_j^2 \text{Var}_{\mathscr{P}^*}(P_j)}{\Sigma_j a_j^2 \text{Var}_{\mathscr{P}^*}(P_j)}$, where $\text{Var}_{\mathscr{P}^*}$ indicates the variance over $\mathscr{P}^*$ (see Fig. 5a).

Clearly, S^{IE}/S^{EE} and $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$ are the two most salient factors for regime determination. The performance of regime separation using these two parameters is shown in Fig. 5c. One can see a trend that smaller values of S^{EI}/S^{EE} and $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$ indicating weak drives to the I-population are more likely to result in inhibitory response reversal. The prediction accuracy by logistic regression using only these two parameters yields a significantly worse accuracy of $\sim 67\%$, indicating that the ignored input dimensions in fact play nonnegligible roles in the prediction, and there is no clean linear separation between the two regimes in the space of P.

As noted in Fig. 4b, r^E and r^I also play important roles in determining the inputs that go into I-neurons, so we experimented next with using P and O together for the prediction of the sign of $\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}} \hat{r}^I$. After logistic regression, we achieved a surprisingly high accuracy of $\sim 97\%$. Moreover, as shown in Fig. 5b, $\eta^{\text{ext,E}}$ and $\hat{r}^E$ stood out as effectively the only key factors that mattered for the prediction. By using only $\eta^{\text{ext,E}}$ and $\hat{r}^E$, one can still achieve a very high prediction accuracy of $\sim 94\%$. This surprisingly good performance is illustrated in Fig. 5d, where the two regimes characterized by the sign of $\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}} \hat{r}^I$ are very well separated by a line of the form $\hat{r}^E = c\eta^{\text{ext,E}}$ for some $c > 0$. Note that, this c is clearly independent of 7 model parameters, however, may depend on other factors like connection probabilities that two randomly picked neurons are connected, fixed in our model.

A regime with a large excitatory response to external drives can be thought of as having high gain. Our results suggest that a natural definition of *high gain* might be $r^E > c\eta^{\text{ext,E}}$ for the critical value of c defined above. With this notion of gain, the above statistical analysis suggests that inhibitory response reversal occurs in a regime of high gain.

It is difficult to compare directly the parameters used in rate models and in networks of integrate-and-fire neurons. In our model, the physiological ranges of the parameters are chosen to be consistent with experimental data[3]. For parameters in this range, we found that sufficiently high gain, i.e., $r^E/\eta^{\text{ext,E}} > c$, is the best condition for inhibitory response reversal. This finding is new, it is quantitative, and it was discovered entirely through our

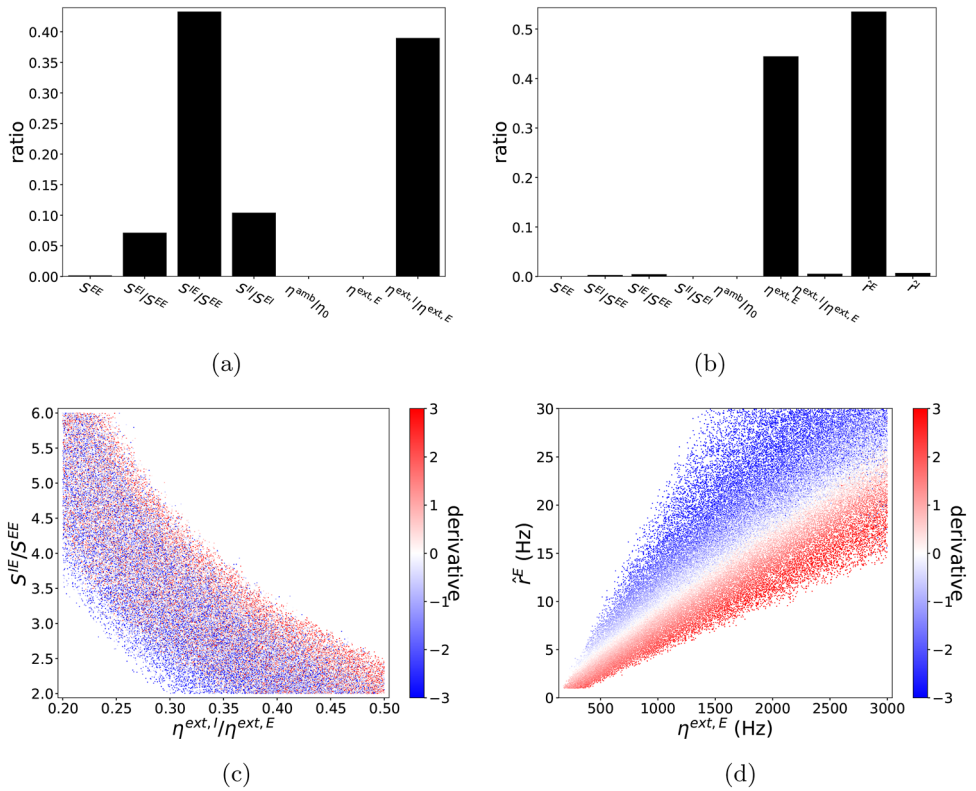


Figure 5. Analysis of inhibitory response reversal. Upper panels: Relative importance of each parameter in the best logistic predictor of the sign of $\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}}\hat{r}^I$ using (**a**) P only, (**b**) both P and $\hat{O}$. Bottom panels: $\partial_{\eta^{\text{ext,I}}/\eta^{\text{ext,E}}}\hat{r}^I$ (red indicates strongly positive and blue indicates strongly negative) over different pairs of (**c**) S^{IE}/S^{EE} and $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$, (**d**) $\hat{r}^E$ and $\eta^{\text{ext,E}}$, both selected based on the importance analysis in (**a**) and (**b**) respectively.

DNN-assisted analysis. The implications of this finding and its relation to ISN need to be explored; that will be done elsewhere. We finish with a direct application of this idea.

