Learning the Architectural Features That Predict Functional Similarity of Neural Networks

Adam Haber and Elad Schneidman*

Department of Brain Sciences, Weizmann Institute of Science, Rehovot 76100, Israel

(Received 25 September 2020; revised 14 July 2021; accepted 22 September 2021; published 3 June 2022)

The mapping of the wiring diagrams of neural circuits promises to allow us to link the structure and function of neural networks. Current approaches to analyzing such *connectomes* rely mainly on graph-theoretical tools, but these may downplay the complex nonlinear dynamics of single neurons and the way networks respond to their inputs. Here, we measure the functional similarity of simulated networks of neurons, by quantifying the similitude of their spiking patterns in response to the same stimuli. We find that common graph-theory metrics convey little information about the similarity of networks' responses. Instead, we learn a functional metric between networks based on their synaptic differences and show that it accurately predicts the similarity of novel networks, for a wide range of stimuli. We then show that a sparse set of architectural features—the sum of synaptic inputs that each neuron receives and the sum of each neuron's synaptic outputs—predicts the functional similarity of networks of up to 1000 neurons, with high accuracy. We thus suggest new architectural design principles that shape the function of neural networks. These architectural features conform with experimental evidence of homeostatic synaptic mechanisms.

DOI: 10.1103/PhysRevX.12.021051

Subject Areas: Biological Physics, Complex Systems
Interdisciplinary Physics

I. INTRODUCTION

Many biological systems can be described as networks of interacting elements where the function of the system is determined by the nature of individual elements, the type of interactions between them, and the emerging individual and collective behavior or phenotype. Mapping the relation between the structure and function of such networks is a key goal in many areas of biology, as well as engineering, social networks, and more. Because the number of possible architectures is combinatorial in the size of the network, analyzing and understanding the design and function of biological networks hinge on finding simplifying principles. Functional "design principles" have been suggested to include robustness to noise [1–3], resilience to attack [4], controllability [5], efficiency [6,7], criticality [8,9], and learnability [10,11]. Structural design principles have implied the nature of network growth [12], use of small subnetwork motifs [13], modular organization and power-law scaling [14,15], sparseness of activity [16], random connectivity [17,18], and centrality or percolation properties of networks [19,20]. However, how these functional and structural design principles relate to one another is not immediately clear [21].

The reconstruction of the detailed wiring diagrams of full neural circuits at single-cell resolution [22–26] would enable direct exploration and characterization of the architectural design of neural modules and even whole brains [27]. Importantly, very different connectivity structures may give rise to very similar function [28]. Thus, the ability to record the joint activity patterns of large populations of neurons [29,30] whose *connectome* has been reconstructed is crucial for linking of neural networks' structure and function [31–34]—which would be central to our understanding of development, coding, plasticity, and learning in biological neural networks.

Understanding the relations between network topology and the activity of networks of neurons requires ways to measure both the functional similarity of networks and their architectural similarity and to map the relations between these two, potentially very different, metrics. It is not obvious what is the correct measure for either one or how we may extend tools from graph and network theories [12,15,35]. For some classes of networks and of interacting elements, links between the topology of a network and the nature of its dynamics have been elucidated, such as the number of fixed points or classes of attractor dynamics and their stability [36–40]. Toward the study of real neural networks, we present here a general framework for linking the topology of networks and their population spiking patterns for arbitrary classes of network architectures, in response to a wide range of stimuli. This framework can be extended to many other biological and

*elad.schneidman@weizmann.ac.il

nonbiological systems that can be described by networks of interacting elements.

Rather than assuming or guessing which structural metric should be used to compare networks, we learn a functional similarity metric based on the structural differences of the networks. We thus characterize the space of neural networks in terms of their function and then seek the architectural principles that govern the organization of that space. We develop this framework and validate it by studying simulated networks of spiking neurons, where we have complete control over all parameters, no limits on experimental design and length, and the "ground truth" is known. We show that we can *learn* the informative structural features that shape a network's function and that a structural metric, based on these features, significantly outperforms a wide range of graph-theoretical measures in predicting the functional similarity of neural circuits. We then show that the informative structural features that we identify for small networks are highly informative also for networks of up to 1000 neurons—suggesting them as a general principle for the comparison of networks of neurons.

II. RESULTS

To study the relation between structure and function in networks of neurons, we simulate the responses of tens of thousands of small networks of spiking neurons to a wide range of stimuli [Fig. 1(a)]. Direct characterization of the space of all network architectures is impossible for large networks, since for a group of N neurons there are $2^{N(N-1)}$ different directed graphs of interactions (topologies); considering neurons of different types or diverse strengths of

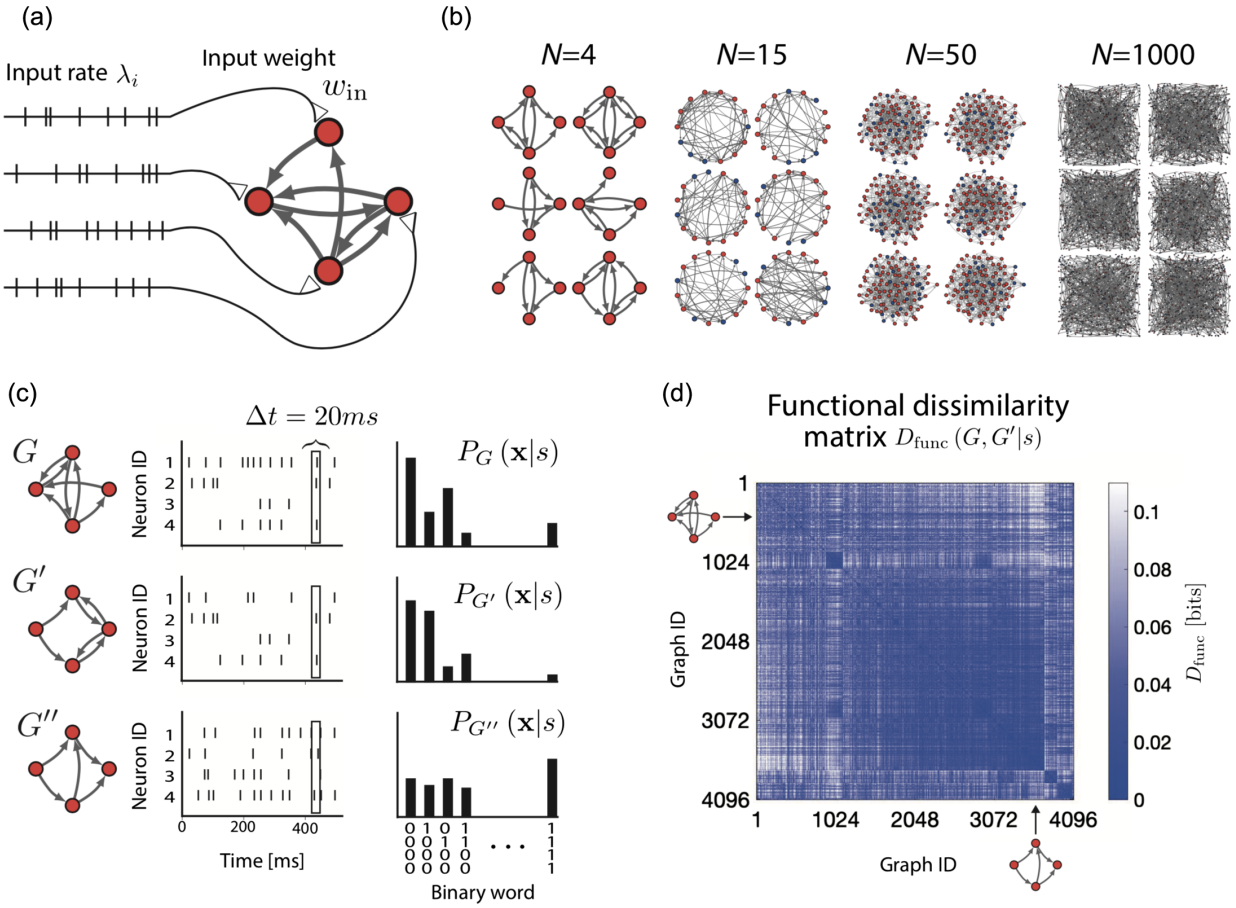

FIG. 1. Simulating the responses of ensembles of networks to the same stimuli and computing functional similarity between networks. (a) Each of the N neurons in the simulated networks receive as an input a Poisson spike train with rate λ_i; corresponding neurons in all networks in the ensemble receive the same exact ith spike train as input. (b) Examples of the simulated networks: networks of four neurons with identical excitatory synapses and networks of 15, 50, or 1000 excitatory and inhibitory neurons where the synaptic weights are drawn from a log-normal distribution. (c) Examples of segments of the spike train responses of different networks (left) that are presented with the same stimulus. The activity of the neurons is discretized into bins of length $\Delta t = 20$ ms and binarized (middle; see the text). The different responses of the networks result in different distributions of population activity patterns (right). (d) The functional dissimilarity matrix $D_{\rm func}(G, G'|s)$ for networks of size $N = 4$. The normalized stimulus strength is $\eta = 1.5$ (see Sec. IV), which results in firing rates ranging between 10 and 20 Hz over all networks in the ensemble. The structure of this matrix implies a low-dimension organization of the functional space of networks (see the text).

synaptic connections makes this number considerably larger. We, therefore, start by seeking the principles that govern small networks with different topological and sign properties and increase their size later [Fig. 1(b)]. We first consider the two exhaustive sets of all 4096 topologies of networks of four neurons comprised of all excitatory or all inhibitory leaky integrate and fire neurons, with all synapses having the same strength (we simulate both current- and conductance-based models for the neurons and both alpha and delta activation functions for the synapses—see the Appendix). The next ensemble is comprised of networks of 15 excitatory and inhibitory neurons, with 20% inhibitory neutrons [41], where synaptic strengths are drawn from a log-normal distribution [42] and the average excitation and inhibition are balanced. We then consider another ensemble of balanced networks of 15 neurons in which the connections are drawn from a 3D geometric random graph model, such that the probability of synaptic connections is distance dependent (see Sec. IV), resulting in different clustering profiles and motif distributions within networks (Supplemental Fig. S1 [43]). Even for 15 neurons, there are approximately 10^{63} directed topologies, and so we use a random sample of 10 000 networks of each size and rely on cross-validation of new networks to verify our results and models (see Sec. IV).

