Fine-Grained System Identification of Nonlinear Neural Circuits

Dawna Bagherian
California Institute of Technology
Pasadena, CA, USA
dawna@caltech.edu

James Gornet
California Institute of Technology
Pasadena, CA, USA
jgornet@caltech.edu

Jeremy Bernstein
California Institute of Technology
Pasadena, CA, USA
bernstein@caltech.edu

Yu-Li Ni
California Institute of Technology
Pasadena, CA, USA
ynni@caltech.edu

Yisong Yue
California Institute of Technology
Pasadena, CA, USA
yyue@caltech.edu

Markus Meister
California Institute of Technology
Pasadena, CA, USA
meister@caltech.edu

ABSTRACT

We study the problem of sparse nonlinear model recovery of high dimensional compositional functions. Our study is motivated by emerging opportunities in neuroscience to recover fine-grained models of biological neural circuits using collected measurement data. Guided by available domain knowledge in neuroscience, we explore conditions under which one can recover the underlying biological circuit that generated the training data. Our results suggest insights of both theoretical and practical interests. Most notably, we find that a sign constraint on the weights is a necessary condition for system recovery, which we establish both theoretically with an identifiability guarantee and empirically on simulated biological circuits. We conclude with a case study on retinal ganglion cell circuits using data collected from mouse retina, showcasing the practical potential of this approach.

CCS CONCEPTS

• **Networks** → *Network structure*; • **Applied computing** → *Systems biology*; **Biological networks**; • **Computing methodologies** → *Modeling methodologies*; *Neural networks*.

KEYWORDS

nonlinear system identification, neural networks, neuroscience

1 INTRODUCTION

Sparse system estimation or identification is a core problem in science and engineering [22, 33, 34, 61], where the goal is to recover from training data a sparse set of non-zero parameters in a high-dimensional function class. Examples include estimating dynamical systems [16, 33], graphical models or inverse covariance matrices [34, 61], and linear effects models [22]. Thus far, the bulk of prior work has focused on linear models, which have limited applicability when the underlying system is inherently nonlinear.

In many domains there is a growing need for reliably estimating a sparse nonlinear system from data. We are particularly motivated by applications to circuit neuroscience, where the goal is to estimate the fine-grained structure and connection weights of a neural circuit. A typical data collection approach is to apply stimuli to the sensory neurons in the input layer of the circuit, and measure the responses of the output neurons [38]. Recent advancements in biotechnology have yielded increasingly powerful methods for collecting neuronal data, including genetic tools for labeling neurons [5], dense electrode arrays for recording action potentials from living cells [24], genetically encoded fluorescent activity reporters [9], and complex surgical procedures to provide access to specific brain regions [14].

Motivated by the neuroscience modeling application, we aim to estimate a sparse nonlinear model from a high-dimensional compositional function class, i.e., a deep neural network. The key question we ask is: *under what conditions can one recover the parameters of a sparse neural network?* To keep our inquiry practically grounded, we restrict to exploring conditions that abstractly capture conventional neuroscience domain knowledge.

In this paper, we study conditions for successful sparse parameter recovery of a nonlinear composition function with skip connections. Notably, we show that having each layer of weights be sign constrained is a necessary condition (in addition to other more standard necessary conditions). We establish this result both theoretically with an identifiability guarantee for a specific class of neural networks, and empirically on a broader class of simulated biological circuits. Our experiments show a striking dichotomy: ℓ_1-regularized regression fails to recover the correct parameters without sign constraints, and succeeds with them. Our results also provide guidance as to the dependence of system identification on the training dataset, the structure of the circuit under study, and the quantity of domain foreknowledge provided to the neural network.

We conclude with a case study of the mouse retina. We measure responses to visual input from four types of retinal ganglion cells in the live mouse retina using a multi-electrode array [38]. We confirm the current scientific understanding of the structure of one of these four and suggest an additional structural component. We also provide a hypothesis as to the structure of the other three

types, which can later be confirmed with biological experiments. These results showcase the practical potential of this approach.

2 RELATED WORK

System identification and structure recovery. Broadly speaking, system identification is the use of statistical methods to build models of systems from measured data [33]. System identification is a key tool in modeling dynamical systems, which includes early work on neural system identification for control systems [29]. A fundamental issue that arises is *identifiability* [8]—that is, when can one uniquely recover the true system. This affects the reliability of the results for downstream scientific analysis.

Identifiability has been studied theoretically for nonlinear neural networks [3, 13, 44, 46, 52, 57], "linear networks" in the context of matrix factorization [10, 23, 30], and linear autoregressive models [16]. In Section 4, we introduce and discuss some of the theoretical concepts that have a bearing on identifiability. As stated before, our main goal is to provide a thorough theory-to-practice investigation grounded in real neuroscience modeling challenges.

A related concept is support recovery, where the main goal is to discover which parameters of a model are non-zero [22, 31, 48, 51, 54]. Support recovery can be thought of as a subgoal of full system identification. It is commonly studied in sparse linear systems that have few non-zero parameters. While biological neural networks are also sparse [19, 39], they are nonlinear multi-layer models for which theoretical results in sparse linear support recovery do not directly apply. Nonetheless, we show that, under suitable conditions, one can employ ℓ_1-regularized regression (that is commonly used for estimating sparse linear models [54]) to reliably estimate sparse neural networks with limited training data; an interesting future direction would be to establish *sparsistency* guarantees [31].

Another related concept is structure discovery in (causal) graphical models [34, 45, 50, 55, 61]. A typical setting is to recover the structure of a directed acyclic network (i.e., which edges are non-zero) that forms the causal or generative model of the data. This setting is very similar to ours with a few differences. First, the goal of structure discovery in graphical models is to recover the direction of the edges in addition to the weights, whereas in our setting all the edge directions are known a priori. Second, the training data for structure discovery is typically fully observed in terms of measuring every node in the network, whereas for our setting we only observe the inputs and final outputs of the network (and not the measurements of nodes in the hidden layers). Like in sparse system identification, most prior work in structure discovery of graphical models is restricted to the linear setting.

Interplay between neuroscience and deep learning. The interplay between neuroscience and deep learning has a rich history that ebbs and flows between one field driving progress in the other. For instance, convolutional neural networks (CNNs) are originally inspired by biology [17], and were designed to have computational characteristics of the vertebrate visual system. More recently, researchers have endeavored to understand biological, and specifically neuronal, systems using artificial neural networks [2, 28, 35, 36, 46, 47, 53, 59]. Such work has been largely restricted to coarse-grained analyses that characterize computation of regions of the brain, rather than that of individual neurons interacting within

a circuit, or to the revelation of microcircuit motifs within trained CNNs that mirror those found in biology.

In contrast, we are interested in fine-grained circuit modeling where the neurons in the learned neural network have a one-to-one correspondence with individual biological units. Preliminary efforts in this direction have been made for a smaller model of one layer of retinal synapses [46], but to the best of our knowledge, this work covers only parameter estimation, not structure estimation, and the general problem has not been studied theoretically or empirically via systematic simulations, nor has it been tested on real data from deeper circuits.

3 PROBLEM STATEMENT

We formulate the problem of system identification in nonlinear feedforward networks as follows. Consider a function $f(x; W, b)$ known as the *network*. The network is parameterized by an unknown weight vector W and bias b. We assume that we can query the network's nonlinear input-output mapping:

$$x \mapsto f(x; W, b) \equiv y. \tag{1}$$

That is, we may apply input stimuli x and record the output y. Section 6 describes how we collect such data from the mouse retina.

Our goal is to recover the true W and b given such (x, y) queries. A first question that arises is whether recovering the true W and b is possible, even with infinite data—i.e., whether W and b are uniquely identifiable. Other questions include how to accurately recover W and b (or at least their non-zero support) given finite and noisy data, which data points to query, and what types of domain knowledge can aid in this process.

In practice, estimating the true W and b given training data is tackled as a regression problem. We are particularly interested in the case where W is sparse. The conventional way to encourage sparsity is to use ℓ_1-regularization [54], which we will also employ. We discuss in Section 5 practical considerations through extensive evaluation of simulated biological circuits.