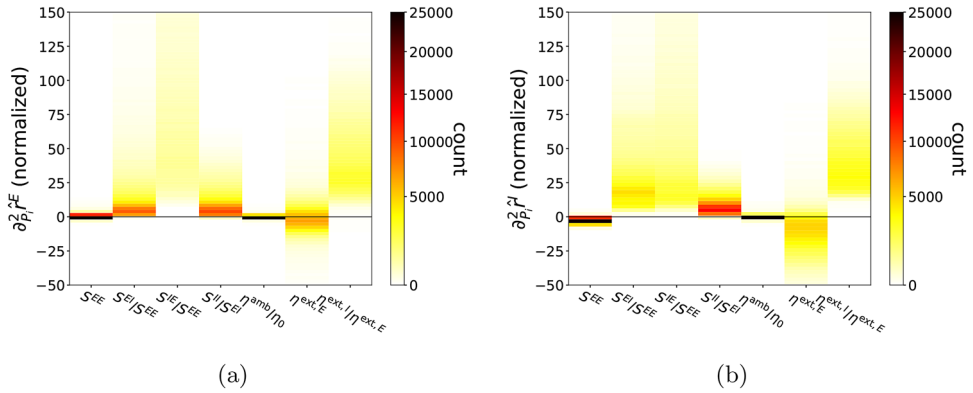
Plausible explanation for surround suppression. Surround suppression is a well documented visual phenomenon. It refers to the fact that a neuron's sensitivity to a stimulus is modulated by the extent of the stimulus outside of its classical receptive field. The discussion below is far from a systematic study of surround suppression, which is a wide-ranging and important topic in its own right. We wish to point out only a plausible explanation for the suppression associated with spatially extended stimuli that follows from the observations above.

To briefly review the phenomenon, consider an excitatory neuron in the primary visual cortex, V1. Drifting gratings of various sizes aligned with the neuron's orientation preference and centered at its receptive field are presented. It has been observed that while the neuron spikes vigorously in response to smaller gratings, its response peaks at a certain grating radius and decreases as the size of the grating continues to increase, leveling off eventually when the stimulus is many times the size of its classical receptive field[36]. This decrease in firing rate of a neuron at the center when the surround is also stimulated is called surround suppression. Experimental measurements of a quantity called suppression index indicates that the suppression of E-neurons can be quite strong depending on layer within V1[37]. For some layers, firing rates for large gratings may be no more than half those for smaller gratings. A similar phenomenon has been found to hold for I-neurons, though the decline in firing rate is smaller[28].

Here is how our results may be relevant:

Consider a local population located at the center, receiving external input from feedforward and feedback sources as well as from within its own layer via long-range connections. We hypothesize that for E-neurons in this population, as the size of the stimulus increases, $\eta^{\text{ext,E}}$ first increases and then saturates as the size of the grating continues to increase, whereas input to the I-population, $\eta^{\text{ext,I}}$, increases for a while longer saturating at a larger grating radius. This means that $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$ is at first constant and later increases. We further hypothesize that the circuit is always in a high gain state, i.e., $r^E/\eta^{\text{ext,E}}$ is always larger than the critical value c defined above.

When $\eta^{\text{ext,E}}$ and $\eta^{\text{ext,I}}$ are both increasing and $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$ is constant, our derivative analysis asserts that both r^E and r^I should be increasing, consistent with experimental observations before the size-tuning curves peak. When $\eta^{\text{ext,E}}$ saturates and $\eta^{\text{ext,I}}$ continues to increase, we are in the situation where the partial derivatives with respect to $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$ becomes relevant, and if the population is in a high gain state, then our derivative analysis predicts that r^I would decrease though not as steeply as r^E, a prediction in agreement with experimental data.



(a) (b)

Figure 6. Second derivative statistics. Histograms of second order partial derivatives for the physiological parameters yielding physiological output, i.e., as P vary over $\mathscr{P}$ * is presented. (**a**) $\frac{1}{2}\frac{\partial^2 \hat{r}^E}{\partial P_i^2}\Delta_i^2$, (**b**) $\frac{1}{2}\frac{\partial^2 \hat{r}^I}{\partial P_i^2}\Delta_i^2$ for index $i = 1, \ldots, 7$, where Δ is the length of physiological domain for each parameter in P. Here $\frac{1}{2}\Delta^2$ is used to fix second partial derivatives to a dimensionless unit with physiological domain of P scaled as a unit box.

To summarize, our proposed explanation suggests that it is entirely possible to have both r^E and r^I decrease while $\eta^{\text{ext,E}}$ and $\eta^{\text{ext,I}}$ are both increasing, provided the relative rates of increase in $\eta^{\text{ext,E}}$ and $\eta^{\text{ext,I}}$ are as above. We have also identified the property that is the key to what makes this possible, namely that the population should be in a state of high gain.

Analysis of second derivatives. Second derivatives reflect the acceleration and deceleration of the output in response to changes in parameters. In this section, we study the statistics of second derivatives, and investigate the model's capability to produce nonlinear outputs in response to increasing drive.

Distribution of second derivatives. Figure 6 displays the distribution of second partial derivatives of $\hat{r}^E$ and $\hat{r}^I$ with respect to each dimension of P. As an example of what these histograms tell us, consider the fact that $\frac{\partial^2 \hat{r}^E}{(\partial S^{EI}/S^{EE})^2}$ and $\frac{\partial^2 \hat{r}^I}{(\partial S^{EI}/S^{EE})^2}$ are always positive. Combined with our earlier result that $\frac{\partial \hat{r}^E}{\partial S^{EI}/S^{EE}}$ and $\frac{\partial \hat{r}^I}{\partial S^{EI}/S^{EE}}$ are both negative, we get the following picture: As S^{EI} increases (with all other parameters fixed), $\hat{r}^E$ and $\hat{r}^I$ both decrease, and the graphs are convex. The effect of S^{IE} is curious: As S^{IE} increases, $\hat{r}^E$ decreases and the graph is (quite strongly) convex. The graph of $\hat{r}^I$ is also convex, but since $\frac{\partial \hat{r}^I}{\partial S^{IE}/S^{EE}}$ can change sign, there is the possibility that it can decrease first and later increase.

In general, the following response properties can be inferred from the statistics of second derivatives.

(1) *Outputs are not describable by second-order polynomials.* Fig. 6a,b rule out the possibility that r^E and r^I can be as simple as second-order polynomials of P. Most second partial derivatives are clearly nonconstant, and some have quite a wide spread.
(2) *Insignificance of dependence on S^{EE} and η^{amb}/η_0.* This is consistent with results from our first derivative analysis.
(3) *Convexity of r^E and r^I as functions of all parameters in P except for $\eta^{\text{ext,E}}$.* This property further supports the simplicity of the mapping $P \mapsto O$.
(4) *Nonlinearity of gain curves.* We are concerned here with the second derivatives of r^E and r^I with respect to $\eta^{\text{ext,E}}$, i.e. when both $\eta^{\text{ext,E}}$ and $\eta^{\text{ext,I}}$ are increasing with the ratio of $\eta^{\text{ext,I}}/\eta^{\text{ext,E}}$ fixed. Firing rates almost always increase monotonically by our first derivative analysis, but they can accelerate or decelerate as our second derivative analysis shows. A more quantitative analysis reveals the following, however: While a typical change of $\hat{r}^E$ is $>$ 30Hz over the input domain, the normalized second derivative $\frac{1}{2}\frac{\partial^2 \hat{r}^E}{\partial (\eta^{\text{ext,E}})^2}\Delta_6^2$ is typically between $\pm$10Hz. The smallness of the second derivative compared to the first suggests that gain curves are statistically more likely to be fairly linear for our model with physiological parameters.