To map the functional similarity between networks, we simulate their responses to the same stimulus s and compare their respective population activity patterns [Fig. 1(c)]. We denote each network by a weighted connectivity graph or the matrix of synaptic connection G, where G_{kl} is the strength of the synapse from neuron k to l. The external stimulus s to the networks is defined individually for each neuron, such that the ith neuron in each network receives as an input a 30-s Poisson-distributed spike train with a rate λ_i, weighted by an input synaptic weight w_{input} [Fig. 1(a)]. While the rate λ_i of inputs to all neurons is identical, each neuron in the network receives its own realization of incoming spikes with that rate (whereas corresponding neurons in different networks receive the *exact* same input patterns, i.e., same realization). We explore the responses of the networks to a wide range of stimuli (see the Appendix and Supplemental Fig. S2 [43]) and focus henceforth on stimulus parameters and synaptic strength values for which the networks' behavior is not pathological, i.e., epileptic or completely silent. The length of the stimulus is chosen to give an adequate sample of the responses that this class of stimuli would elicit. To disentangle the effects of architectural differences between networks from the effects of the initial conditions of networks on their responses, the same set of initial conditions is used for each network, and all network measures and similarity measures between networks are the average over many different sets of random initializations.

We discretize the spiking patterns of the neurons in the network in response to the stimulus into small temporal bins of size $\Delta t = 20$ ms, such that the activity of the network in each time bin is given by a binary vector $\mathbf{x} = (x_1, \ldots, x_N)$, where $x_k = 1$ if neuron k spikes in that bin and 0 otherwise [Fig. 1(c), middle]. We then summarize the response of the network whose synaptic connections are given by G, to stimulus s by the distribution of population activity patterns over the whole length of the stimulus [Fig. 1(c), right], which we denote by $P_G(\mathbf{x}|s)$. For small networks, we estimate these distributions by direct sampling of the networks' "vocabulary," whereas for large ones we fit a pairwise maximum entropy model [44] for the population activity (see Sec. IV). Using different temporal bin size gives similar results for the analyses that follow (see the Appendix and Supplemental Fig. S3 [43]).

We quantify the functional similarity of pairs of networks whose synaptic weights are given by G and G' using the overlap of their population responses, conditioned on the stimulus s:

$$D_{\text{func}}(G, G'|s) = D_{\text{JS}}[P_G(\mathbf{x}|s)||P_{G'}(\mathbf{x}|s)], \tag{1}$$

where D_{JS} is the Jensen-Shannon divergence, a symmetric and bounded measure of the distinguishability of distributions, ranging from 0 bits for identical distributions to 1 bit for nonoverlapping ones (see Sec. IV). Notably, this comparison of the networks' population vocabulary in response to a particular stimulus or class of stimuli is more general than overlap measures based only on individual firing rates, as it also considers differences in the correlations between neurons. As this does not take into account the temporal structure of the responses, we also compare networks based on the similarity of the post-stimulus-time histogram (PSTH) of the corresponding neurons in the two networks to the same stimulus, D_{PSTH} (see the Appendix). We find that D_{PSTH} is strongly correlated with D_{func} (see the Appendix and Supplemental Fig. S4 [43]), and we, therefore, focus on the latter for the rest of the analyses.

We compute the functional dissimilarity between all networks in our ensembles. Figure 1(d) shows a typical example of the resulting matrix of dissimilarity values between pairs of networks, $D_{\text{func}}(G, G'|s)$, for s with a normalized input rate of $\eta = 1.5$ and $w_{\text{input}} = 20$ pA (which is equivalent to a total rate of incoming spikes to each neuron of 8 kHz; see Sec. IV). The structure of the matrix reflects a low-dimensional organization of the space of networks based on their response properties—evident by the spectrum of the eigenvalues of the matrix, which decay significantly faster than shuffled controls (see the Appendix and Supplemental Fig. S5 [43]). We then ask what are the structural properties of the networks that underlie this functional organization.

A. Common structural metrics fail to capture the functional similarity of networks of neurons

Given the plethora of graph-theory measures of similarity, it might seem that a smart choice of one such

measure might be sufficient to predict the functional similarity of neural networks. But, which one should we choose? The simplest intuitive way to compare the topology of networks of the same size is by counting the fraction of links or synapses they share. For weighted directed graphs, this can be interpreted in different ways, and we consider here two options: first, the graph edit or Hamming distance between G and G':

$$d_{\text{Hamming}}(G,G') = \sum_{k,l=1}^{N} \mathbb{1}_{\text{sgn}G_{kl} \neq \text{sgn}G'_{kl}}, \tag{2}$$

where $\mathbb{1}$ is an indicator function, which compares the *type* of synaptic connections between neurons (inhibitory, excitatory, or absent) and, second, the L_2 norm or Euclidean distance between the corresponding synapses in the networks:

$$d_{L_2}(G,G') = \sqrt{\sum_{k,l=1}^{N} (G_{kl} - G'_{kl})^2}, \tag{3}$$

where G_{kl} is the strength of the synapse from neuron k to l in network G. Both of these metrics prove to be poor predictors of the functional similarity of networks