Summary of Domain Knowledge. Our goal is to not only establish conditions where system identification of nonlinear feedforward networks is possible, but also that those conditions be practically relevant. Guided by neuroscience domain knowledge, we study circuits with the following properties:

(i) The nonlinearity of individual neurons is well understood, and can be well modeled using ReLUs or leaky ReLUs [6].
(ii) The dominant computation of the neural circuit is feedforward, which is true for circuits found in the retina [19, 58].
(iii) Many potential weights can be preemptively set to zero, due to the implausibility of neurons in various spatial configurations being connected to one another [11, 49].
(iv) All weights can be sign constrained (either non-negative or non-positive), due to knowing the inhibitory or excitatory behavior of each neuron [19, 42, 58].

Items (i) & (ii) above imply knowing the functional form (i.e., multilayer perceptron with known number of layers, maximal number of neurons in each layer, and form of the nonlinearity). Item (iii) implies that one can reduce the number of free parameters in the model, thus easing the burden of learning (although still requiring high-dimensional sparse estimation). Item (iv) is perhaps the most interesting property, as it effectively constrains the weights to be
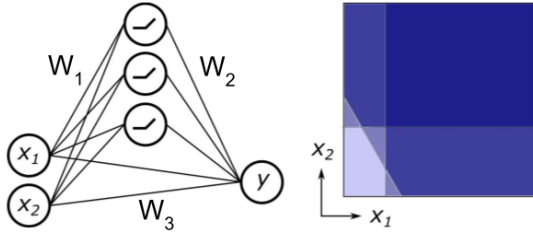
Figure 1: Left: A one hidden layer ReLU network with skip connection (Equation 2). Right: The nonlinearity transitions occur on hyperplanes in input space.

within a single known orthant. We show in Section 4 that the sign constrained condition is an important sufficient condition for proving identifiability, and we show empirically in Section 5 that ℓ_1-regularized regression can succeed in system identification on sign constrained networks and fails on unconstrained networks.

4 THEORETICAL ANALYSIS

Motivated by the neuroscience domain knowledge discussed in Section 3, we now establish factors that govern fine-grained identifiability of neural networks—a topic that has been studied since the 1990s [3, 13, 52]. To develop the core ideas, it will help to consider the following neural network:

$$f(x) := W_2 \max(0, W_1 x + b_1) + W_3 x + b_2, \tag{2}$$

for weights $W_1 \in \mathbb{R}^{n_1 \times n_0}$, $W_2 \in \mathbb{R}^{n_2 \times n_1}$, $W_3 \in \mathbb{R}^{n_2 \times n_0}$ and biases $b_1 \in \mathbb{R}^{n_1}$, $b_2 \in \mathbb{R}^{n_2}$.

There are three reasons why this is a sensible network to consider. First, the network's relatively simple structure facilitates theoretical insight. Second, this network bears a resemblance to neural (sub-)circuits found in the retina (see Fig. 2). And third, if one ignores the skip connections, biases, and nonlinearity, then the identification problem reduces to factorization of $W_2 W_1$, allowing us to compare to results on matrix factorization [10, 23, 30].

We shall now provide a result on identifiability for this class of nonlinear networks. The proof of Theorem 1 is given in Appendix A.4.

THEOREM 1. *Suppose that network (2) satisfies the following three conditions:*
  (i) no column of W_2 or row of W_1 is entirely zero;
  (ii) no two rows of W_1 are collinear;
  (iii) W_1 is nonnegative.
Let P denote an unknown permutation matrix and D an unknown positive diagonal matrix. Then, by input-output queries of the form (1), we may recover DPW_1, DPb_1, $W_2 P^{-1} D^{-1}$, W_3 and b_2.

The appearance of permutation matrix P reflects the invariance of a two-layer network to permutations of its hidden units. Also, by positive homogeneity of the max function, the output synapses of a hidden unit (columns of W_2) may be scaled up by some $\alpha > 0$ provided that the input synapses (rows of W_1) are scaled down by $1/\alpha$. This gives rise to the diagonal matrix D. These symmetries are innate to two-layer systems—the same issue is present in matrix factorization [23, Definition 4].

This theorem is somewhat typical of the results in the identifiability literature. For example, in 1992 Sussmann [52] proved a similar result for a network with one hidden layer, tanh nonlinearity, and no skip connections. In contrast, our result applies to networks with skip connections and relu nonlinearity.

Discussion of preconditions and assumptions. Condition (i) imposes that every hidden unit must be connected to both the input and output of the circuit, and condition (ii) imposes that no hidden unit is computationally redundant with another. These conditions are intuitively important for identifiability. The most substantive preconditon is condition (iii), which imposes a sign constraint on the synapses at the first layer. In the neuroscience context, this would correspond to knowing that all of the synapses in the first layer are excitatory as is the case for the outgoing synapses of bipolar cells in the retinal circuit (see Fig. 2). This demonstrates that domain knowledge of neuronal cell types can simplify both the theoretical analysis as well as the recovery of the neural network's weights.

Our theoretical analysis has some notable limitations. For instance, our result requires querying the network on inputs x lying in the negative orthant, which is not biologically plausible in cases where x is modeling the output of a layer of rectified nonlinear neurons. Additional constraints on the bias b_1 would be required to guarantee identifiability given only non-negative input queries. Furthermore, our result makes no comment on sample complexity or dealing with non-convexity of the underlying optimization problem, which are all interesting directions for future work. We do, however, provide an extensive empirical study of such questions in Section 5.

Connection to identifiability in other systems. The preconditions of Theorem 1 are mild in comparison to results in the nonnegative matrix factorisation literature, which require strong sparsity conditions on W_1 to guarantee identifiability of $W_2 W_1$ [10, 23, 30]. Theorem 1 suggests that—surprisingly—the presence of nonlinearities can make system identification *easier*. This is because the location of the nonlinear thresholds in input space is what reveals the entries of b_1 and W_1 up to a scaled permutation. We illustrate this in Figure 1.

5 EMPIRICAL ANALYSIS & PRACTICAL CONSIDERATIONS

We now empirically analyze conditions under which ℓ_1-regularized regression can lead to successful fine-grained identification of neural circuits, and cross-reference with the theoretical results from Section 4 where appropriate. Practical system identification of neural networks can be challenging even with significant constraints imposed (e.g., many weights preemptively constrained to zero). Most applications of artificial neural networks (ANNs) are not concerned with identification of the "true" network, but rather focus on predictive performance (which can be achieved by many parameterizations). Therefore, it is important to understand, operationally, when system identification can be reliably performed.

Since we assume that the true neural circuit is sparse, we employ ℓ_1-regularized regression. Given a training set $S = \{(x, y)\}$, tuning parameter λ, neural circuit $f(\cdot; W)$, and set of allowable weights
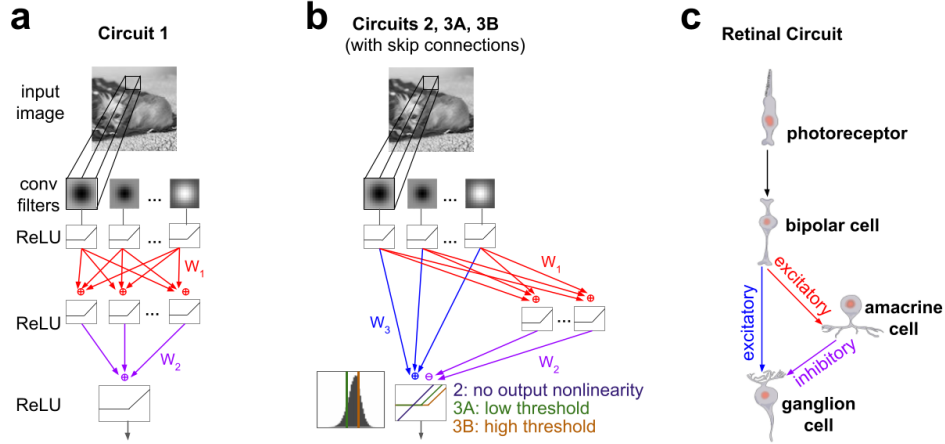
Figure 2: Networks studied in simulation. (a) Simulated circuit. (b) Three simulated circuits with skip connections, varying in the use of output nonlinearity. (c) The retinal circuit that inspired the models in (a) and (b). The photoreceptor layer is treated as linear, and the convolutional layers in (b) and (c) represent the combined computation of photoreceptors and bipolar cells.