    As mentioned in the Overview of Results, one of the uses of a surrogate model is to inform on the limitations of the original neuronal network model. In real cortex, gain curves have been observed to be sigmoidal in shape. Item (4) in the second derivative analysis above raises the question of whether neurons in the model described in "Materials and methods" ("I&F neuronal model" are capable of producing such nonlinear gain curves. We now investigate this question more systematically using the DNN surrogate.

Generation of nonlinear gain curves. Gain curves capture changes of r^E in response to changes in external input. For convenience, we let P^- denote all the parameters of P except for $\eta^{\text{ext,E}}$, and study the gain curve $r^E_{P^-}(\eta^{\text{ext,E}}) = r^E(\eta^{\text{ext,E}}; P^-)$. In physiological experiments, sigmoidal gain curves are often observed[38], and neurotheories hinging on the shapes of gain curves have been proposed[39]. In this section, we study with the help of the DNN surrogate whether the model described in "Materials and methods" ("I&F neuronal model" is capable of producing gain curves that are sigmoidal in shape.

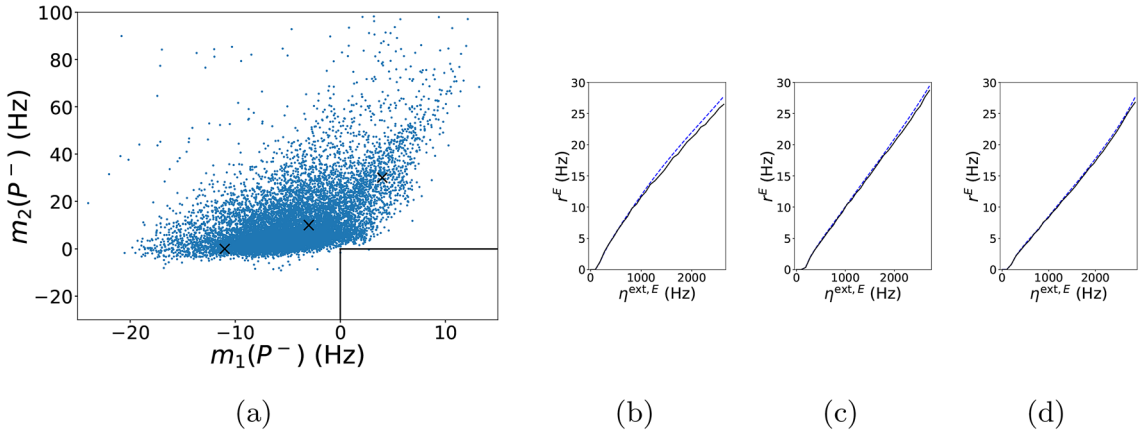


Figure 7. Model capability in generating sigmoidal gain curve. (**a**) m_1 vs. m_2 (see main text for notation) for physiologically plausible P^-'s. (**b**–**d**) are example gain curves corresponding to the three typical data points $[-11, 0]$, $[-3, 10]$ and $[4, 30]$ marked in (**a**), respectively. Blue dashed lines indicate surrogate gain curves $\hat{r}^E_{P^-}(\eta^{\text{ext,E}})$ whereas black lines indicate the simulated gain curves $r^E_{P^-}(\eta^{\text{ext,E}})$.

To capture the sigmoidal property, we require, for definiteness, that r^E as a function of $\eta^{\text{ext,E}}$ be accelerating for $r^E \in [5\,\text{Hz}, 15\,\text{Hz}]$, and decelerating for $r^E \in [20\,\text{Hz}, 30\,\text{Hz}]$. For each P^- in the physiological range, we increase $\eta^{\text{ext,E}}$, and as $r^E_{p^-}(\eta^{\text{ext,E}})$ increases, we identify the intervals J_1, J_2 of $\eta^{\text{ext,E}}$ that correspond to $r^E_{p^-}(\eta^{\text{ext,E}})$ falling in $[5\,\text{Hz}, 15\,\text{Hz}]$ and $[20\,\text{Hz}, 30\,\text{Hz}]$ respectively. We then compute the mean values of $\frac{1}{2}\frac{\mathrm{d}^2 r^E_{P^-}}{(\mathrm{d}\eta^{\text{ext,E}})^2}(\eta_{30\,\text{Hz}})^2$ on J_1 and J_2, and call them $m_1(P^-)$ and $m_2(P^-)$. Here, $\eta_{30\,\text{Hz}}$, which is determined by solving $r^E_{p^-}(\eta_{30\,\text{Hz}}) = 30\,\text{Hz}$, is used to normalize the second derivative to a unified dimensionless unit.

In Fig. 7, the x and y-axes show the m_1 and m_2 values for each plausible P^- satisfying (i) $[P^-, \eta^{\text{ext,E}}] \in \mathscr{P}$ and (ii) $O_{P^-}(\eta^{\text{ext,E}}) \in \mathscr{O}$ for $\hat{r}^E_{P^-}(\eta^{\text{ext,E}}) \in [5\text{Hz}, 30\text{Hz}]$. The lower right box bounded by the two black lines describes the region with the desired sigmoidal properties. As one can see, very few data points lie in this box. Some examples of gain curves are displayed in Fig. 7b–d, where results from the DNN surrogate and firing rates simulated directly from the neuronal network are superimposed. At least in these examples, our DNN surrogate quite accurately emulates the true behavior of the network model.

We conclude that the integrate-and-fire model described in "Materials and methods" ("I&F neuronal model") without further enhancement is incapable of producing gain curves that are sigmoidal in shape and that deviate substantially from a straight line. This is a limitation of the model. The present study should serve to inform the modeling community that to produce a sigmoidal gain curve with more pronounced curvature (as has been observed experimentally), some other mechanisms must be incorporated. In the V1 network model in[6], for example, mechanisms such as synaptic depression of I-neurons and potassium currents that prevent E-cells from firing repeatedly in rapid succession were implicated in contrast response properties.

Discussion

A broader aim of this work is to promote the use of machine-learning approaches in biological modeling. We propose that these more systematic methods can be useful not as replacement of but as supplement to conventional modeling techniques[9]. To demonstrate the efficacy of this approach, we considered a neuronal network built to resemble local circuits in the cerebral cortex, and illustrated how via the use of a surrogate DNN combined with data analysis (such as "Correlation analysis and logistic regression"), rich statistical structures can be extracted from limited data generated by simulation.