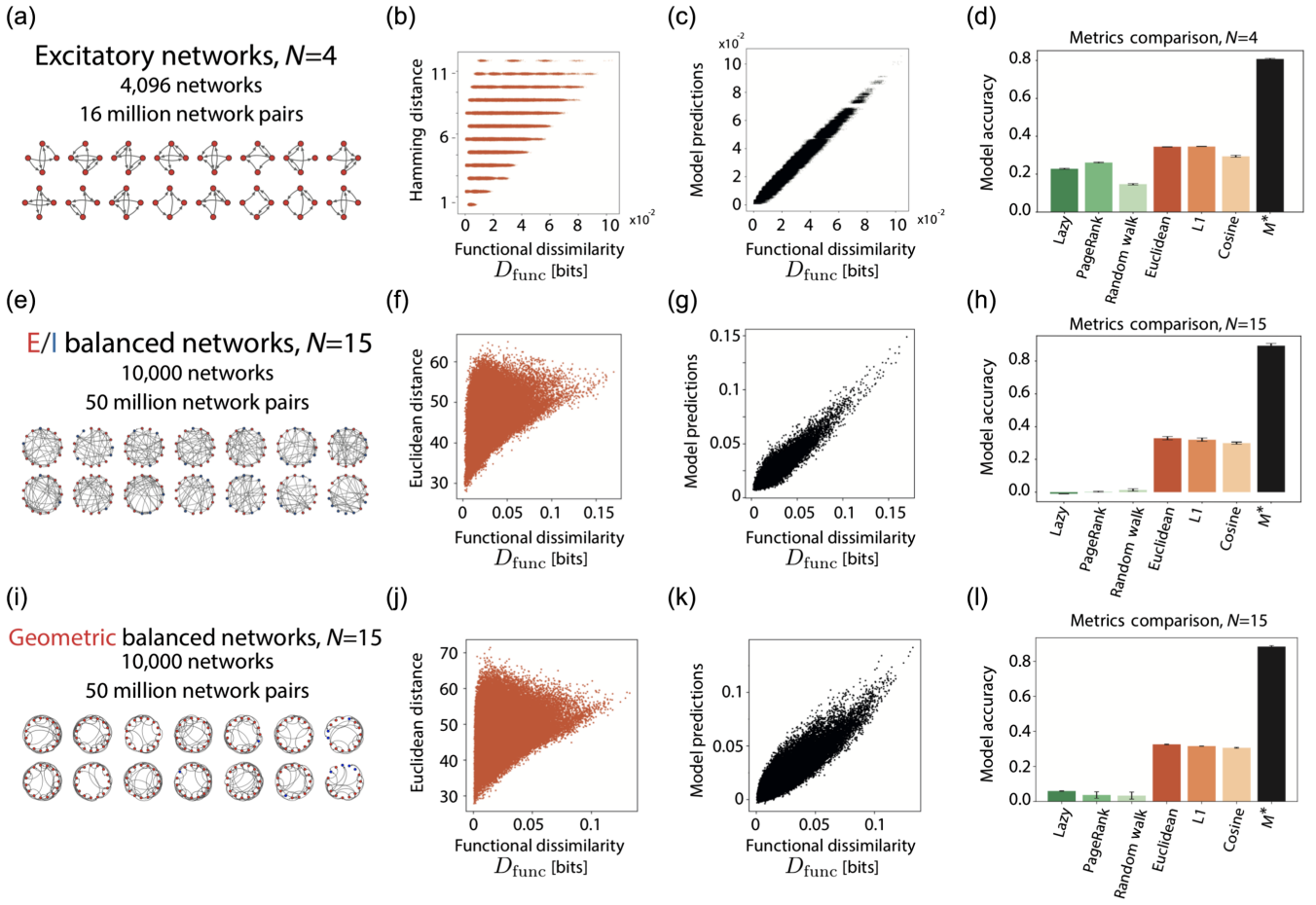

FIG. 2. Common structural metrics fail to predict the functional similarity between networks of neurons, whereas a learned bilinear metric succeeds. (a) The first simulated ensemble, consisting of 4096 networks of four neurons with constant weights. (b) Hamming distances between all pairs of 4096 networks of size $N = 4$ are plotted against the computed D_{func} between networks (y values are slightly jittered to show the density of points). We note the points in the upper-left correspond to pairs of networks that have almost no overlap in terms of synaptic connections yet are functionally very similar. (c) Predictions of the bilinear model on held-out test data, i.e., networks that are not used in finding M^*. (d) We compare the prediction accuracy of different structural metrics for networks of size $N = 4$, by computing the mean Pearson correlation of their predictions with the functional dissimilarity of the networks D_{func} (averaged over 30 different initial conditions). Our learned model (black) outperforms multiple graph-based models (green) and vector-based ones (orange). Error bars represent one standard deviation. (e) The second simulated ensemble, consisting of 10 000 networks of 15 neurons with balanced excitatory and inhibitory weights (on average). (f)–(h) The same as (b)–(d) but for an ensemble of networks with $N = 15$. Since synaptic weights in this ensemble are continuous, the Euclidean distance between synaptic weights is used instead of Hamming (see the main text). (i) This simulated ensemble consists of 10 000 "geometric" networks, each with 15 neurons with both excitatory and inhibitory weights that are balanced (on average), but with distance-dependent connectivity (see the main text). (j)–(l) The same as (f)–(h) but for the ensemble of geometric networks.

[Figs. 2(b), 2(f), and 2(j)]. In particular, while low Hamming or Euclidean distance does imply low D_{func}, medium and high values of Hamming or Euclidean distance are practically noninformative of functional similarity.

We examine a wide variety of other common structural metrics, including ones that treat networks as vectors of synaptic weights and compare them as vectors in $R^{N(N-1)}$, as well as explicit metrics of graphs, such as the spectral distances between directed graph Laplacians. All the structural metrics in our comparisons give poor results, for both $N = 4$ and $N = 15$—as reflected by the green and orange bars in Figs. 2(d), 2(h), and 2(l) (a larger set of failed metrics can be found in the Appendix and Supplemental Fig. S6 [43]). We, therefore, ask whether, instead of trying common similarity measures, we could learn a metric directly from the data.

B. Learning to predict the functional similarity of networks from their synaptic connections

Metric learning approaches vary significantly in their assumptions and applications [45] and are used to model perceptual distances [46,47] and to study the structure of neural codes [48–50]. Here, we ask how well D_{func} between networks' responses to the same stimulus s can be approximated by a bilinear function of the synaptic differences between the networks, which we quantify as the difference of their connectivity matrices $\Delta G_{kl} = G_{kl} - G'_{kl}$. To simplify mathematical notation, we denote by

$$g = \begin{pmatrix} G_{12} \\ \vdots \\ G_{N(N-1)} \end{pmatrix} \in \mathbb{R}^{N(N-1)}$$

the vector corresponding to the "flattened" representation of the matrix G, without its diagonal elements (which are all zero), and so $\Delta g = g - g'$ is the vector of synaptic differences between two networks. The problem of finding the optimal bilinear function is then translated into seeking the optimal matrix $M^*(s)$, which is given by

$$M^*(s) = \underset{M \succeq 0}{\operatorname{argmin}} \langle [D_{\text{func}}(G, G'|s) - \Delta g^T \cdot M \cdot \Delta g]^2 \rangle_{G,G'} + \alpha \|M\|_2, \tag{4}$$

where M is an $N(N-1) \times N(N-1)$ positive-semidefinite matrix (so defined in order for the distances between networks to be non-negative), known as the Mahalanobis matrix [45]. The regularization term and its control parameter α are chosen by cross-validation (see the Appendix and Supplemental Fig. S7 [43]). Fortunately, $M^*(s)$ we seek is the solution to a convex constrained optimization problem, which is, therefore, guaranteed to be the global optimum. We note the dependence of M^* on s and stress that different stimuli might require different metrics, which we investigate below.

To find $M^*(s)$, we randomly split the networks in the ensemble into a training set and a test set (75% and 25% of the networks, respectively) and use the train set to find $M^*(s)$, using conjugate gradient descent on the manifold of positive-semidefinite matrices (see Sec. IV). To assess how well this learned measure captures the functional dissimilarity between networks, we use it to predict the pairwise distances between all pairs of networks in our held-out test set and compare these predictions to the empirical D_{func} values [Figs. 2(c), 2(g), and 2(k)]. We find that M^* captures functional dissimilarity significantly better than all other metrics [Figs. 2(d), 2(h), and 2(l)], for networks of size $N = 4$ and $N = 15$. In particular, this is true for different random graph models (randomly connected ones with different probability of connections, as well as graphs with distance-dependent connectivity) and for networks that have balanced excitatory and inhibitory connections (on average), as well as completely unbalanced ones, namely, all-excitatory or all-inhibitory networks. We emphasize that the predictive accuracy of M^* stems from its ability to capture the geometry of the functional space of networks and not from the computational expressive power of this model (see the Appendix and Supplemental Fig. S8 [43]).

The sparse structure of the Mahalanobis matrix M^* for networks of four excitatory neurons [Fig. 3(a)] reflects the architectural features that the model relies on. To uncover what these features are, we use the fact that M^* is a positive-semidefinite matrix that can be decomposed uniquely into the product of a lower triangular matrix and its conjugate transpose, or *Cholesky factor*, $M^* = RR^T$, such that R is a lower-triangular matrix. We can use R to find a decomposition of M^* that is easier to interpret: $M^* = LL^T$, with $L = RU$ and U is unitary matrix that makes L as sparse as possible. In other words, right multiplying R by any unitary matrix U results in a decomposition $M^* = (RU)(RU)^T$, which means $\|R^T g - R^T g'\|^2 = \|(RU)^T g - (RU)^T g'\|^2$. We then solve the constrained optimization problem:

$$L = \underset{U \in \{Q | QQ^T = I\}}{\operatorname{argmin}} \|RU\|_1 \tag{5}$$

over the manifold of all possible unitary matrices [51,52] and find a matrix U such that $L = RU$ is maximally sparse and yet remains an exact decomposition of M^*. Using this sparse decomposition, the distance between networks can be rewritten as $\|L^T g - L^T g'\|_2^2$, which means that L^T implements a linear transformation that represents each network using a set of structural features. In this view, our model measures the squared Euclidean distance between networks in the feature space induced by L^T. The interpretation of the structural features extracted by L [Fig. 3(b)]