$\mathcal{W}$, the optimization problem can be written as:

$$W_k \leftarrow \operatorname{argmin}_{W \in \mathcal{W}} \sum_{(\boldsymbol{x}, y) \in S} \|f(\boldsymbol{x}; W) - y\|^2 + \lambda |W|. \qquad (3)$$

Due to the nonconvex nature of this optimization, it is preferable to run it with multiple (K) random initializations, thereby producing a set of solutions $W_1, \dots, W_K$.

Using controlled simulated settings, we performed a systematic study to understand the dependence of successful system identification on the following factors:

(i) The design of the function class $f(\cdot; W, \boldsymbol{b})$, and in particular the presence of skip connections in the architecture.

(ii) The number of non-zero weights $W \in \mathcal{W}$, via preemptively setting many connections to zero.

(iii) The use of sign constraints on the weights $W \in \mathcal{W}$.

(iv) The use of ℓ_1 regularization to encourage weight sparsity.

(v) The design of the training dataset, i.e., which data points to query.

In each simulated experiment, we created a sparse oracle network, and then trained a second network to recover that structure via the optimization problem in Equation 3, while varying the constraints, regularizers and datasets.

We define two measures of success:

Definition 5.1 (System Recovery Score). Let the *system recovery score* be defined on the vectorized weight matrices as:

$$R_{\text{sys}} = 2 \times \frac{W_{\text{learned}} \cdot W_{\text{oracle}}}{|W_{\text{learned}}|^2 + |W_{\text{oracle}}|^2}. \qquad (4)$$

Definition 5.2 (Support Recovery Score). We will denote by $\overline{W}$ the vector of the same size as W where $\overline{W}_i = 0$ if $W_i = 0$ and 1 otherwise. Let the *support recovery score* be defined on the vectorized and binarized weight matrices as:

$$R_{\text{supp}} = 2 \times \frac{\overline{W}_{\text{learned}} \cdot \overline{W}_{\text{oracle}}}{|\overline{W}_{\text{learned}}|^2 + |\overline{W}_{\text{oracle}}|^2}. \qquad (5)$$

Both measures are related to cosine similarity and the Jaccard coefficient. The System Recovery Score achieves perfect performance

when the exact system is recovered with the exact weights. The Support Recover Score achieves perfect performance whenever the non-zero support (i.e., the structure) is recovered exactly.

5.1 Summary of findings

Fig. 3 shows our main findings on the practicality of the approach. For all simulated circuits, we are able to reliably recover most of the structure and parameters using ℓ_1-regularized regression, when there are around 500 non-zero, sign-constrained free parameters in the network. As we will see in Section 6, the biological problems we are interested in have a few hundred free parameters, and so are closely modeled by the darkest curves in Figures 3 and 4.

Fig. 4 shows that successful recovery can be contingent on the use of ℓ_1-regularization as well as sign constraints. In particular, we see that ℓ_1-regularization is especially crucial when the circuit has skip connections (as is common in the retina), and that the estimation problem fails completely without sign constraints. The former finding potentially suggests skip connections as an important factor for finite-sample analysis of system identification of nonlinear circuits, and the latter finding supports our theoretical finding that sign constraints are an important condition for guaranteeing identifiability.

5.2 Network architecture

We focus on the most generic network architectures that are prevalent in retinal biology—though similar architectures arise in gene regulatory networks [34]. We call these networks:

- Circuit 1: three layers, with the first layer convolutional and fixed, and the output layer nonlinear (Fig 2a).
- Circuit 2: three layers, with the first layer convolutional and fixed, the output layer linear, including skip connections (Fig 2b).
- Circuit 3A: Circuit 2 with the addition of a low-threshold ReLU output nonlinearity (Fig 2b).
- Circuit 3B: Circuit 2 with the addition of a high-threshold ReLU output nonlinearity (Fig 2b).

Note that Circuit 2 is, in fact, the circuit studied theoretically in Section 4, but with a restricted input domain. In the retina, one can think of the output of the first layer as modeling bipolar cell glutamate release rates, the second layer as modeling amacrine cells, and the third as modeling ganglion cells, though this "feedforward loop" circuit motif also appears in other domains, most prominently in transcriptional regulatory networks [4, 21].

For each of the four architectures, we constructed a specific oracle circuit by selecting a fixed set of weights. Circuit 1 had 36 nonzero synapses. Circuits 2, 3A and 3B had 96 nonzero synapses. For each of these four circuits, we generated a training set by presenting images to the circuit and recording the output responses. The rest of Section 5 will analyze recovery performance of these architectures while varying the other factors of interest.

5.3 Non-zero connection constraints are helpful for identification of some architectures

We first considered varying the number of free parameters in the ANN. Intuitively, with more free parameters, it will be harder to recover the true network. However, constraining to fewer parameters would require increasing amounts of domain knowledge.

Existing neuroscience domain knowledge constrains many connections to zero weight. In the retina, such connectivity constraints are implemented by the precisely organized anatomy of neurons. The so-called inner plexiform layer (IPL) is a meshwork of synapses between different unit types (bipolar, amacrine, and ganglion cells). Each type sends its axons and dendrites into a distinct lamina of the IPL, and neurons may connect only if they co-stratify in at least one lamina. This implements an "address book" of allowable connectivity between cell types in the network [49]. Using anatomical studies in the literature [7, 15, 18, 20, 26, 40, 56, 60], we have compiled such an address book for 36 retinal cell types (see appendix). A strong address book constraint translates to fewer free parameters.

For convenience, the first layer of weights (W_1) were also constrained so that each amacrine cell analogue in layer 2 only received input from a restricted spatial region of the output of layer 1. This is to remove permutation symmetry of W_1 and W_2, and it is also realistic, since it is known that amacrine cells in the retina typically pool inputs over a small region in space, not the entire visual field. Scaling symmetry of the weights was removed by fixing biases in place during training (see A.1.4 for details.)

Fig. 3 shows our main results, where we also constrain the sign of each weight based on whether the corresponding biological synapse should be excitatory or inhibitory. We see that all circuits can be approximately identified ($R_{sys} \geq 0.9$) given enough training data and if sufficiently constrained in the number of free parameters (around 500). Circuit 1, which is the only circuit without skip connections, always yields high system and support recovery (given enough training data) irrespective of the number of free parameters. However, the successful recovery of Circuits 2, 3A, and 3B exhibits a strong dependence on the number of free parameters. Our theory directly implies that Circuit 2 should be identifiable, which suggests that the non-convex optimization landscape of ANN training and the restricted input domain may be a significant factor here.
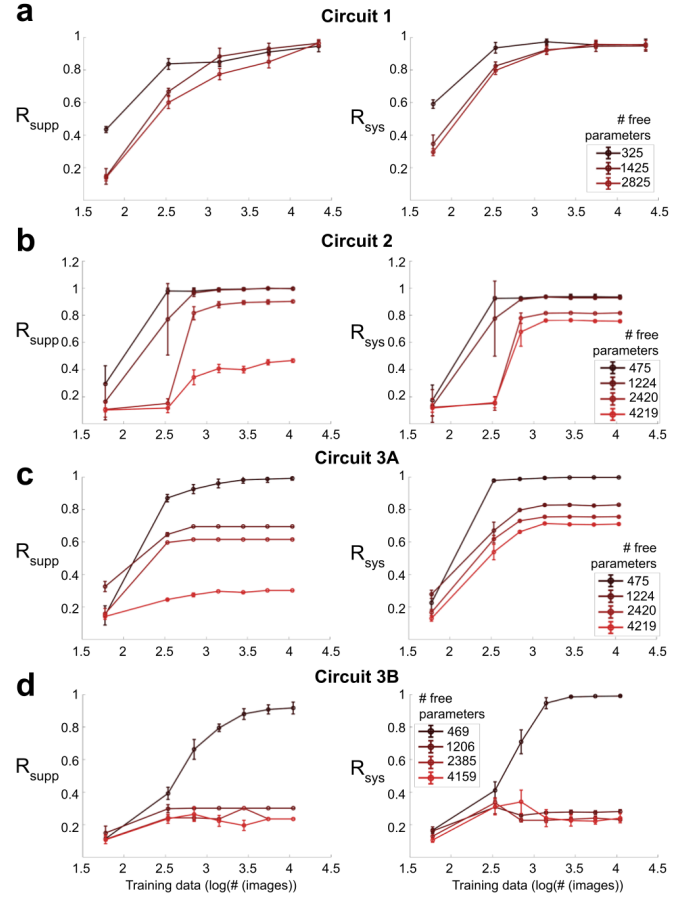