A specific approach that we are proposing here is the following: While biological processes are typically extremely complex, if one is able to build a model of the system modulo a finite—possibly very large—number of unknown parameters and identify a finite number of key quantities that best describe what goes on, then the modeling problem can be framed in terms of discovering the mapping from

$$\text{parameter space} \times \text{input space} \to \text{output space}.$$

Such input-output relations are especially well suited to data-driven inferences using neural nets. The statistical analysis of DNN surrogates in general suggests rather than proves any specific behavior of the target mapping, due to the presence of uncertainties intrinsic to any data-driven approach. Nevertheless, compared to heuristic arguments and *ad hoc* numerical explorations of parameter space, these results are quantitative in nature and provide strong supporting evidence for the conclusions they suggest.

On surrogate-based modeling and DNN. After a surrogate learns from data, it allows highly efficient manipulation including evaluation, differentiation, optimization (e.g. parameter tuning) and statistical analysis. Among a rich class of conventional surrogate models, many of which may serve our purpose equally well, DNN is convenient to use for a number of reasons: there are rich and sophisticated open source libraries (e.g., Tensorflow, Keras, Pytorch); DNN is faithful to data, with low training error; it is robust, generalizes well, and

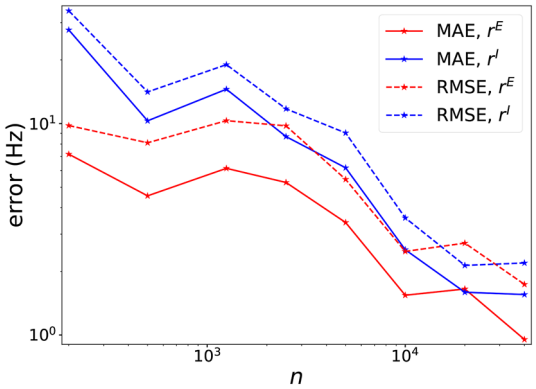


Figure 8. Performance of DNN trained on a large domain. Performance of DNN trained by training datasets of size n sampled from $\mathscr{P}_{\mathrm{L}}$ tested on a test dataset of size 10000 sampled from $\mathscr{P}$ is presented. MAE (dashed line) and RMSE (solid line) for r^E (red) and r^I (blue) are exhibited.

often does not require extra regularization; finally it is flexible, with universal approximation capability and rich architecture.

In engineering, the use of surrogate-based modeling to assist in the analysis and exploration of complex experiments and designs is well established. In spite of its huge success in tasks related to image, audio, and video recognition and processing, DNN up until now has largely remained a black box. It is only in recent years that researchers from different scientific disciplines have begun to exploit its many potentials. We believe that DNN surrogates can potentially be of great use in biological modeling, and it is with more complex models in mind that we have embarked in this direction. This paper is a first step to demonstrate, using a network model of a local cortical circuit, the type of statistical analysis made possible by such an approach.

Applications to neuroscience. Many questions remain. For local circuits, the P in our $P \to O$ mapping can include, e.g., connectivity and system size, O can be currents, and an important problem inspired by the balanced-state ideas[40–43] may be to quantify the balancing of currents under different network conditions. Nor must the target O be limited to firing rates and currents. It can include other quantitative measures of firing patterns, such as correlations and degrees of synchrony. A problem of interest is to relate gamma rhythms as characterized by their power spectral densities to network parameters[44,45], as gamma rhythms are known to be altered by disease, drugs and other physiological states[46–48] These are all potential applications of the methodology proposed.

Populations of homogeneously connected, i.e., the probability of connection is fixed depending on the connection type, and homogeneously driven neurons are ideal starting points for theoretical studies. A natural next step is to consider multi-component networks, beginning with source-target populations and progressing to more complicated network motifs with feedback loops. Neuronal networks in the real cortex are in fact not abstract graphs; they have spatial structures (see e.g.[49]). An ultimate use of DNN-assisted surrogates may be to reveal the mapping

$$\text{network structures} \times \text{stimulus} \quad \to \quad \text{temporal dynamics}$$

Outlook on the use of surrogates in biological (neural) modeling. High degrees of complexity and a low ratio of knowns to unknowns is characteristic of biological modeling. A case in point is the modeling of neuronal circuits. Network models that incorporate neuroanatomy and physiology are necessarily very complicated because of the large numbers of neurons (on the order of 10^{11} in the human cerebral cortex), the many neuron types, their detailed and varied modes of interactions, not to mention the complex wiring, with intra/inter-laminar connections, and inter-areal connections with multiple feedforward and feedback loops.

This level of complexity implies (i) any realistic model will contain a large number of unknown parameters; (ii) *a priori* constraints for many of these parameters are hard to obtain, and (iii) simulation time is long, limiting the number of training sets possible. The issues above exacerbate one another. For example, when parameter space has dimension $d \gg 1$, a search domain that is k times larger in each dimension will result in a volume that is k^d times larger; and if the actual physiological domain is small relative to the search domain, then with high probability, a reasonable-sized sample will not contain a single point in the actual physiological domain.

In Fig. 8 we used a parameter domain $\mathscr{P}_{\mathrm{L}}$ with $k \approx 5$ compared to $\mathscr{P}$, the domain used in Results. Using a training set of 40000 points sampled randomly from $\mathscr{P}_{\mathrm{L}}$, it was very likely that none was $\mathscr{P}$. This figure shows, however, our well-trained DNN still achieved a good accuracy of $\sim$ 1 Hz. Compared to Fig. 1b, a larger training set was needed, and the accuracy was lower, but it performed satisfactorily nevertheless.

In "Results" ("Viability of parameters and DNN performance") and again in Fig. 8, the reason why small training sets sufficed was the simplicity of the mapping from input to output, a fact we confirmed in subsequent sections. Obviously one cannot conclude from this one study that such mappings always have simple structures, but modeling experience of the authors suggests that even in large-scale biologically realistic network models (e.g.[6]) neuronal responses tend to depend fairly smoothly on parameters. This means that locally in parameter

space, the dependence of target mappings on parameters is relatively simple, not unlike those revealed in our derivative analysis.

These observations offer hope to the feasibility of surrogate-based approaches for more complex neuronal circuit models. They also point to the need for good *a priori* bounds on physiological ranges to help simplify the structure of input-output maps, and this is where biology enters. The judicial use of biological facts and experimental data to partially constrain parameters in advance will increase the chances of success for machine-learning approaches.