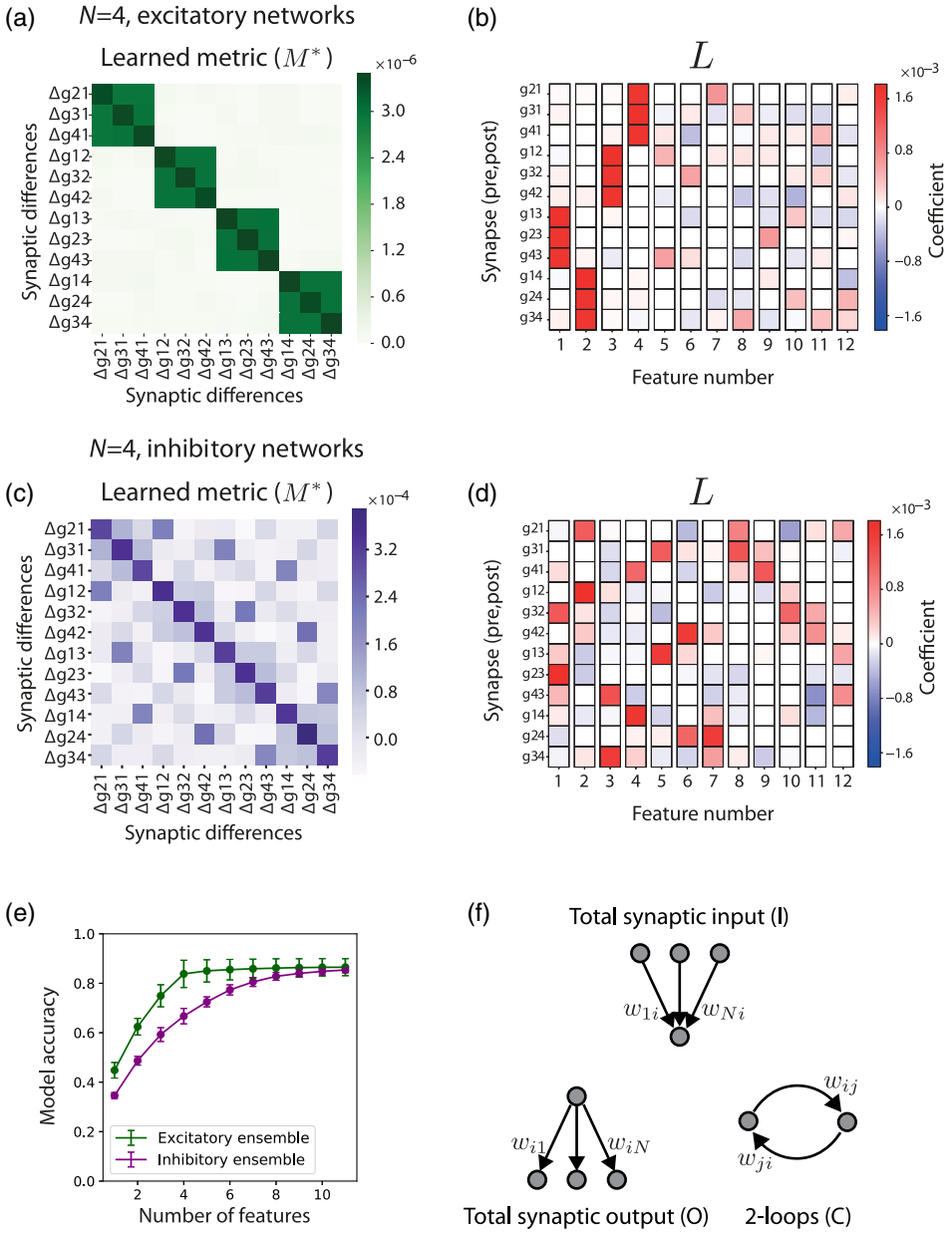

FIG. 3. Identifying the informative structural features that govern the M^* learned model of network similarity. (a) The optimal Mahalanobis matrix M^* for networks of $N = 4$ excitatory neurons. Each entry in the matrix is the weight assigned to a pairwise term in the bilinear function. (b) The sparsest L matrix from Cholesky factor-based decomposition of M^* from (a). Each column represent a single structural feature. Matrix entries correspond to the weights of the different synapses in each feature. The first four features correspond to the total synaptic inputs to each of the neurons; columns are sorted by their Euclidean norm. (c),(d) The same as (a),(b) but for networks of four inhibitory neurons. Here, loops of length 2 emerge as the dominant structural features. (e) Accuracy of the model when using only the k most important features (ones that have the largest Euclidean norm), for both ensembles of networks of size $N = 4$ (error bars represent standard deviations over different stimuli). Accuracy is defined as the Pearson correlation between the model's prediction and D_{func}. (f) An illustration of the most informative types of structural features.

is facilitated by the fact that each column of L defines a linear combination of synaptic weights. For the ensemble of excitatory networks of size $N = 4$, a dominant type of structural features stands out—one of the form $\sum_l G_{lk}$, or the *sum of synaptic inputs* to the kth neuron (this corresponds to the *indegree* in unweighted networks). Thus, for example, the left column of L in Fig. 3(b) is akin to the sum of synapses going into the third neuron. We identify two other types of features when using other stimuli and for networks with inhibitory neurons: the sum of outgoing synapses from a neuron, $\sum_k G_{lk}$, and the sum of synaptic weights along loops of size 2, $G_{kl} + G_{lk}$; the last type is dominant, for example, for the ensemble of all inhibitory networks of size $N = 4$, as shown in Figs. 3(c) and 3(d).

Since each column of L corresponds to a single structural feature, we further ask how many features are needed to accurately approximate the functional dissimilarity. We sort the columns of L by their norms (as vectors) and approximate M^* using the kth highest norms columns, for different values of k. We find that, for the ensemble of purely excitatory networks, the number of required structural features is small and close to the number of *neurons* rather than the number of *synapses*, whereas for the ensemble of purely inhibitory ones, a larger number of features is required [Fig. 3(e)]. For balanced networks of 15 neurons, a similar saturation of performance for a small number of features is observed [Fig. 4(a)].

Based on these results, we fit a new model that uses just the dominant structural features of L to approximate $D_{\text{func}}(G, G'|s)$:

$$D_{\text{features}}(G, G'|s) = \sum_l \alpha_l \cdot \underbrace{\left(\sum_k \Delta G_{kl}\right)^2}_{\substack{l\text{th neuron} \\ \text{synaptic input}}} + \sum_k \beta_k \cdot \underbrace{\left(\sum_l \Delta G_{kl}\right)^2}_{\substack{k\text{th neuron} \\ \text{synaptic output}}} + \sum_{k<l} \gamma_{kl} \underbrace{(\Delta G_{kl} + \Delta G_{lk})^2}_{kl\text{-neuron pair}}, \tag{6}$$

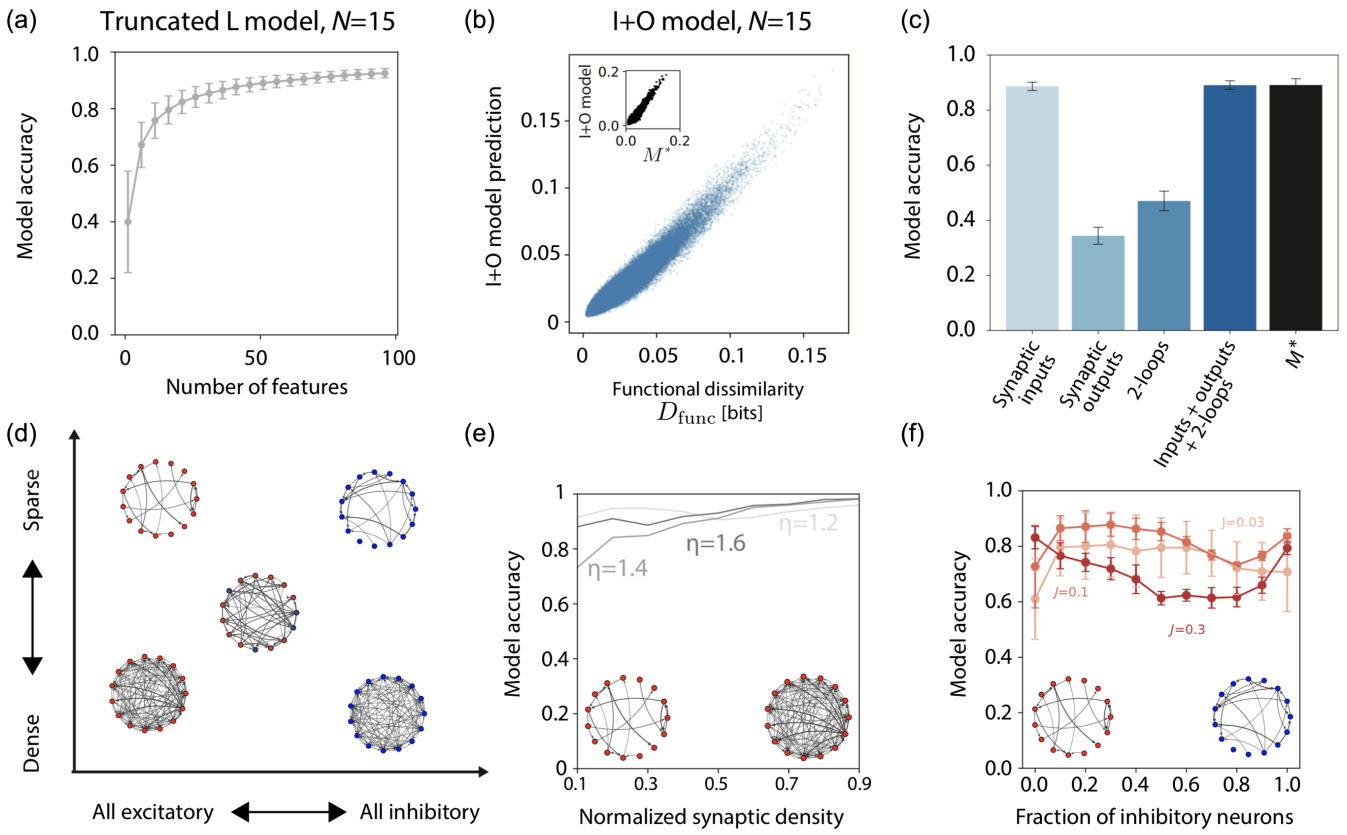

FIG. 4. High performance of the models of the functional similarity of networks, based on our learned architectural features, across different structural parameters. (a) Accuracy of the model when using the first k most important features (largest Euclidean norm), for the ensemble of balanced networks of $N = 15$ neurons (error bars represent standard deviations over ten different stimuli, spaced between $\eta = 1.0$ and $\eta = 1.8$). Accuracy is defined as the Pearson correlation between the model's prediction of similarity and D_{func}. (b) The performance of a model based only on the sum of synaptic inputs to each neuron and the sum of synaptic outputs of each neuron ($I + O$ model) predicts the functional similarity of networks of $N = 15$ neurons as well as the full model using the full matrix M^*; the inset shows the high correlation between the M^* model and the $I + O$ model (stimulus strength $\eta = 1.5$). (c) Prediction accuracy of models using different subsets of structural features, measured by the mean Pearson correlation coefficient with D_{func}. (d) Illustration of the space of network parameters for which we compute the accuracy of the feature-based model. The set of network parameters explored is characterized by the probability of forming a synapse between any two neurons (sparseness) and the ratio between excitation and inhibition in the networks. The drawn networks show examples of networks with the corresponding parameters in different parts of the space. (e) Accuracy of the M^* model as a function of sparseness of connectivity in the networks, for 3 different stimulus values. (f) Accuracy of the feature-based model that uses the sum of synaptic inputs and sum of synaptic outputs per each neuron ($I + O$), as a function of the ratio of excitation and inhibition in the network for three different values of J, the mean strength of excitatory synapses in the ensemble (error bars correspond to standard deviations across stimuli).

where α_l, β_k, and γ_{kl} are learned from the data to minimize the mean squared difference between the two measures. For many stimuli, the feature-based approximation is close to the performance of the full learned matrix M^*. Furthermore, an even simpler class of models that relies just on the first two terms of Eq. (6)—namely, the set of sums of synaptic inputs and sums of synaptic outputs of each of the neurons—is nearly as good, as shown in Figs. 4(b) and 4(c).