Figure 3: System and support recovery scores (Definitions 5.1 and 5.2) for Circuits 1, 2, 3A and 3B respectively, using ℓ_1-regularization and sign constraints. Each curve represents an ANN with 325–4219 free parameters (see legend). The ANN was given varying quantities of free parameters (weights associated with synapses) and training data. For each circuit structure, a single training dataset was generated, then networks with varying numbers of weights at initialization were given access to some fraction of the training set and trained. System and support recovery scores were computed in each case. Mean and standard deviation of these scores are plotted for 10 training runs, each with a different random initialization. One can reliably recover most of the system structure and parameters when there are around 500 free parameters (corresponding to putative synapses) in the system.

5.4 Sign constraints are necessary for system identification

Another way to constrain the network is to use sign constraints, due to abundant neuroscience domain knowledge about whether certain neurons are either excitatory or inhibitory [19, 42, 58]. In fact, for the retinal circuits we study in Section 6, we have sufficient knowledge to sign constrain every weight. Fig. 4 c&d shows the results of an ablation study that removes the sign constraints. We
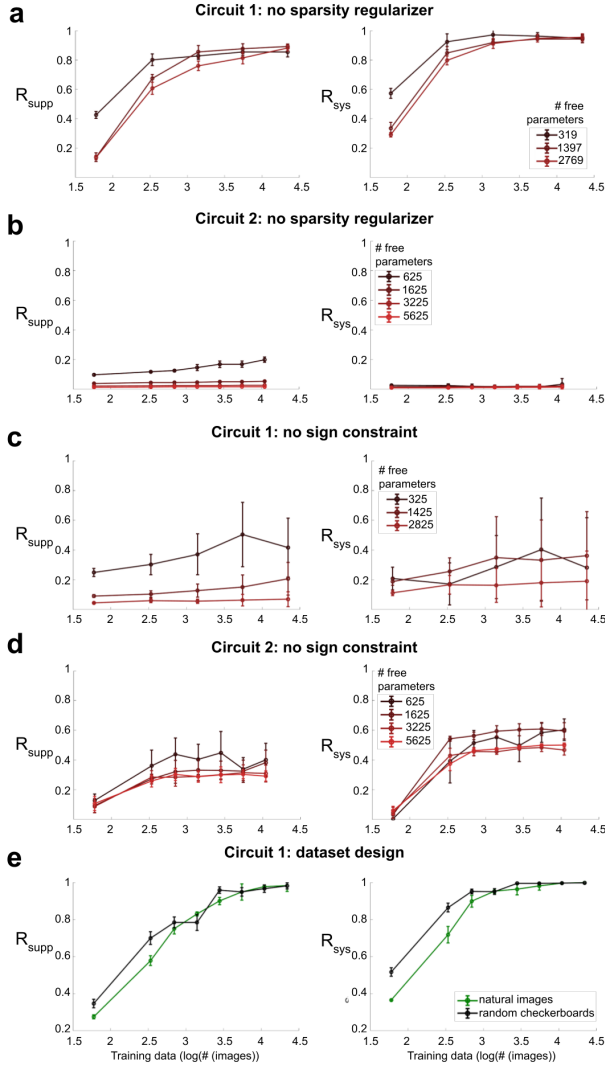
Figure 4: Results of ablation studies. Successful system identification can be contingent on regularization, sign constraint and dataset type. System identification for simulated retina-style circuits: (a) for Circuit 1 proceeds as normal without ℓ_1-regularization; (b) for Circuit 2 is impaired without ℓ_1-regularization; (c) for Circuit 1 is impaired without the sign constraint; (d) for Circuit 2 is impaired without the sign constraint. In (e) we see that random noise images lead to better system identification of Circuit 1 than natural images when data is scarce.

see a sharp contrast compared to Fig. 3 in that removing the sign constraints causes system identification via (3) to fail. This finding is consistent with our theoretical exploration and Theorem 1, where the sign constraint was a crucial element in identifying the half planes generated by the ReLU units.

5.5 Sparsity regularization is helpful for identification of some architectures

We next investigate the benefits of using ℓ_1-regularization in practice. Fig. 4 a&b show the results of an ablation study that omits the

use of ℓ_1-regularization. Compared to Fig. 3, we see that regularization never hurts system identification, and sometimes appears to be crucial in practice. In particular, one can still perform identification without ℓ_1-regularization for Circuit 1, which has no skip connections, but not for Circuit 2, which has skip connections. We hypothesize that a theoretical characterization of sample-efficient system identification of neural circuits could depend in a non-trivial way on the presence or absence of skip connections.

5.6 Dataset design matters when data is scarce

In visual neuroscience experiments one must choose among many possible images or movies with which to query the neural circuit to collect training data. Practically, biological experiments are time-limited, and therefore, the training dataset size is limited. Unsurprisingly, system identification of all four circuits showed a strong dependence on the size of the training dataset (Fig 3a-d).

Our final analysis in this section is to compare the efficacy of two very different data collection approaches: white-noise images vs natural photographs. White noise has a long history for system identification in engineering, whereas natural images better reflect the signal domain that the retina evolved to handle. White-noise stimuli were generated as random checkerboards. Each image was 100×100 pixels, and consisted of 5×5 grayscale checkers of random intensity. Natural grayscale images were taken from the COCO (Common Objects in Context) dataset [32]. An overparameterized artificial neural network with architecture matching Circuit 1 (Fig. 2a) was trained with varying quantities of data from both sets. When training data were plentiful, performance on the two datasets was comparable. However, when data were more scarce, random checkerboard stimuli led to better system identification (Fig. 4e). Interpolating, we find that a system recovery score of at least $R_{\text{sys}} = 0.9$ can be achieved with about 400 checkerboard images, but requires 630 natural images.

6 CASE STUDY ON THE MOUSE RETINA

We conclude this paper with a case study on the mouse retina. Our goal here is to demonstrate the practicality of fine-grained neural system identification in real biology settings. In many cases, one does not have the actual ground truth neural circuit, and so we evaluate by confirming that we can recover models that are consistent with known neuroscience domain knowledge.

We specifically studied four types of mouse retinal ganglion circuits: the sustained OFF, sustained ON, transient OFF, and transient ON subtypes of the alpha retinal ganglion cell (sOFFα, sONα, tOFFα, and tONα RGCs). The circuitry of interneurons that provide input to these cells is not yet known in retinal biology. However, there is strong evidence that the tOFFα cell pools excitatory input from OFF bipolar cells and inhibitory input from ON bipolar cells via amacrine cells [40]. These cell types present an interesting challenge for system identification. Namely: can we use this technique to (1) confirm what is known about the tOFFα cell circuit (2) provide new specific circuit hypotheses for further study?