We do not pretend to have a roadmap going forward, but our analysis has shown that DNN surrogates may have a role to play in complex biological modeling when used in conjunction with other techniques. We finish with a discussion of how this might work. A major obstacle to using surrogate modeling directly is the large number of parameters in complex biological models in relation to the relatively small training sets that can be obtained through simulations. In the local cortical network model studied in this paper, DNN surrogates performed well with smaller-than-expected training sets ("Results", "Performance of DNN surrogate"). This strong performance can be explained by the simplicity of the input-output map, a fact confirmed in our derivative analysis: firing rate (E or I) were shown to vary monotonically (increasing or decreasing), or were mostly indifferent, with respect to parameter increases in all but two or three instances. One cannot expect input-output maps in complex biological networks to always possess such simple structures, but some degree of regularity can be expected. In large cortical models, for example, we have found outputs to be fairly smooth due probably to the large numbers of neurons and the averaging effects of random noise. As smooth maps have relatively simple local structures dominated by their derivatives, this gives reason to hope that after (most) parameters have been localized to small enough intervals, DNN surrogates and the sensitivity analysis made possible by them can offer insight into properties of input-output relations.

In other words, we believe that surrogate methods can be useful when a priori bounds on parameters are known. This is not to downplay the challenges in locating such bounds, but it is a different kind of problem requiring different methods, such as leveraging information from biology, practicing smart parameter tuning (e.g. invoking experiments that involve as few parameters as possible to stabilize baseline values). Machine learning techniques such as evolution and genetic algorithms may also be useful at this stage. The more parameters one is able to localize and the better constrained they are, the more effective surrogate modeling techniques will be.

Materials and methods

We first describe the neuronal model that was used for illustration throughout the paper. Then, we define the deep neural network that was used as surrogate for this model. At last, we briefly introduce "Correlation analysis and logistic regression".

I&F neuronal model. In this work, we consider a homogeneously connected network of integrate-and-fire (I&F) neurons that can be thought of as a generic model of a local neuronal population. The network has $N_E = 225$ excitatory neurons (E-neurons) and $N_I = 75$ inhibitory neurons (I-neurons) with a ratio of $N_E/N_I = 3$. Each E-neuron is postsynaptic to another E-neuron with probability 10% and to an I-neuron with probability 50%. Each I-neuron is postsynaptic to any other neuron with probability 50%. These connection probabilities are consistent with those in the visual cortex; see[50] for supporting references. A single realization of the random graph with these connectivities was fixed and used throughout in our numerical experiments.

The dynamics of each neuron in the network is modeled by the I&F equation

$$\dot{V} = -\frac{1}{\tau_{\text{leak}}} V - (V - V_E) g_E - (V - V_I) g_I. \tag{1}$$

Here time is in milliseconds (ms) and V is the membrane potential normalized in a dimensionless unit with a reset value $V_R = 0$ and a spiking threshold $V_T = 1$, so that when V reaches V_T, the neuron fires a spike; then V is reset to V_R and will remain there for an absolute refractory period of 2.5ms. In these normalized units, $V_E = 14/3$ and $V_I = -2/3$ are excitatory and inhibitory reversal potentials, and $\tau_{\text{leak}} = 20$ms is the leak rate[51]. For any neuron n of type $Q \in \{E, I\}$, $g_E, g_I \geq 0$ are its excitatory and inhibitory conductances governed by

$$\tau_E \dot{g_E} = -g_E + \beta^{QE} S^{QE} \sum_{i=1}^{\infty} \delta(t - t_i^{\text{syn,E}}) + S^{QE} \sum_{i=1}^{\infty} \delta(t - t_i^{\text{ext,Q}}) + S^{\text{dr}} \sum_{i=1}^{\infty} \delta(t - t_i^{\text{dr}}), \tag{2}$$

$$\tau_I \dot{g_I} = -g_I + S^{QI} \sum_{i=1}^{\infty} \delta(t - t_i^{\text{syn,I}}), \tag{3}$$

where $\tau_E = 2$ms and $\tau_I = 3$ms are decay rates for excitatory and inhibitory conductances respectively. Synaptic inputs from other neurons within the network are described in the second terms on the right sides of Eqns (2) and (3): $\{t_i^{\text{syn,E}}\}_{i=1}^{\infty}$ and $\{t_i^{\text{syn,I}}\}_{i=1}^{\infty}$ are the spike times of all the E- and I-neurons presynaptic to neuron n, and $\delta(\cdot)$ is the dirac delta function indicating an instantanous jump of conductance g_E or g_I upon the arrival of an E or I-spike, with amplitude equal to $\beta^{QE} S^{QE}/\tau_E$ and S^{QI}/τ_I respectively. The quantity $S^{QE} \sum_{i=1}^{\infty} \delta(t - t_i^{\text{ext}})$ models the independent excitatory drive to neuron n from another region of the brain with Poisson kicks at rate $\eta^{\text{ext,Q}}$ arriving at times $\{t_i^{\text{ext,Q}}\}_{i=1}^{\infty}$. In addition, neuron n receives an independent Possion drive with strength $S^{\text{dr}} = 0.005$, rate η^{amb} and arrival times $\{t_i^{\text{dr}}\}_{i=1}^{\infty}$; this term is intended to represent "ambient" modulatory influences from other parts of the brain or body. Note that we do not model synapses individually, and to simulate the effect of

synaptic failure between E-neurons, at each spike a random number β^{EE} is picked from the uniform distribution on [0.8, 1]; we have set $\beta^{IE} = 1$, i.e., no synaptic failure for the synapses from E- to I-neurons is assumed.

The undetermined parameters of this model are the synaptic coupling weights among model neurons, S^{EE}, S^{EI}, S^{IE} and S^{II}, and inputs parameters to the population $\eta^{\text{ext,E}}$, $\eta^{\text{ext,I}}$ and η^{amb}.

Synaptic weights of real cortical neurons are not known, but physiologically plausible ranges can be estimated from a combination of indirect measurements (such as *in vitro* experiments and the firing rates of neurons) together with some analysis (see[3], Methods). In this paper, following Ref.[3], we will assume the physiologically plausible ranges to be

$$S^{EE} \in [0.02, 0.03], \qquad S^{EI}/S^{EE} \in [1.5, 3],$$
$$S^{IE}/S^{EE} \in [0.2, 0.5], \qquad S^{II}/S^{EI} \in [0.5, 1]\,.$$

We have chosen to normalize the other quantities by S^{EE} because it has been observed from parameter tuning (in e.g.[3]) that the 4 synaptic weights $S^{QQ'}$ can be adjusted up and down together without having a strong effect on the system; this point will be justified later on in our analysis. Note that S^{II} is normalized by S^{EI} with a ratio less than 1 to account for electrical coupling among I-neurons, which effectively weakens the self-inhibition of the I-population.