Theoretical studies of networks that have balanced excitation and inhibition link the spectral properties of the connectivity matrix of such networks and their dynamics [53–56]. To investigate whether the predictive power of the sum of synaptic inputs per neuron and the sum of synaptic outputs per neuron originate or depend on the assumption of *E-I* balance, we repeat the analysis above for ensembles of networks of 15 neurons with varying degrees of sparseness and different balances of excitation and inhibition. We span the full range from all-excitatory networks, through balanced networks, to all inhibitory-networks [Fig. 4(d)] and explore different stimulus' strengths, as well as different synaptic strengths [Figs. 4(e) and 4(f)]. We find that the accuracy of the feature-based model in predicting the similarity of networks is consistent across all these cases and parameter values. In particular, for larger values of synaptic strengths, the accuracy is higher for the unbalanced networks, suggesting that the theoretical framework of firing-rate-based balanced *E-I* networks is not sufficient to explain the accuracy of our models.

C. Scaling and universality of the functional space of networks for different classes of stimuli

The results above show the success of learning a feature-based metric for networks for a specific set of stimuli, where all neurons receive inputs from different realizations of Poisson-distributed spike trains with the same rate. Next, we ask how the functional similarity between networks depends on the nature of the stimulus used and, in particular, whether the functional metric we learn for one class of stimuli would generalize to other stimuli. We, therefore, repeat the mapping of functional distances between networks, D_{func}, for a wide range of Poisson inputs with different mean values—from weak stimuli that elicit almost no responses to strong stimuli that elicit very high firing rates (see the Appendix and Supplemental Fig. S2 [43]) and from uncorrelated inputs to the different neurons to highly correlated ones [Fig. 5(a)]. We find that the distances between networks are highly correlated for different stimuli: Figure 5(b) shows an example of the tight correlation of $D_{\text{func}}(G, G'|s)$ for two different stimuli, over all pairs of networks in the ensemble. Figure 5(c) shows the strong correlation between the distances of networks for many different pairs of stimuli, which are high and significant for all stimulus classes we test (see the Appendix and Supplemental Fig. S9 [43] for the correlation values of the dissimilarity of all the stimuli we use throughout the main text). This implies that, on average, changing the stimulus changes the dissimilarity between networks in a way that preserves the relations between their relative distances: Networks that are relatively close (far) under one stimulus *tend* to be relatively close (far) under a different stimulus. Figure 5(d) shows this explicitly by using 2D embedding of networks based on their functional distances, using an example of the same randomly selected four networks under five different stimuli—reflecting that the structure of relations between networks is preserved, while the overall map of distances may stretch or shrink.

The approximate scaling of the map of functional distances with stimulus strength suggests that the structure of the Mahalanobis matrix M^* should also remain stable, up to a stimulus-dependent multiplicative factor. Figures 5(e) and 5(f) show the accuracy of the models based on the sum of synaptic inputs and sum of synaptic outputs, across the space of stimuli we explore. Indeed, for the vast majority of stimulus parameters, the prediction of our model and the empirical D_{func} values are highly correlated. Interestingly, the relative importance of the total synaptic inputs and total synaptic outputs changes across stimulus space, and a transition from the domination of the total synaptic output values to the total synaptic inputs occurs as the stimuli become stronger (see the Appendix and Supplemental Fig. S10 [43]).

In the analyses above, the inputs to the networks are sampled from Poisson-distributed spike train that have a similar input rate for all neurons. While commonly used in many studies of models of neural activity, this set of inputs lacks the heterogeneity of stimuli impinging onto real circuits and misses on the complex temporal structures that characterize *in vivo* neural activity. To verify that our result does not depend on these simplified sets of stimuli, we repeat the analysis using more biologically realistic inputs: We use real spike trains recorded *in vivo* from neurons in the visual cortex of mice presented with a natural movie (see Sec. IV). Each of the neurons in the networks in our ensembles receives spikes from a different randomly sampled subset of the recorded units. Thus, each neuron receives its unique time-varying input with differently modulated mean rate [Fig. 6(a)]. Again, we find that our feature-based model is highly predictive of the functional dissimilarity of networks [Fig. 6(b)], generalizing our results to more biologically realistic stimuli. Model accuracy remains high for different values of input parameters, namely, the feed-forward connection probability and synaptic strength (see Supplemental Fig. S11 [43]). Results are shown for networks with $J = 1$ mV; different values of synaptic strength ($J = 0.1, 0.3, 0.5$ mV) result in different distributions of pairwise functional similarities, but the accuracy of the $I + O$ model remains high.

To reflect the effect of the temporal structure of the stimulus on the measured functional distances, we choose three specific networks from the same ensemble, such that

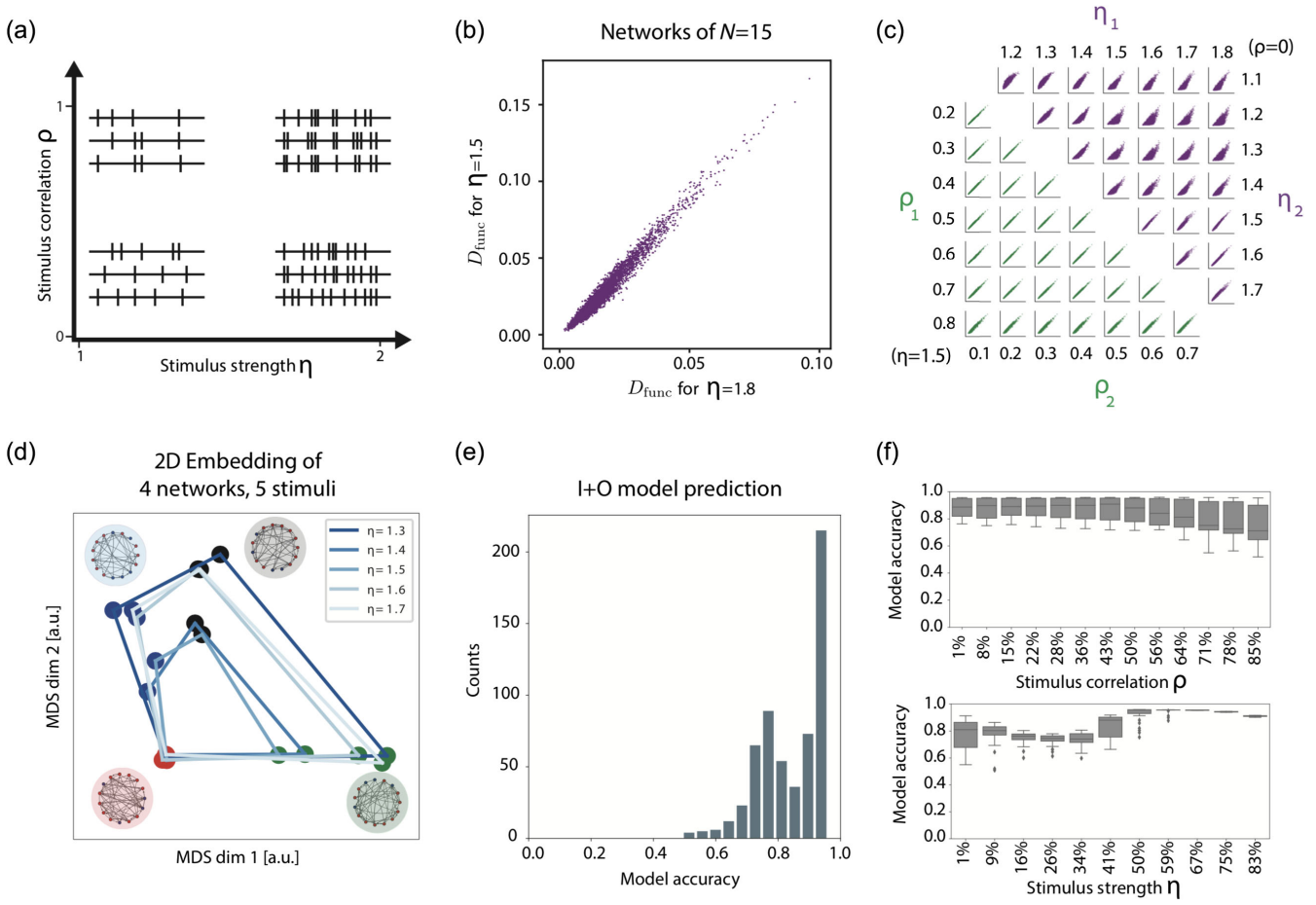

FIG. 5. The structure of the map of functional similarities between networks is conserved across stimuli. (a) Illustration of the space of different stimuli for which we compute the pairwise similarity between networks. The set of stimuli explored is characterized by the average rate of the external Poisson input to the neurons, η, and the correlation between different inputs to different neurons, ρ. (b) An example of the correspondence of $D_{\rm func}$ for 1 million pairs of networks in the ensemble of networks of size $N = 15$ is shown for two different stimuli; every dot represents one pair of networks. (c) The same as (b) but for many pairs of stimuli, from the parametric space of stimuli, described in (a). The correlations between the distances between all pairs of networks are shown for the whole range of ρ and η values. Green panels show the case of varying ρ, at fixed $\eta = 1.5$; purple panels varying η at fixed $\rho = 0$. (d) Five overlaid 2D MDS embeddings of four example networks of size $N = 15$, based on the functional dissimilarity between them. The networks in each case are the same and are marked by a dot of the same color. The colors of the lines between nodes denotes which of the five stimuli this embedding relates to. Overlaying is done by anchoring the network marked by the red dot. Overlaid maps show the geometric organization is preserved across stimuli space. (e) Pearson correlation between the computed functional similarity $D_{\rm func}$ between networks and predictions of the models based on the sum of synaptic inputs and sum of synaptic outputs of each neuron, for 900 different combinations of (η, ρ) for networks of size $N = 15$. (f) The interquartile ranges of the data shown in (e) are shown by aggregating the values shown in (e) over η (top) and ρ (bottom). Percentage values of each bar denote the parameter value between lowest strength or correlation (0%) and highest ones (100%).