We studied four hypothetical models of the α cell based on current domain understanding: the X-Cell Model, the Y-Cell Model, the XY-Cell Model, and the Push-Pull Model model (Fig. 5a-d). These
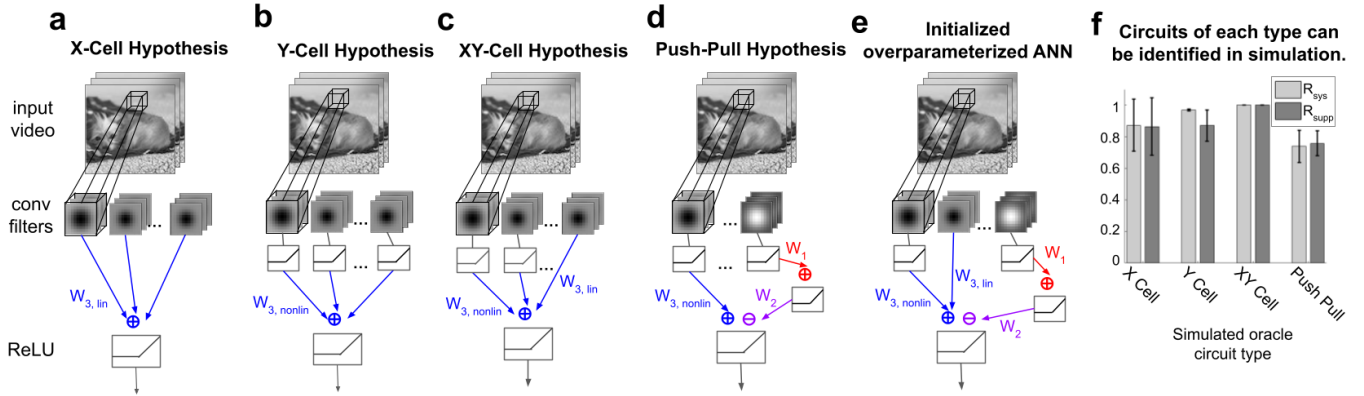
Figure 5: Hypothesized circuit architectures for the alpha ganglion cell. Each structure convolves the input video with a bank of spatiotemporal filters in the first layer. (a) X-Cell circuit: utilizes only linear units in the first layer. (b) Y-Cell circuit: utilizes only nonlinear units in the first layer. (c) XY-Cell circuit: utilizes both linear and nonlinear units in the first layer. (d) Push-pull circuit: utilizes nonlinear units in the first layer. Includes an additional inhibitory pathway. Structurally analogous to circuits 3A and 3B in Fig. 2b. (e) The ANN is initialized to include all components and connections used in all four hypotheses. (f) Structure and support recovery scores when the ANN in (e) is trained on data generated by simulated circuits with architecture matching (a)-(d).

hypotheses are specific to alpha cell physiology, and therefore include some variations on the more generic architectures studied in Section 5. The X- and Y-Cell models were inspired by retinal circuits commonly described in the literature [12]. They are characterized by their use of linear or nonlinear units in layer 1 respectively. These units correspond to bipolar cells, and it is known that some bipolar cells report light intensity linearly, while others do so nonlinearly [6]. The XY-Cell model is a hybrid of these two, and the first layer of this circuit includes both linear and nonlinear units. The Push-Pull model uses both excitation from nonlinear units as well as inhibition from other nonlinear units through an intermediate nonlinear unit (structurally similar to Circuits 3A and 3B in Fig. 2b).

6.1 Additional practical considerations

Leveraging insights from Section 5, we designed the learning problem to maximize the probability of success. The architecture of the artificial neural network (ANN) was designed to include the necessary synapse types for each of the four hypothesized circuits. The ANN was initialized as a superposition of all components and connections included in the X-Cell, Y-Cell, XY-Cell, and Push-Pull models. It therefore had the opportunity to converge to one of the four hypotheses, any hybrid thereof, or a completely different circuit. We initialized the ANN with enough artificial "bipolar" and "amacrine" cell units to tile the entire stimulus (though we know that the true circuit will only tile a small portion of it) while adhering to known properties of retinal cells including receptive field size and approximate synaptic convergence ratios [11]. The reason for this design choice is to ensure that the model has the capacity to contain the true circuit. This network had 417 free parameters (comparable to the darkest curves in Figures 3 and 4).

The ANN also included a temporal processing component to account for the temporal correlations inherent to ganglion cell responses. The stimuli used were temporal sequences of white

noise images and the response collected from the ganglion cells was in the form of firing rate over time (Fig. 6b). The training sets were comprised of 4500-7000 2-second videos and corresponding ganglion cell responses. Our results from Section 5 suggest that, under these conditions, most of the circuit should be recovered.

6.2 Sanity check simulation: each circuit structure can be correctly identified

As a preliminary test, each of the four structural models (Y-cell, X-cell, XY-cell and Push-pull) was simulated. A training set of ganglion cell firing rates in response to visual stimuli was generated. An ANN initialized to include all connections necessary to replicate any of the four models, as well as various other "in-between" structures (Fig. 5e), was trained in turn on each training set. It correctly selected the appropriate circuit hypothesis each time it was trained, regardless of random initialization, and achieved high values of R_{sys} and R_{supp} in every case (Fig. 5f). By comparison, R_{sys} scores for hypotheses different from the model that generated the data were considerably lower (Fig. 8). We therefore conclude that one can reliably identify which structural model gave rise to a dataset.

6.3 System identification with real data confirms the known circuitry of the tOFFα cell and suggests a new pathway

Prior research in neuroscience suggests that the tOFFα cell uses a Push-Pull motif in its circuitry [40]. Training the ANN described above (Fig. 5e) on data from mouse tOFFα cells corroborates this. The network was trained on data from 14 tOFFα cells with 5-10 random initializations for each cell. In every case, the resulting structure was not of any of the four types hypothesized, but instead a hybrid we call the XY-Push-Pull type. Thus the neural network replicated the results of the biological experiments performed in
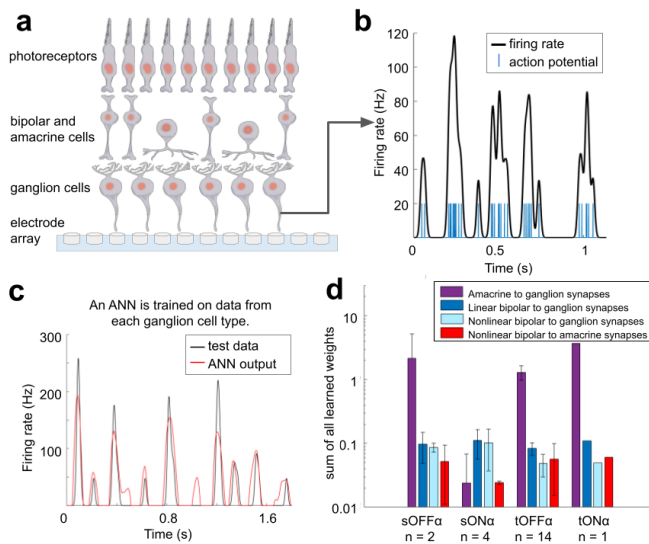
Figure 6: (a) Schematic of mouse retina on microelectrode array. The retina is positioned so that the ganglion cell layer is flush with the array of electrodes. Hundreds of ganglion cells can be simultaneously recorded. (b) Example of signal from a single mouse retinal ganglion cell. Blue ticks are single action potentials. Black curve shows the smoothed firing rate of the neuron over time. (c) Example of quality of fit to ganglion cell firing rate. (d) Summed weights for artificial neural networks trained on datasets recorded from each of the four ganglion cell types in mouse retina.

[40], but also suggested an additional pathway, namely, the inclusion of linear OFF bipolar cells which excite the tOFFα ganglion cell. Training the same network on data from other alpha cell types suggested the same circuit motif, but with varying relative weights. Further biological experiments could be conducted to confirm this finding. Thus, fine-grained identification of neural circuits can serve as a guide for these experimental efforts, allowing a biologist to prioritize hypotheses and direct future investments.

7 CONCLUSION

We have demonstrated the potential to perform accurate fine-grained system identification of nonlinear neural circuits. It can both validate existing understanding of a biological system, as well as put forth hypotheses for biological circuitry. Future theoretical directions include studying finite-sample guarantees and understanding the dependence of such guarantees on skip connections. Future modeling directions include combining with biophysical models [46] into a unified framework. More generally, this method of system identification may serve as a useful tool for scientists aiming to gain microcircuit-level understanding of systems about which only coarse-grained information is known.

Biology has seen an explosion of technological development in the past decade. Alongside this advancement in electrophysiological technologies and anatomical tools, we will need to put forth new, more powerful computational techniques that can convert large datasets into biological insights. While the retina is still among the best-understood regions of the brain, we anticipate that this level of understanding may soon be gained about other biological systems, as well as in other scientific domains. As this begins to unfold, system identification using neural networks will serve as a powerful tool for hypothesis generation and elimination. Rather than develop highly specialized algorithms for each specific application, perhaps accessible deep learning tools can be combined with regularizing domain knowledge, and used more commonly by scientists to uncover the secrets of nonlinear circuits.