With regard to the input parameters, in this paper we will assume the plausible ranges are

$$\eta^{\text{ext,E}} \in [25, 3000]\,\text{Hz}, \qquad \eta^{\text{ext,I}}/\eta^{\text{ext,E}} \in [2, 6],$$
$$\eta^{\text{amb}} \in [1/3, 2/3] * 1200\,\text{Hz}.$$

The range for $\eta^{\text{ext,E}}$ is large as it is intended to include input strengths that range from spontaneous to strong drive, and we have coupled the drive to E and to I-neurons because most synaptic input will affect both. The quantity $\eta_0 = 1200$ Hz is the threshold for causing a neuron to spike in the absence of other inputs, and η^{amb} in real cortex is known to be below this threshold.

From here on, we will refer to the parameters above as $P = [P_{\text{S}},\ P_{\text{I}}]$, where

$$P_{\text{S}} = [S^{EE},\ S^{EI}/S^{EE},\ S^{IE}/S^{EE},\ S^{II}/S^{EI}]$$

are network synaptic parameters and

$$P_{\text{I}} = [\eta^{\text{amb}}/\eta_0,\ \eta^{\text{ext,E}},\ \eta^{\text{ext,I}}/\eta^{\text{ext,E}}]$$

are input parameters, and we will say $P = [P_{\text{S}},\ P_{\text{I}}]$ is in our physiological domain $\mathscr{P}$, if all 7 parameters fall within the ranges above.

Given P, we let r^E and r^I denote the mean firing rates of the E- and I-populations at steady state, and our model output is taken to be

$$O = [r^E,\ r^I]\,.$$

Model firing rates are computed through numerical simulation. In our simulations, each trial runs for 3s, the last 2s of which are used to compute the system's (empirical) firing rates. We assume, based on physiological experiments, that in an active state of the cortex,

$$r^E \in [5, 30]\,\text{Hz}, \qquad r^I/r^E \in [2.5, 5.5]\,,$$

and we will say O is in our physiological domain $\mathscr{O}$ if both r^E and r^I fall in the ranges above.

We reiterate that $\mathscr{P}$ consists of *a priori* biological constraints either deduced from indirect experimental measurements or learned from previous modeling results. It is necessary to partially constrain parameter space, and these are effectively educated guesses. The domain $\mathscr{O}$ consists of firing rates that correspond roughly to what is observed in the laboratory under a variety of circumstances. There is no guarantee whatsoever that $P \in \mathscr{P}$ will produce $O \in \mathscr{O}$.

DNN surrogate. First we review the general setup for a DNN. For the regression problem of fitting a training dataset $\{(x_i; y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}^{d'}$ for each i, a fully connected DNN of H layers, $H \geq 2$, is defined as follows. Let $h_j^{[l]}$ be the output of the jth node of the lth layer of the DNN. Then

$$h_j^l(x) = \sigma\,(W_j^{[l]} \cdot h^{[l-1]}(x) + b_j^{[l]})\,, \quad j = 1, \ldots, m_l\,,$$

where $x \in \mathbb{R}^d$, m_l is the number of neurons in layer l ($m_1 = d$, $m_H = d'$), $b_j^{[l]} \in \mathbb{R}$, $W_j^{[l]} \in \mathbb{R}^{m_{l-1}}$, and $h^{[l-1]}(x) = [h_j^{[l-1]}(x)]_{j=1}^{m_{l-1}} \in \mathbb{R}^{m_{l-1}}$. For the jth neuron of the output layer of the DNN,

$$h_j^{[H]}(x) = W_j^{[H]} \cdot h^{[H-1]}(x) + b_j^{[H]}\,.$$

The DNN is abbreviated as $h(x; \theta) = h^{[H]}(x)$, where

$$\theta = [W^{[H]}, b^{[H]}, W^{[H-1]}, b^{[H-1]}, \ldots, b^{[1]}]$$

is the set of parameters of the DNN. In this work, the activation σ is fixed to the sigmoid function, i.e., $\sigma(s) = 1/(1 + e^{-s})$. The loss function is fixed to the mean-square error (MSE)

$$L(\theta) = \sum_{i=1}^{n} \big(h(x_i;\theta) - y_i\big)^2.$$

During training, the parameters of the DNN in each epoch t can be updated using gradient descent as

$$\theta^{t+1} = \theta^t - \alpha \nabla_\theta L(\theta^t),$$

where α is the learning rate. To speed up the training process, we use a popular accelerated gradient-based optimizer of Adam in our experiments[24].

Here is how the DNN will be used in this work: We train a sigmoid-DNN $h(x;\theta)$ of hidden layer sizes $800-200-200$ on training dataset $\mathscr{D}_{\text{train}}^{n} = \{(P_i; O_i)\}_{i=1}^{n}$ obtained from n trials of simulations (for various values of n), where each P_i is randomly drawn from a uniform distribution in its physiological domain $\mathscr{P}$. The accuracy of the DNN $h(\cdot;\theta_n)$, where θ_n is the weight of DNN well-trained on $\mathscr{D}_{\text{train}}^{n}$, is evaluated on a testing dataset $\mathscr{D}_{\text{test}}$ consisting of 10000 (P, O)-pairs where P was drawn independently from $\mathscr{P}$ and O was computed from simulations. Mean-absolute error (MAE) defined as $\frac{1}{n}\sum_{i=1}^{n} ||h(P_i;\theta_n) - O_i||_1$ and root-mean-square error (RMSE) defined as $\sqrt{\frac{1}{n}\sum_{i=1}^{n} ||h(P_i;\theta_n) - O_i||_2^2}$ are used for accuracy quantification. A DNN trained on $\mathscr{D}_{\text{train}}^{20{,}000}$, donoted by $\hat{O}(P) = [\hat{r}^E(P), \hat{r}^I(P)] = h(P;\theta_{20,000})$, serves as a surrogate of the neuronal circuit for all later analysis.

We remark on the following known properties of the DNN that make it a powerful tool: (i) DNN is a universal approximator. It has been proved that a sufficiently wide neural network of at least one hidden layer can approximate any continuous function to any desired accuracy[52–54]. (ii) Empiricial and theoretical studies indicate that the DNN approach is free from the curse of dimensionality, i.e., error decay can be bounded by a scaling $\sim n^{-\frac{1}{2}}$ independent of the input dimension[18,19]. (iii) It has been observed in practice that DNNs in general do not overfit even in an overparameterized setting without explicit regularization[55]. Non-overfitting combined with the universal approximation property makes DNN a highly robust and flexible approach for capturing general nonlinear mappings. (iv) It has been shown by the discovery of the Frequency Principle that DNNs are especially effective in learning low frequency functions from training data[20,21,56,57]. Therefore, very good accuracy can be achieved if the target mapping is dominated by low frequencies.