network A and network B have similar responses (low $D_{\rm func}$), whereas networks A and C have dissimilar responses (high $D_{\rm func}$)—shown in Fig. 6(c). Each of these specific networks has three different inhibitory neurons. For reference, we also simulate an "empty" network where none of the neurons are connected to each other (all weights are zero), which, by construction, gives the neuronal activity due to the stimulus alone. A sample of the responses of the corresponding neurons in these different networks is shown in Fig. 6(d). To assess the impact of temporal correlations between neurons on the similarity of their activity, we also consider mock responses of the networks, where we shuffle each spiking pattern of each neuron in the network in time, giving responses that preserve firing rates of each neuron but no stimulus-induced or structure-induced correlations. We then compute the dissimilarity matrix between all eight response distributions (four original networks and their corresponding shuffled responses). The functional similarity of the responses to the natural stimulus due to structural features $[D_{\rm func}(A, C)$ and $D_{\rm func}(B, C)]$ are similar to the functional similarity between a network and its shuffled variant

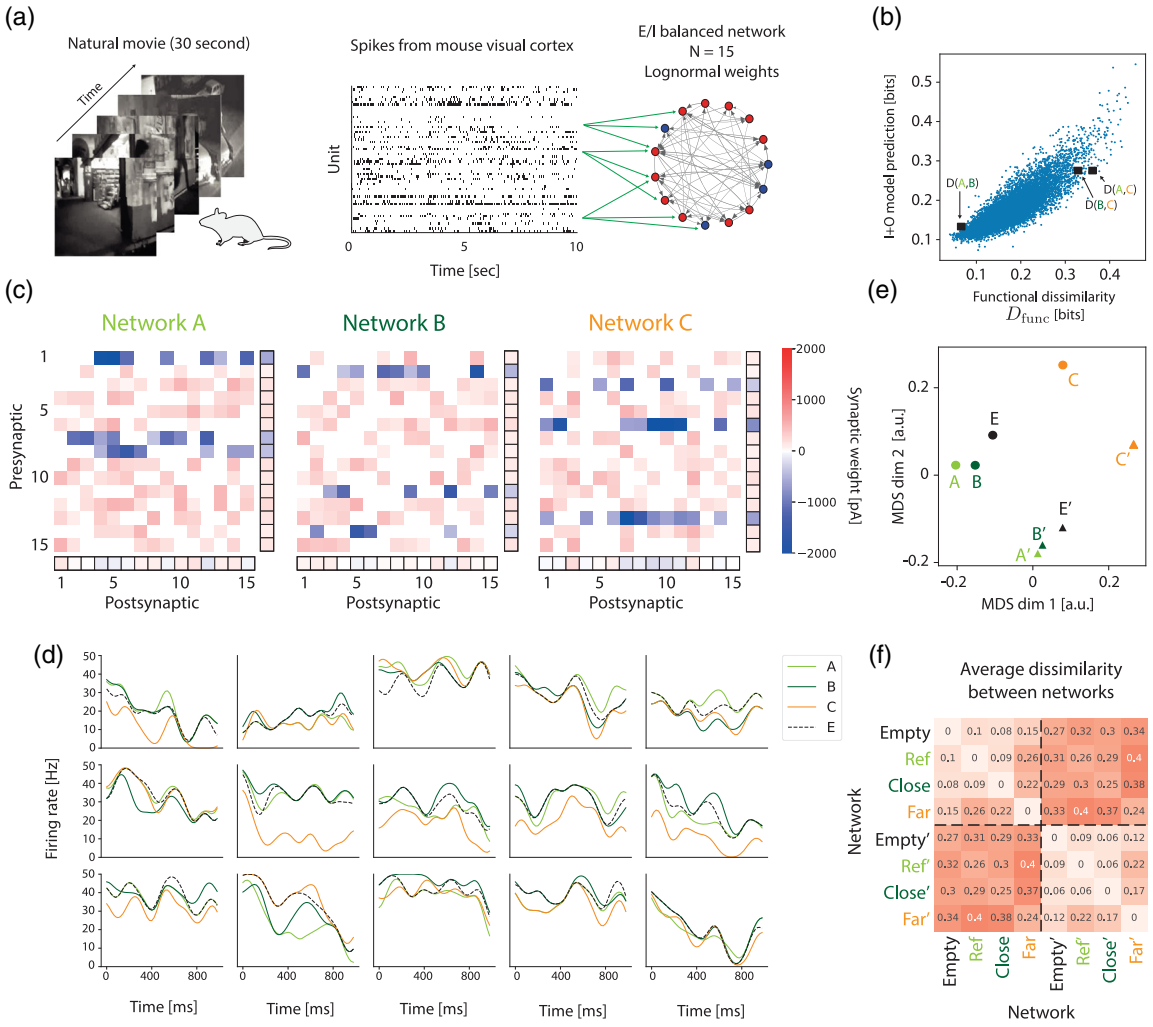

FIG. 6. Feature-based model predicts functional similarity under natural stimuli. (a) Our ensemble of networks are stimulated with spike trains recorded from mouse *in vivo* (see Sec. IV). Each neuron in the networks is stimulated by a randomly selected set of approximately 15 neurons out of the 59 recorded units from the visual cortex of a mouse presented with a 30-s-long natural movie. (b) The prediction of the feature-based $I + O$ model for the pairwise similarity of 10 000 balanced networks of 15 neurons, stimulated with spike trains as described in (a), is highly correlated with the functional similarity of the responses of networks, D_{func}. (c) Connectivity matrices for three specific networks from the ensemble (described in the main text). Each entry is a synaptic weight (red, excitatory; blue, inhibitory). A and B have similar responses to the stimuli, as measured by D_{func}, while A and C have dissimilar responses. The connectivity matrix of the empty network is not shown. The mean synaptic input (bottom) and output (right) of each neuron are shown for each network. (d) Firing rates of the 15 neurons in all four networks in the first second of the simulation, obtained by convolving individual spike trains with a Gaussian kernel (A, green; B, dark green; C, orange; empty, black dashed line). Kernel standard deviation is 60 ms (three time bins). (e) 2D MDS embedding of networks A, B, and C, the "empty" network, and time-shuffled variants of their responses (denoted as A', B', C', and empty′; see the main text). (f) The functional dissimilarity matrix between a reference network (similar to A above), the functionally closest and farthest networks with the same number of inhibitory neurons (the same as B/C above), the empty network (E above), and the shuffled variants of their responses, averaged over 100 reference networks.

[e.g., $D_{\text{func}}(A, A_{\text{shuffled}})$] as shown by multidimensional scaling (MDS) embedding of the eight networks [Fig. 6(e)]. This relation replicates over many different groups of such networks [Fig. 6(f)], namely, that shuffling the spikes of individual neurons in a network results in network responses that are as remote from the original network as the differences between two remote networks responding to the same stimulus. Moreover, it is especially interesting to note that the correlation between the functional similarity of networks measured by the overlap of the PSTH of corresponding neurons and the functional similarity measured by D_{func} as in Eq. (6) is 0.85, whereas the correlation between D_{func} and the prediction of the $I + O$ model is also 0.85 (the correlation between the $I + O$ model and PSTH similarity is 0.83)—reflecting that our feature-based model is as accurate in predicting the networks' responses as does knowing the overlap of the PSTHs of the neurons of networks. We also repeat the analysis for networks in which the identity of the

inhibitory and excitatory neurons in the network is fixed (neurons 1–12 are excitatory, and neurons 13–15 are inhibitory for all networks in the ensemble) and obtain similar results (see Supplemental Fig. S12 [43]).