REFERENCES

[1] Martín Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

[2] Reza Abbasi-Asl, Yuansi Chen, Adam Bloniarz, Michael Oliver, Ben DB Willmore, Jack L Gallant, and Bin Yu. 2018. The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *bioRxiv* (2018), 465534.

[3] Francesca Albertini, Eduardo D. Sontag, and Vincent Maillot. 1993. UNIQUENESS OF WEIGHTS FOR NEURAL NETWORKS.

[4] Uri Alon. 2006. *An introduction to systems biology: design principles of biological circuits.* CRC Press/Taylor & Francis Group.

[5] Andrew, Xu, Jian, Deniz, Cherian, Steven H., Zhu, Yongling, and Suraj. 2018. Intersectional Strategies for Targeting Amacrine and Ganglion Cell Types in the Mouse Retina. https://www.frontiersin.org/articles/10.3389/fncir.2018.00066/full

[6] Stephen A. Baccus and Markus Meister. 2002. Fast and Slow Contrast Adaptation in Retinal Circuitry. *Neuron* 36, 5 (2002), 909–919. https://doi.org/10.1016/s0896-6273(02)01050-4

[7] J. Alexander Bae, Shang Mu, Jinseop S. Kim, Nicholas L. Turner, Ignacio Tartavull, Nico Kemnitz, Chris S. Jordan, Alex D. Norton, William M. Silversmith, Rachel Prentki, and et al. 2017. Digital museum of retinal ganglion cells with dense anatomy and physiology. *Cell* 173, 5 (2017), 1293–1306. https://doi.org/10.1101/182758

[8] Ror Bellman and Karl Johan Åström. 1970. On structural identifiability. *Mathematical biosciences* 7, 3-4 (1970), 329–339.

[9] Tsai-Wen Chen, Trevor J. Wardill, Yi Sun, Stefan R. Pulver, Sabine L. Renninger, Amy Baohan, Eric R. Schreiter, Rex A. Kerr, Michael B. Orger, Vivek Jayaraman, and et al. 2013. Ultrasensitive fluorescent proteins for imaging neuronal activity. https://www.nature.com/articles/nature12354

[10] David Donoho and Victoria Stodden. 2004. When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts? In *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf (Eds.). MIT Press, 1141–1148.

[11] Felice A. Dunn and Rachel O. L. Wong. 2014. Wiring patterns in the mouse retina: collecting evidence across the connectome, physiology and light microscopy. *The Journal of Physiology* 592, 22 (2014), 4809–4823. https://doi.org/10.1113/jphysiol.2014.277228

[12] Christina Enroth-Cugell and J. G. Robson. 1966. The contrast sensitivity of retinal ganglion cells of the cat. *The Journal of Physiology* 187, 3 (Jan 1966), 517–552. https://doi.org/10.1113/jphysiol.1966.sp008107

[13] Charles Fefferman. 1994. Reconstructing a neural net from its output. *Revista Matematica Iberoamericana* 10 (1994), 507–555.

[14] Evan H. Feinberg and Markus Meister. 2014. Orientation columns in the mouse superior colliculus. https://www.nature.com/articles/nature14103

[15] Katrin Franke, Philipp Berens, Timm Schubert, Matthias Bethge, Thomas Euler, and Tom Baden. 2017. Inhibition decorrelates visual feature representations in the inner retina. Nature 542, 7642 (2017), 439–444. https://doi.org/10.1038/nature21394

[16] Renee Fry and Adrian Pagan. 2011. Sign restrictions in structural vector autoregressions: A critical review. Journal of Economic Literature 49, 4 (2011), 938–60.

[17] K Fukushima. 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. https://www.ncbi.nlm.nih.gov/pubmed/7370364/

[18] Krishna K. Ghosh, Sascha Bujan, Silke Haverkamp, Andreas Feigenspan, and Heinz Wässle. 2003. Types of bipolar cells in the mouse retina. Journal of Comparative Neurology 469, 1 (2003), 70–82. https://doi.org/10.1002/cne.10985

[19] Tim Gollisch and Markus Meister. 2010. Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. Neuron 65, 2 (2010), 150–164. https://doi.org/10.1016/j.neuron.2009.12.009

[20] Matthew J. Greene, Jinseop S. Kim, and H. Sebastian Seung. 2016. Analogous Convergence of Sustained and Transient Inputs in Parallel On and Off Pathways for Retinal Motion Computation. Cell Reports 14, 8 (2016), 1892–1900. https://doi.org/10.1016/j.celrep.2016.02.001

[21] Rong Gui, Quan Liu, Yuangen Yao, Haiyou Deng, Chengzhang Ma, Ya Jia, and Ming Yi. 2016. Noise Decomposition Principle in a Coherent Feed-Forward Transcriptional Regulatory Loop. Frontiers in Physiology 7 (2016). https://doi.org/10.3389/fphys.2016.00600

[22] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. 2015. Statistical learning with sparsity: the lasso and generalizations. CRC press.

[23] K. Huang, N. D. Sidiropoulos, and A. Swami. 2014. Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition. IEEE Transactions on Signal Processing 62, 1 (2014), 211–224.

[24] James J. Jun, Nicholas A. Steinmetz, Joshua H. Siegle, Daniel J. Denman, Marius Bauza, Brian Barbarits, Albert K. Lee, Costas A. Anastassiou, Alexandru Andrei, Çağatay Aydın, and et al. 2017. Fully integrated silicon probes for high-density recording of neural activity. https://www.nature.com/articles/nature24636

[25] Justin Keat, Pamela Reinagel, R.clay Reid, and Markus Meister. 2001. Predicting Every Spike: A Model for the Responses of Visual Neurons. Neuron 30, 3 (2001), 803–817. https://doi.org/10.1016/s0896-6273(01)00322-1

[26] Jinseop S. Kim, Matthew J. Greene, Aleksandar Zlateski, Kisuk Lee, Mark Richardson, Srinivas C. Turaga, Michael Purcaro, Matthew Balkam, Amy Robinson, Bardia F. Behabadi, and et al. 2014. Space–time wiring specificity supports direction selectivity in the retina. Nature 509, 7500 (2014), 331–336. https://doi.org/10.1038/nature13240

[27] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[28] David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. 2017. Neural system identification for large populations separating "what" and "where". In Advances in Neural Information Processing Systems. 3506–3516.

[29] John G Kuschewski, Stefen Hui, and Stanislaw H Zak. 1993. Application of feedforward neural networks to dynamical system identification and control. IEEE Transactions on Control Systems Technology 1, 1 (1993), 37–49.

[30] Hans Laurberg, Mads Græsbøll Christensen, Mark D. Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. 2008. Theorems on Positive Data: On the Uniqueness of NMF. Computational Intelligence and Neuroscience 2008 (2008), 788 – 791. https://doi.org/10.1155/2008/764206

[31] Yen-Huan Li, Jonathan Scarlett, Pradeep Ravikumar, and Volkan Cevher. 2015. Sparsistency of ℓ_1-Regularized M-Estimators.. In AISTATS.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014 Lecture Notes in Computer Science (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[33] Lennart Ljung. 1999. System identification. Wiley encyclopedia of electrical and electronics engineering (1999), 1–19.

[34] Calvin McCarter and Seyoung Kim. 2014. On sparse Gaussian chain graph models. Advances in Neural Information Processing Systems 27 (2014), 3212–3220.

[35] Lane McIntosh and Niru Maheswaranathan. 2015. A Deep Learning Model of the Retina.