The evaluation of output using DNN is extremely efficient, especially when a large batch of input parameters is passed all at once to the DNN to best exploit the parallel computing capability of GPU. For our DNN of size $800-200-200$, evaluation of 10000 inputs takes $\sim$ 1 s on Nvidia GTX1080 using Tensorflow. The evaluation of outputs using simulation is much slower. A 3 s simulation of our 300-neuron network takes $7 \sim 10$ s on Intel i7 6800K using Brian2. Our simulation of neuronal networks can be speeded up with better optimization for parallel computing, but it is impossible to close such a gap of over 10^4 in efficiency difference. For a more realistic neuronal network of over 10000 neurons, the gap in efficiency will be much larger.

It is widely known that the choice of DNN architecture and hyperparameters can have a large impact on the training and generalization performance of a DNN. Because the mapping we consider in this work lacks structures that can take advantage of architectures like CNN or RNN, we have used a vanilla fully-connected network. Empirically, we found that the depth of the network is not crucial for our problem; however, a moderate depth, say 4 layers as used in this work, can help accelerate the training and reduce the requirement of width, i.e., number of neurons in each layer. In addition, the performance of the DNN is not sensitive to width as long as the network is sufficiently overparameterized, i.e., the number of parameters is larger than the size of training samples, to ensure a very low training error. In this work, we found a mysterious dependence of DNN performance on the scale of output. The test error of our DNN can be over 2 times larger if we scale the output by a factor of 0.01. This phenomenon is currently poorly understood in both theoretical and experimental studies of DNN and is out of the scope of this paper. For the optimization algorithm, we have stuck to Adam, which significantly improves the convergence rate during the training in comparison to gradient descent. In general, hyperparameter search can improve the training efficiency and test accuracy of DNN, though that is not crucial for the present study.

Finally, as noted in the Introduction, other machine learning approaches like support vector regression (kernel method) and gaussian process regression (kriging) may also serve our purposes of surrogate modeling. However, we anticipate that, for more complex biological networks, the flexibility of DNN surrogates may be a great advantage in application.

Correlation analysis and logistic regression. Correlation between two variables $x_i, x_j \in \{-1, 1\}$ is defined by

$$c_{ij} = \frac{\mathbb{E}(\tilde{x}_i \tilde{x}_j)}{\sqrt{\mathbb{E}(\tilde{x}_i^2)\mathbb{E}(\tilde{x}_j^2)}} \in [-1, 1].$$

where $\tilde{x} = x - \mathbb{E}(x)$. $|c_{ij}|$ is also a good indicator of how accurate x_i and x_j can predict one another.

Logistic regression solves a classification problem as follows. Model $f(x;\theta) = 1/(1 + e^{-(\mathbf{a}\cdot x_i + b)})$ with $\theta = [\mathbf{a}, b]$ is fitted to data $\{(x_i \in \mathbb{R}^d, y_i \in \{0, 1\})\}_{i=1}^{n}$ by maximizing the log-likelihood function, i.e.,

$$\theta^* = \max_{\theta} \Sigma_{i=1}^{n} \big[y_i \log f(x_i;\theta) + (1 - y_i)\log(1 - f(x_i;\theta))\big].$$

Then, for any x, if $f(x;\theta^*) > 0.5$, the output is predicted as 1, otherwise as 0. A high prediction accuracy indicates that input domains correspond to different outputs are linearly separable, whereas low prediction accuracy

($\lesssim$ 50%) indicates a complex structure not linearly separable. In Results, to use logistic regression for the prediction of sign of derivatives, we map positive sign to 1, negative sign to 0 and solve the optimization problem above.

Received: 13 May 2020; Accepted: 20 October 2020
Published online: 18 November 2020