D. Predicting the similarity of large networks from their structural features

We next ask whether the structural features we identify for small networks generalize to large ones. Notably, for networks of 100 neurons, there are 2^{9900} possible unsigned topologies and 2^{100} possible activity patterns of the network. Thus, whereas mapping all networks is intractable already for 15 neurons, here even measuring the similarity of networks using the overlap of their sampled response distributions becomes impractical. We, therefore, fit second-order maximum entropy models for each network's activity (see Sec. IV) and use them to measure the functional dissimilarity between networks G and G' by the average of the log-likelihood ratio of the spiking response of G under a model that is learned from the spiking responses of G', and vice versa:

$$D_{\text{func}}(G, G'|s) = \frac{1}{2T}\sum_{i=1}^{T}\left[\log\frac{P_G(x_i)}{P_{G'}(x_i)} + \log\frac{P_{G'}(x_i')}{P_G(x_i')}\right], \quad (7)$$

where x_i (x_i') denotes the ith activity pattern of network G (G'). This measure converges to the symmetric

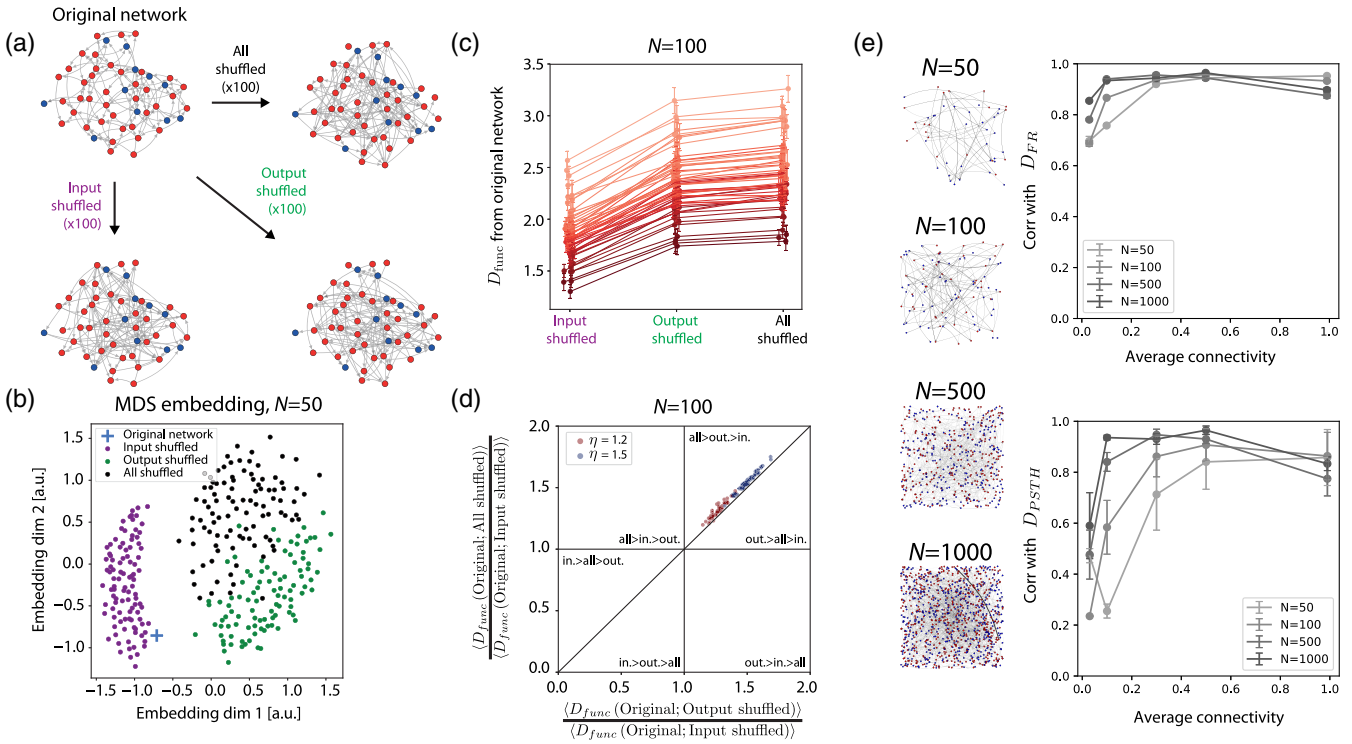

FIG. 7. The architectural features identified in small networks predict functional similarity of networks of 50, 100, and 1000 neurons. (a) We pick 100 randomly connected networks of size $N = 50$ and $N = 100$ and use each of these networks as a "template." From each such original network, we create 100 variants preserving the sum of synaptic inputs to each neuron, by shuffling the source of the synapses into each neuron (red), 100 variants preserving the sum of synaptic outputs of each neuron, by shuffling outgoing synapses (green), and 100 variants where all synapses shuffle (black). (b) A 2D MDS embedding of one original network of 50 neurons and its 300 variants, based on their log-likelihood-based functional similarity values. (c) D_{func} between each template network of 100 neurons and its three types of shuffled variants. Each of the 50 lines using different shades of red connects the values of the mean D_{func} between one original template network and its 100 shuffled variants of each type. Error bars represent one standard deviation over the set of variants. (Small random horizontal jitter is added to the points for clarity.) (d) For each of the original networks, we show the ratio between its average D_{func} to all the output shuffle variants against the average of its D_{func} to all the shuffle variants, normalizing both values for each network by its average D_{func} to input shuffled variants. The colored dots mark the values for the networks under two different stimuli, showing that all 100 original networks reside in the part of the "phase space" where their input-shuffled networks are closer to the original network than the output-shuffled ones, which are closer than the all-shuffled ones. (The relations between variants in other parts of the phase space are labeled in each of the corresponding parts of the figure.) (e) Left: examples of E-I balanced networks of 50, 100, 500, and 1000 neurons, for which we compare functional similarity measures and feature-based models. Right: The dissimilarity between networks based on the respective firing rates of the neurons in the networks (top) and by the dissimilarity of their PSTH (bottom) is presented against the prediction of the $I + O$ feature-based models, for different levels of sparseness of synaptic connectivity and strengths of the stimulus.

Kullback-Leibler (KL) divergence between conditional response distributions, as the number of samples increases.

We use randomly sampled networks of 50 and 100 neurons and create 300 variants for each one of these origin networks, by manipulating their connectivity graphs [Fig. 7(a)]: (i) 100 variants in which the sum of synaptic inputs to each neuron is preserved, but the incoming synapses to each neuron are assigned to a random neuron in the network (shuffle inputs). (ii) 100 variants that preserve the sum of synaptic outputs of each neuron, but the outgoing synapses of each neuron are randomly connected to the other neurons (shuffling the outputs). (iii) 100 variants in which all synapses are randomly shuffled, regardless of their presynaptic or postsynaptic neuron. As before, all networks are presented with the same stimulus and have identical initial conditions. Figure 7(b) shows the 2D MDS embeddings of all the variant networks of one original network of 50 neurons, based on their functional dissimilarity matrix—where closer points represent networks that have similar conditional response distributions. Figure 7(c) shows for 50 randomly chosen original networks, each with $N = 100$ neurons, that variant networks with the same total synaptic inputs to each neuron (input shuffle) are functionally more similar (smaller D_{func}) to their original network compared to networks with the same total synaptic outputs for each neuron, which, in turn, are more similar than networks where all synapses of the original network are shuffled. Figure 7(d) shows this is the case for all the networks we test and that this result replicates under different stimuli.

We then aim at scaling the analysis to even larger networks. Learning a functional metric as in Eq. (4) and the decomposition of such matrices is impractical, since M^* scales as $O(N^4)$. Moreover, even measuring the functional similarity of networks of more than 100 neurons is challenging. Instead, we ask whether the architectural features we identify for small networks would generalize to large ones. We find that, using models based on these features [Eq. (6)], we can accurately predict functional dissimilarity between networks of 50, 100, 500, and 1000 neurons with different probabilities of synaptic connectivity [Figs. 7(e) and 7(f)], both when we quantify their functional dissimilarity using the differences in firing rates of the respective neurons and when we use a PSTH-based dissimilarity measure (see Sec. IV).

III. DISCUSSION

Measuring the functional similarity of small neural networks in terms of their population spiking patterns allows us to learn a structural metric on the space of networks that predicts functional similarity with high accuracy. Our learned metric outperforms a wide range of commonly used graph-theory metrics, by very large margins. We then use the learned metric to identify features of the topology of networks that govern their function. Surprisingly, while the full synaptic connectivity map for n neurons is of size n^2, only a small number of architectural features emerge as shaping the function of networks: the total synaptic sum of inputs and total synaptic sum of outputs for each neuron. This set of $2n$ features is highly predictive for small networks and generalized well to predict the functional similarity of networks of up to 1000 neurons. We further find that the "geometrical organization" of distances between networks is highly correlated over a wide range of stimulus strengths and correlations. We emphasize that, unlike previous studies, our work quantitatively describes and models the relations between structure and function, and the metric-learning-based approach allows us to provide a geometrical interpretation for fundamental neural processes such as learning, development, and plasticity in brain networks.

We note that, while many of the ensembles of networks we analyze above are balanced on average in terms of their excitatory and inhibitory synapses, our results do not rely on an assumption of *E-I* balance. In particular, our learned metric and features are accurate for all-excitatory or all-inhibitory ensembles of spiking neural networks, as well as different combinations of excitatory and inhibitory rations, for different random graph models, and for different classes of stimuli, including inputs that are taken from real spike trains. We hope that future theoretical work would bring together the theoretical work on rate-based *E-I* balanced networks [39,41] and that on inhibitory threshold linear networks [57] with our results on a spiking neural network under nonbalanced networks and realistic stimuli. Another potential direction of theoretical analysis would relate our results to the functional similarity of other biological networks that differ in their connectivity patterns and single-element dynamics (e.g., Ref. [37]).

Our results imply that within the space of networks there are subspaces or manifolds that retain similar functional properties and that these manifolds are the ones which contain networks that have the same sum of synaptic inputs and sum of synaptic outputs. A network that changes its synaptic connections along a trajectory contained within such a manifold could, therefore, be very different structurally, without changing its function. This implies a "neutral evolution" path for neural networks' organization and learning dynamics [58]. Interestingly, our results conform with the synaptic homeostatic mechanisms that have been extensively studied both experimentally and theoretically [59,60]. In particular, our model predicts that homeostatic mechanisms that redistribute synaptic weights but preserve the total synaptic inputs to a neuron or its outputs may not only shape the computational properties of single cells, but play a crucial role in learning, plasticity, and development at the level of the network. We reiterate, however, that the models based on the total synaptic inputs and total synaptic outputs capture most but not all of the functional similarity between networks. Thus, "perfect

homeostatic" changes would not result in functionally identical networks.