[36] Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. 2016. Deep Learning Models of the Retinal Response to Natural Scenes. In Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 1369–1377. http://papers.nips.cc/paper/6388-deep-learning-models-of-the-retinal-response-to-natural-scenes.pdf

[37] Markus Meister and Michael J. Berry. 1999. The Neural Code of the Retina. Neuron 22, 3 (1999), 435–450. https://doi.org/10.1016/s0896-6273(00)80700-x

[38] Markus Meister, Jerome Pine, and Denis A. Baylor. 1994. Multi-neuronal signals from the retina: acquisition and analysis. Journal of Neuroscience Methods 51, 1 (1994), 95–106. https://doi.org/10.1016/0165-0270(94)90030-2

[39] A. Mizrahi. 2013. Synaptogenesis in the Adult CNS–Olfactory System. Cellular Migration and Formation of Neuronal Connections (2013), 739–755. https://doi.org/10.1016/b978-0-12-397266-8.00112-5

[40] Thomas A Münch, Rava Azeredo Da Silveira, Sandra Siegert, Tim James Viney, Gautam B Awatramani, and Botond Roska. 2009. Approach sensitivity in the retina processed by a multifunctional neural circuit. Nature Neuroscience 12, 10 (Jun 2009), 1308–1316. https://doi.org/10.1038/nn.2389

[41] Marius Pachitariu, Nicholas A Steinmetz, Shabnam N Kadir, Matteo Carandini, and Kenneth D Harris. 2016. Fast and accurate spike sorting of high-channel count probes with KiloSort. In Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 4448–4456.

[42] Elio Raviola and Giuseppina Raviola. 1982. Structure of the synaptic membranes in the inner plexiform layer of the retina: A freeze-fracture study in monkeys and rabbits. The Journal of Comparative Neurology 209, 3 (1982), 233–248. https://doi.org/10.1002/cne.902090303

[43] Esteban Real, Hiroki Asari, Tim Gollisch, and Markus Meister. 2017. Neural Circuit Inference from Function to Structure. Current Biology 27, 2 (2017), 189–198. https://doi.org/10.1016/j.cub.2016.11.040

[44] David Rolnick and Konrad P. Kording. 2020. Identifying Weights and Architectures of Unknown Re{LU} Networks.

[45] Basil Saeed, Snigdha Panigrahi, and Caroline Uhler. 2020. Causal Structure Discovery from Distributions Arising from Mixtures of DAGs. In International Conference on Machine Learning. PMLR, 8336–8345.

[46] Cornelius Schröder, David Klindt, Sarah Strauss, Katrin Franke, Matthias Bethge, Thomas Euler, and Philipp Berens. 2020. System Identification with Biophysical Constraints: A Circuit Model of the Inner Retina. (2020). https://doi.org/10.1101/2020.06.16.154203

[47] Katja Seeliger, Luca Ambrogioni, Yağmur Güçlütürk, Umut Güçlü, and Marcel AJ van Gerven. 2019. End-to-end neural system identification with neural information flow. bioRxiv (2019), 553255.

[48] Jie Shen and Ping Li. 2017. On the iteration complexity of support recovery via hard thresholding pursuit. In Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 3115–3124.

[49] Sandra Siegert, Brigitte Gross Scherf, Karina Del Punta, Nick Didkovsky, Nathaniel Heintz, and Botond Roska. 2009. Genetic address book for retinal cell types. Nature Neuroscience 12, 9 (Feb 2009), 1197–1204. https://doi.org/10.1038/nn.2370

[50] Kyung-Ah Sohn and Seyoung Kim. 2012. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In Artificial Intelligence and Statistics. PMLR, 1081–1089.

[51] Raghav Somani, Chirag Gupta, Prateek Jain, and Praneeth Netrapalli. 2018. Support recovery for orthogonal matching pursuit: upper and lower bounds. In Advances in Neural Information Processing Systems. 10814–10824.

[52] Héctor J. Sussmann. 1992. Uniqueness of the weights for minimal feedforward nets with a given input-output map. Neural Networks (1992).

[53] Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, and Surya Ganguli. 2019. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. In Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 8537–8547.

[54] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 1 (1996), 267–288.

[55] Robert E Tillman and Frederick Eberhardt. 2014. Learning causal structure from multiple datasets with similar variable sets. Behaviormetrika 41, 1 (2014), 41–64.

[56] Yoshihiko Tsukamoto and Naoko Omi. 2017. Classification of Mouse Retinal Bipolar Cells: Type-Specific Connectivity with Special Reference to Rod-Driven AII Amacrine Pathways. Frontiers in Neuroanatomy 11 (2017). https://doi.org/10.3389/fnana.2017.00092

[57] Verner Vlacic and Helmut Bölcskei. 2019. Neural network identifiability for a family of sigmoidal nonlinearities. ArXiv abs/1906.06994 (2019).

[58] H. Wassle and B. B. Boycott. 1991. Functional architecture of the mammalian retina. Physiological Reviews 71, 2 (1991), 447–480. https://doi.org/10.1152/physrev.1991.71.2.447

[59] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111, 23 (2014), 8619–8624.

[60] Yifeng Zhang, In-Jung Kim, Joshua R. Sanes, and Markus Meister. 2012. The most numerous ganglion cell type of the mouse retina is a selective feature detector. https://www.pnas.org/content/109/36/E2391

[61] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. Advances in Neural Information Processing Systems 31 (2018), 9472–9483.

A APPENDIX

A.1 Reproducibility: Empirical study

A.1.1 Computing methods and data repository. All code for this project was run on Amazon Web Services p2.xlarge instances, equipped with the Deep Learning AMI. All trainings and simulations were run using Tensorflow v1.15.0 [1] and Python 2.7. All trainings were done using the Adam optimizer [27]. Precise hyperparameters used to generate the figures in this paper can be found in the KDD_readme file. Training code and evaluation code can be found at https://github.com/dbagherian/neural-sys-id. Data files and examples of trained models can be found at https://doi.org/10.22002/D1.1989.

All hyperparameters were selected via cross validation. During cross-validation, the learning rate was varied between 10^{-4} and 10^{-2}, while the ℓ_1-regularizer strength parameter, λ, was varied between 10^{-7} and 10^2. The README file in the link above lists all hyperparameter values needed to recreate every figure in this paper.

A.1.2 Complexity and runtimes. The complexity of system identification can be understood in units of "network training runs." The number of free parameters in each network varied between 300 and 6000. Each training in figures 3, 4, 5 and 6 was run 5-10 times with different random initializations. Average runtime for these simulated trainings varied between 5 minutes and 4 hours, depending on the quantity of training data used.

A.1.3 Datasets. In the simulations in section 5, between 60 and 22000 training examples were used, and between 2000 and 2080 examples were used as a test set. Stimuli used for training were of two forms: random checkerboard and natural images. Random checkerboard stimuli were 100×100 pixel images with 5×5 pixel grayscale checkers of continuously varying intensity in the empirical simulations. Natural images were derived from the COCO (Common Objects in Context) dataset [32]. Images from this set were made grayscale and randomly cropped to 100×100 pixels.

A.1.4 Additional details of empirical simulations. The simulated biological circuits studied in section 5 were inspired by the retina in multiple ways. In the first layer of these circuits, spatial convolutional filters are applied to input images. These filters were inspired by retinal spatial receptive fields with antagonistic surrounds [37], and composed of two opposing 2D Gaussian functions. Three different convolutional filters were used in the oracle networks with Gaussians of varying diameters. 28 different convolutional filters were included in the ANN. 14 of these were OFF filters, activated by dark spots in the image, and 14 were ON filters, activated by light spots in the image.

The remaining two layers took linear combinations of the units in the prior layer. All layers included ReLU nonlinearities. The output of the network was a continuously-varying "firing rate" value for the single output neuron, modeling the rate at which the mouse retinal ganglion cell fires action potentials.

In order to remove some degeneracy from the space of circuit models being searched, it was necessary to fix the biases in place during training. The toy example in Fig. 7 illustrates the need for this. In order to efficiently compute the structure recovery score,
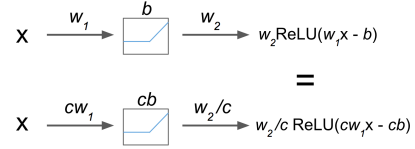

Figure 7: Two toy networks with differing parameter sets which produce the same output for an input, x.