References

1. McLaughlin, D., Shapley, R., Shelley, M. & Wielaard, D. J. A neuronal network model of macaque primary visual cortex (V1): orientation selectivity and dynamics in the input layer 4C. *Proc. Natl. Acad. Sci.* **97**, 8087–8092 (2000).
2. Markram, H. *et al.* Reconstruction and simulation of neocortical microcircuitry. *Cell* **163**, 456–92 (2015).
3. Chariker, L., Shapley, R. & Young, L.-S. Orientation selectivity from very sparse LGN inputs in a comprehensive model of macaque V1 cortex. *J. Neurosci.* **36**, 12368–12384 (2016).
4. Schmidt, M. *et al.* A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas. *PLOS Comput. Biol.* **14**, e1006359 (2018).
5. Billeh, Y. N. *et al.* Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *Neuron* **106**, 388-403.e18 (2020).
6. Chariker, L., Shapley, R. & Young, L.-S. Contrast response in a comprehensive network model of macaque V1. *J. Vis.* **20**(4), 16. https://doi.org/10.1167/jov.20.4.16 (2020).
7. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
8. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, 2016).
9. Carleo, G. *et al.* Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).
10. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
11. Bhosekar, A. & Ierapetritou, M. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. *Comput. Chem. Eng.* **108**, 250–267 (2018).
12. Razavi, S., Tolson, B. A. & Burn, D. H. Review of surrogate modeling in water resources. *Water Resour. Res.*https://doi.org/10.1029/2011WR011527 *(2012).*
13. Sun, G. & Wang, S. A review of the artificial neural network surrogate modeling in aerodynamic design. Proc. Inst. *Mech. Eng. Part G J. Aerosp. Eng.* **233**, 5863–5872 (2019).
14. Pruett, W. A. & Hester, R. L. The Creation of Surrogate Models for Fast Estimation of Complex Model Outcomes. *PLOS ONE* **11**, e0156574 (2016).
15. Renardy, M., Yi, T.-M., Xiu, D. & Chou, C.-S. Parameter uncertainty quantification using surrogate models applied to a spatial model of yeast mating polarization. *PLOS Comput. Biol.* **14**, e1006181 (2018).
16. Schuecker, J., Schmidt, M., Albada, S. J., Diesmann, M. & Helias, M. Fundamental activity constraints lead to specific interpretations of the connectome. *PLOS Comput. Biol.* **13**, e1005179 (2017).
17. Bahuguna, J., Tetzlaff, T., Kumar, A., Kotaleski, J. H. & Morrison, A. Homologous Basal Ganglia network models in physiological and Parkinsonian conditions. *Front. Comput. Neurosci.* **11**, 79 (2017).
18. E, W., Ma, C. & Wu, L. On the Generalization properties of minimum-norm solutions for over-parameterized neural network models. arXiv:1912.06987 (2019).
19. E, W., Ma, C. & Wu, L. Machine learning from a continuous viewpoint. arXiv:1912.12777 (2019).
20. Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y. & Ma, Z. Frequency principle: fourier analysis sheds light on deep neural networks. arXiv :1901.06523 (2019).
21. Zhang, Y., Xu, Z.-Q. J., Luo, T. & Ma, Z. Explicitizing an implicit bias of the frequency principle in two-layer neural networks. *arXiv preprint*arXiv:1905.10264 *(2019).*
22. Wang, H., Zhang, L. & Han, J. DeePMD-kit: a deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).
23. Zhang, L., Han, J., Wang, H., Car, R. & E, W. DeePCG: constructing coarse-grained models via deep neural networks. *The J. Chem. Phys.* **149**, 034101 (2018).
24. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *arXiv preprint*arXiv:1412.6980 *(2014).*
25. Swadlow, H. A. Efferent neurons and suspected interneurons in binocular visual cortex of the awake rabbit: receptive fields and binocular properties. *J. Neurophysiol.* **59**, 1162–1187 (1988).
26. Cardin, J. A., Palmer, L. A. & Contreras, D. Stimulus feature selectivity in excitatory and inhibitory neurons in primary visual cortex. *The J. Neurosci.* **27**, 10333–10344 (2007).
27. Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J. & McNaughton, B. L. Paradoxical effects of external modulation of inhibitory interneurons. *J. Neurosci.* **17**, 4382–4388 (1997).
28. Ozeki, H., Finn, I. M., Schaffer, E. S., Miller, K. D. & Ferster, D. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron* **62**, 578–592 (2009).
29. Murphy, B. K. & Miller, K. D. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron* **61**, 635–648 (2009).
30. Mahrach, A., Chen, G., Li, N., van Vreeswijk, C. & Hansel, D. Mechanisms underlying the response of mouse cortical networks to optogenetic manipulation. *eLife* **9**, e49967 (2020).
31. Kato, H. K., Asinof, S. K. & Isaacson, J. S. Network-level control of frequency tuning in auditory cortex. *Neuron* **95**, 412–423 (2017).
32. Moore, A. K., Weible, A. P., Balmer, T. S., Trussell, L. O. & Wehr, M. Rapid rebalancing of excitation and inhibition by cortical circuitry. *Neuron* **97**, 1341-1355.e6 (2018).
33. Garcia del Molino, L. C., Yang, G. R., Mejias, J. F. & Wang, X.-J. Paradoxical response reversal of top-down modulation in cortical circuits with three interneuron types. *eLife* **6**, e29742 (2017).
34. Litwin-Kumar, A., Rosenbaum, R. & Doiron, B. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *J. Neurophysiol.* **115**, 1399–1409 (2016).
35. Sadeh, S., Silver, R. A., Mrsic-Flogel, T. D. & Muir, D. R. Assessing the role of inhibition in stabilizing neocortical networks requires large-scale perturbation of the inhibitory population. *J. Neurosci.* **37**, 12050–12067 (2017).
36. Angelucci, A. *et al.* Circuits and mechanisms for surround modulation in visual cortex. *Annu. Rev. Neurosci.* **40**, 425–451 (2017).
37. Sceniak, M. P., Hawken, M. J. & Shapley, R. Visual spatial characterization of Macaque V1 neurons. *J. Neurophysiol.* **85**, 1873–1887 (2001).
38. Albrecht, D. G. & Hamilton, D. B. Striate cortex of monkey and cat: contrast response function. *J. Neurophysiol.* **48**, 217–237 (1982).
39. Rubin, D., VanăHooser, S. & Miller, K. The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron* **85**, 402–417 (2015).
40. Vreeswijk, C. & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**, 1724–1726 (1996).
41. Vreeswijk, C. & Sompolinsky, H. Chaotic balanced state in a model of cortical circuits. *Neural Comput.* **10**, 1321–1371 (1998).

42. Vogels, T. P., Rajan, K. & Abbott, L. Neural network dynamics. *Annu. Rev. Neurosci.* **28**, 357–376 (2005).
43. Harish, O. & Hansel, D. Asynchronous rate chaos in spiking neuronal circuits. *PLOS Comput. Biol.* **11**, e1004266 (2015).
44. Henrie, J. A. & Shapley, R. LFP power spectra in V1 cortex: the graded effect of stimulus contrast. *J. Neurophysiol.* **94**, 479–490 (2005).
45. Chariker, L., Shapley, R. & Young, L.-S. Rhythm and Synchrony in a Cortical Network Model. *J. Neurosci.* **38**, 8621–8634 (2018).
46. Sederberg, P. B., Kahana, M. J., Howard, M. W., Donner, E. J. & Madsen, J. R. Theta and gamma oscillations during encoding predict subsequent recall. *J. Neurosci.* **23**, 10809–10814 (2003).
47. Gonzalez-Burgos, G., Hashimoto, T. & Lewis, D. A. Alterations of cortical GABA neurons and network oscillations in Schizophrenia. *Curr. Psychiatry Rep.* **12**, 335–344 (2010).
48. McCarthy, M. M., Ching, S., Whittington, M. A. & Kopell, N. Dynamical changes in neurological diseases and anesthesia. *Curr. Opin. Neurobiol.* **22**, 693–703 (2012).
49. Young, L. Towards a mathematical model of the brain. *J. Stat. Phys.* **180**, 612–629. https://doi.org/10.1007/s10955-019-02483-1 (2020).
50. Chariker, L. & Young, L.-S. Emergent spike patterns in neuronal populations. *J. Comput. Neurosci.* **38**, 203–220 (2015).
51. Koch, C. *Biophysics of Computation: Information Processing in Single Neurons* (Oxford University Press, Oxford, 2004).
52. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **2**, 303–314 (1989).
53. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991).
54. Leshno, M., Lin, V. Y., Pinkus, A. & Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **6**, 861–867 (1993).
55. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint* arXiv:1611.03530 (2016).
56. Xu, Z.-Q. J., Zhang, Y. & Xiao, Y. Training behavior of deep neural network in frequency domain. *arXiv preprint* arXiv:1807.01251 (2018).
57. Rahaman, N. *et al.* On the Spectral Bias of Deep Neural Networks. *arXiv preprint* arXiv:1806.08734 (2018).

Acknowledgements

Yaoyu Zhang did most of this work at the Institute for Advanced Study supported by NSF Grant No.DMS-1638352 and the Ky Fan and Yu-Fen Fan Membership Fund. Lai-Sang Young was supported in part by NSF Grants 1734854 and 1901009. The authors would like to thank David Hansel, Aaditya Rangan and Robert Sharpley for valuable comments and suggestions.

Author contributions

Y.Z. and L.Y. conceived the method and experiments, analysed the results, prepared and reviewed the manuscript. Y.Z. wrote the software codes.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Z. or L.-S.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.