	X-Cell	Y-Cell	XY-Cell	Push-Pull
X-Cell	1.00	0.00	0.80	0.00
Y-Cell	0.00	1.00	0.40	0.07
XY-Cell	0.80	0.40	1.00	0.04
Push-Pull	0.00	0.07	0.04	1.00

Figure 8: For each simulated circuit used in the simulations in figure 5, the R_{sys} was computed with every other circuit hypothesis. This serves as a "floor" to better contextualize the R_{sys} of the trained ANN's in each case.

R, we needed to remove this type of degeneracy from the space of circuit structures being searched. When b is fixed in place during training, this degeneracy is removed.

A.2 Reproducibility: Mouse retina case study

A.2.1 Datasets. In section 6.2, training data consisted of simulated responses to 60 Hz time-varying randomly flickering bar stimuli with varying degrees of spatial correlation. 85 minutes of simulated data were used to train, and 13 minutes to compute out-of-sample loss. (Bipolar cells in the retina integrate stimulus information in a time window of about 0.2 seconds [6].) In section 6.3, about 3 hours of electrophysiological data were used in each training set. 4.5 min of data were used to compute out-of-sample loss. To collect these data, wild type C57/Bl6 mice were euthanized with cervical dislocation. The retina was dissected out from the eye in an oxygenated nutrient bath, then mounted on a 256-channel electrode array. Visual stimuli were projected onto the retina, while action potentials were recorded extracellularly from the ganglion cell layer [38]. We were able to keep the retina alive for 2-5 hours and to record training data in that time. Spike sorting was performed with Kilosort [41]. α cells were identified post hoc by their responses to a battery of visual stimuli. Training data consisted of responses to time-varying 1D random noise stimuli with varying degrees of spatial correlation. Spike trains were convolved with a Gaussian filter to produce a continuously varying firing rate (Fig. 6b). The trained ANN explained a large fraction of the variance of each dataset. By cell type, these fractions are: sOFFα 0.75 ± 0.04, sONα 0.82 ± 0.17, tOFFα 0.89 ± 0.11, tONα 0.65 ± 0.00.

A.2.2 Parameter reduction in the first convolutional layer. In order to reduce the number of learned parameters and speed up training on the mouse data, we pulled the convolution out of the first layer of the trained network in the following way.

Since the functional form of bipolar cell spatiotemporal filters is well understood in biology [6, 15, 25, 43], one can parameterize the temporal convolutional filters in the first layer of the ANN using a

basis set of stretched sinusoid functions of the form:

$$T_j = \begin{cases} \sin\left(\pi j\left(2\frac{t}{\tau} - \left(\frac{t}{\tau}\right)^2\right)\right), & \text{for } 0 \le t \le \tau \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

for $j = 1, 2, ..., 16$ [25]. Let s_i denote the ith stimulus video in the training set. The result of the first layer convolution could therefore be rewritten as:

$$y_i^1 = \sum_{j=1}^{16} \beta_j (s_i * T_j) \quad (7)$$

where $*$ denotes the convolution operation. The β_j could then be initialized in a biologically realistic regime, derived from bipolar cell electrophysiological recordings [15], but allowed to vary slightly as the network fit the data. The input to the ANN, therefore, can be thought of as $\{s_i * T_j\}_{j=1}^{16}$, rather than s_i. This removes the first layer convolution from the backpropagation during training and reduces the number of parameters to learn.

A.3 Retinal address book

We constructed a "retinal address book" based on an idea first presented in [49]. See section 5 for details.

Because these anatomical data come from a number of studies with varying forms and thoroughness of anatomical data published, this address book was constructed by painstaking manual inspection. For example, different authors divided the IPL into different numbers of sublaminae, and set their boundaries in different locations.

We selected a sublamination scheme that was compatible with each of these. In this scheme, the mouse IPL was divided into six sublaminae: outer marginal (0.0-0.28 normalized depth), outer central (0.28-0.47 normalized depth), inner central (0.47-0.65 normalized depth), inner marginal (0.65-1.0 normalized depth) and the ON and OFF ChAT bands (defined by limits of choline acetyltransferase (ChAT) expression). With this sublamination scheme, we were able to neatly define a binary stratification profile for each cell type under study, and to therefore create a binary address book You can view this address book at https://www.dropbox.com/s/tryk5vo856ny42b/IPL%20Address%20Book.png?dl=0

One could imagine, however, a future in which stratification profiles are given in units of normalized IPL depth, with confidence intervals or some other measure of uncertainty, for every retinal cell type. From such a dataset, one could replace the binary entries of this table with a continuously varying measure of co-stratification and uncertainty.

A.4 Proof of Theorem 1

In this section, we prove the identifiability result of Section 4. See [3, 13, 44, 52, 57] for related prior work. Recall the definition of network (2):

$$f(x) := W_2 \max(0, W_1 x + b_1) + W_3 x + b_2,$$

for weights $W_1 \in \mathbb{R}^{n_1 \times n_0}$, $W_2 \in \mathbb{R}^{n_2 \times n_1}$, $W_3 \in \mathbb{R}^{n_2 \times n_0}$ and biases $b_1 \in \mathbb{R}^{n_1}$, $b_2 \in \mathbb{R}^{n_2}$.

THEOREM 1. *Suppose that network (2) satisfies the following three conditions:*

(i) no column of W_2 or row of W_1 is entirely zero;
(ii) no two rows of W_1 are collinear;
(iii) W_1 is nonnegative.
Let P denote an unknown permutation matrix and D an unknown positive diagonal matrix. Then, by input-output queries of the form (1), we may recover DPW_1, DPb_1, $W_2 P^{-1} D^{-1}$, W_3 and b_2.

PROOF. To begin, observe that each component of the vector output $f(x)$ is a piecewise linear function of x. We will first show that the transitions of the max function may be identified with the transitions between linear regions, thereby allowing identification of W_1 and b_1.

Since condition (i) excludes the possibility that any row of W_1 is entirely zero, every row of W_1 will cause a transition of the max function. For the jth row of W_1, the transition occurs on a hyperplane H_j in the input domain:

$$H_j := \left\{ x : \sum_{k=1}^{n_0} W_1^{(jk)} x^{(k)} + b^{(j)} = 0 \right\} \subset \mathbb{R}^{n_0}.$$

The transition of the max function corresponding to the jth row of W_1 will only affect output component $f^{(i)}$ if $W_2^{(ij)} \ne 0$. We know by condition (i) that such an index i exists since the jth column of W_2 is not entirely zero. Therefore all rows of W_1 correspond to a nonlinear transition in at least one output component of f. Since the rows of W_1 are not collinear by condition (ii), we are now sure that the boundaries between linear regions of f correspond to n_1 distinct hyperplanes that partition input space. Since the formula of a hyperplane is unique up to scalings, we may recover b_1 and the rows of W_1 up to scalings. Since we do not know in which order the hyperplanes should be listed, the recovery is only unique up to scalings *and* permutations.

Since we have identified W_1 and b_1 up to scaled permutations of the rows of W_1, we may now query f in the region of input space where all components of max return zero. This region surely exists since by conditions (i) and (iii) combined, all rows of W_1 point into the positive orthant. Therefore, for x sufficiently far into the negative orthant, we have that:

$$f(x) = W_3 x + b_2.$$

The gradient $\nabla_x f(x)$ in this region identifies W_3, at which point b_2 may be identified via:

$$b_2 = f(x) - W_3 x.$$

All that remains is to identify W_2. Consider a point x^* on the jth hyperplane H_j that is far away from the other hyperplanes $\{H_i\}_{i \ne j}$. Such a point exists by condition (ii). Let x^+ and x^- denote points in the local neighbourhood of x^* but on the positive and negative side of H_j, respectively. Observe that:

$$\frac{\partial f^{(i)}}{\partial x^{(k)}}(x^+) - \frac{\partial f^{(i)}}{\partial x^{(k)}}(x^-) = W_2^{(ij)} W_1^{(jk)}.$$

Therefore $W_2^{(ij)}$ may be identified by measuring the change in gradient of $f^{(i)}$ across the jth hyperplane H_j. Of course, since the rows of W_1 are known only up to a permutation and scale, we will be unsure of the columns of W_2 up to the same symmetries.

With W_3 and b_2 identified exactly, and W_2, W_1 and b_1 identified up to a permutation and scale, we are done. □