While there are many potential extensions of our work in terms of more biologically realistic neurons and neuronal classes, we note that our analysis mostly focuses on cases where all neurons in a network receive external stimuli that have similar statistical properties. Extending our work to the general case of arbitrary stimuli would require learning a joint metric for stimulus space and the space of network architectures. These are likely to reveal new, more intricate design features of neural circuits. In addition, while our analysis focuses on stimulus values that do not result in pathological behavior of the network (silence or epilepsy), the framework we introduce may be applied to identify which network topologies would be susceptible to such events. We also note that our ensembles of networks regard the neurons as "identified" in the sense that there is a unique correspondence between neurons in different networks. The comparison of networks without such identification would be interesting. An even more general case to consider would be the comparison of networks of different sizes.

The metric we learn gives accurate results by relying only on the differences between corresponding synapses of two networks. To explain the small residual part of the functional similarity between networks that our metric does not account for would require to go beyond pairwise relations between synapses. Such extensions could elucidate the functional importance of longer loops and global structural properties, which our current model cannot account for. Moreover, our metric also implies the functional difference between networks would be "translation invariant" in δG, i.e., $D_{\text{func}}(G, G'|s) = D_{\text{func}}(G + \delta G, G' + \delta G|s)$ for any δG. This is unlikely to be true, in general, especially for large δG. We, thus, expect that refinements of the approach we present here, such as learning different local metrics, would be important for analyzing larger networks and real networks, and likely reveal additional design principles.

The growing abundance of connectomics data makes it imperative to combine theoretically grounded computational frameworks and experimental measurements in understanding networks' structure and function. The approach we present here is a step toward building such a framework. While we rely on simulated networks, these are models of spiking neurons, and so we believe our results reflect fundamental design principles of real neural networks. Moreover, we have focused here mostly on small networks, where sampling and evaluation of our models is relatively simple, but even for networks of 1000 neurons that we study here there are 2^{999000} possible directed topologies. Thus, sampling and evaluating the success of our approach for even larger networks present considerable computational challenges—which would require scaling of the learning and cross-validation, at the least. Ultimately, this kind of approach would be extendable to asking how to design neural circuits that would perform a desired function or how to interface and control neural circuits in the brain. Finally, we note that the framework we present here is immediately extendable to studying artificial neural networks and other biological and nonbiological networks.

IV. MATERIALS AND METHODS

A. Simulating neural networks

All networks are simulated using the NEST simulator for spiking neural network models [61]. Networks of four neurons—either all excitatory or all inhibitory ones—are simulated using the following parameters for an integrate and fire neuron model (which are the default integrate and fire model parameters in NEST; "iaf psc alpha"): resting membrane potential $E_L = -70$ mV; membrane capacity $C_m = 250$ pF; membrane time constant $\tau_m = 10$ ms; refractory period $\tau_{\text{ref}} = 2$ ms; spike threshold $V_{\text{th}} = -55$ mV; reset potential $V_{\text{reset}} = -70$ mV; rise time of the synaptic alpha function $\tau_{\text{syn}} = 2$ ms. Networks of 15 excitatory and inhibitory neurons are simulated using the biophysical parameters taken from Ref. [41]: $E_L = 0$ mV; $C_m = 250$ pF; $\tau_m = 20$ ms; $\tau_{\text{ref}} = 2$ ms; $V_{\text{th}} = 20$ mV; $V_{\text{reset}} = 0$ mV; $\tau_{\text{syn}} = 0.5$ ms. Networks of 50 and 100 neurons are simulated using the same biophysical parameters as for networks of 15 neurons.

For the given values of the synaptic time constant, membrane time constant, and membrane capacitance, we compute the amplitude of the postsynaptic potential J_{unit} for a synaptic input current of 1 pA. Mean synaptic weights are then chosen such that an input spike to a neuron results in a 0.1 mV increase in its membrane potential, corresponding to $w_{\text{syn}} = (0.1/J_{\text{unit}})$ [42]. Other mean synaptic weights (ranging from 0.05 to 0.5 mV increase per spike) are also considered and give similar results.

The external stimulus to each neuron is a sequence of spikes drawn from a Poisson distribution with rate λ_i, with a fixed synaptic strength w_{input}. The synaptic weight from the inputs to the circuit is chosen such that an input spike to a neuron results in a 0.1 mV increase in its membrane potential, resulting in $w_{\text{input}} = 20$ pA for the biophysical parameters used for networks of $N = 15$ neurons and $w_{\text{input}} = 8$ pA for the parameters of networks of $N = 4$. These biophysical parameters and input weights induce threshold rates (the external rate that fixes the membrane potential of the receiving neuron around its threshold) of 8.8 and 7.2 kHz for networks of $N = 15$ and $N = 4$, respectively. Input rates are then defined as $\lambda_i = \eta \cdot \lambda_{\text{th}}$, where λ_{th} is the threshold rate and η is the ratio between the input rate and the threshold rate, as in Ref. [41]. Simulated stimulus values, therefore, correspond to inputs from thousands of external neurons (see the Appendix and Supplemental Fig. S2 [43]). Correlated stimuli are generated using a multiple interaction process [62]; we

consider stimulus correlations ranging from 0 (independent stimuli) to 0.99.

B. Generating ensembles of networks

For the all-excitatory and all-inhibitory networks of four neurons, we simulate all 4096 unweighted and directed topologies. For networks of 15 neurons, the number of different unweighted directed topologies is larger than 10^{63}. We, therefore, estimate the similarity metrics for an ensemble of 10 000 unweighted directed networks, sampled from an Erdős-Rényi random graph model [12], where the probability of forming a synapse between any two neurons in a network is $p_{\mathrm{syn}} = 0.5$. Each of the 15 neurons is randomly assigned a "type"—excitatory or inhibitory—where the probability of a random cell being inhibitory is $p_{\mathrm{inhib}} = 0.2$. This determines the sign of each synapse in the network. Synaptic weights are drawn from a log-normal distribution with parameter $\sigma = 0.5$ [42]. For $w_{\mathrm{syn}} = 20$ pA ($J = 0.1$ mV), the resulting standard deviation of the log-normal distribution is approximately 9 pA. The mean strength of inhibitory synapses is scaled to ensure $p_{\mathrm{inhib}}\langle w_{\mathrm{inhib}}\rangle = (1 - p_{\mathrm{inhib}})\langle w_{\mathrm{ex}}\rangle$, where the average is over the entire ensemble of networks. Networks of 50 or 100 neurons are randomly generated in a similar manner such that the probability of forming a synapse between two arbitrary neurons is $p_{\mathrm{syn}} = 0.1$ and $p_{\mathrm{inhib}} = 0.2$. ($p_{\mathrm{syn}} = 0.5$ and $p_{\mathrm{syn}} = 0.05$ are also considered and give similar results.) "Geometric networks" are randomly generated by sampling 15 points in $[0, 1]^3$ (uniform sampling, different points for each network), computing their Euclidean pairwise distances, and sampling a synapse from neuron i to neuron j with probability $e^{-\|\mathbf{r_i}-\mathbf{r_j}\|/D}$. D is chosen such that the marginal edge probability is 0.5, similarly to the value used for the nongeometric ensemble.

C. Permutation of synapses for networks of 50 and 100 neurons

The original random networks (see the main text) are shuffled to generate variants that retain the total sum of synaptic input to the ith neuron, by permuting all 49 (or 99) off-diagonal elements of the ith column in the network's connectivity matrix. Variants that retain the total sum of synaptic outputs are similarly generated by permuting off-diagonal elements of each rows. Control networks are generated by shuffling all off-diagonal elements, irrespective of their original row or column.

D. Fitting pairwise maximum entropy models

For large networks, direct estimation of $P_{\mathrm{emp}}(\mathbf{x}|s)$ requires prohibitive or unrealistic long stimuli (since, for a network of N neurons, the number of possible activity patterns is 2^N). We, therefore, use pairwise maximum entropy models to describe the response distributions. These models are shown to be highly accurate for the activity patterns of groups of this scale [10], and we validate that this was the case for our networks. For each network and stimulus, we learn the maximum entropy model of its responses to the stimulus based on the firing rates and pairwise correlations between cells, given by $P(\mathbf{x}|s) = (1/Z)\exp[-\sum_i \alpha_i x_i - \sum_{i<j}\beta_{ij}x_i x_j]$; models are learned using the MaxEnt toolbox software [63].

E. Finding a sparse representation of structural features based on Cholesky decomposition

The Cholesky decomposition of a Hermitian positive-definite matrix [64] factorizes it uniquely into the product of a lower triangular matrix and its conjugate transpose. To interpret the structure of the matrix M^* based on the decomposition $M^* = R \cdot R^T$, we use the fact that right multiplying R by any unitary matrix U results in a decomposition $M^* = (RU)(RU)^T$, which means $\|R^T g - R^T g'\|^2 = \|(RU)^T g - (RU)^T g'\|^2$. We, therefore, solve the constrained optimization problem: $L = \mathrm{argmin}_{U\in\{Q|QQ^T=I\}}\|RU\|_1$ over the manifold of all possible unitary matrices [51,52] and find a matrix U such that $L = RU$ is maximally sparse and yet remains an exact decomposition of M.

F. Multidimensional scaling of networks

Given an $N \times N$ dissimilarity matrix D_{func} and a desired number of dimensions p ($p = 2$ in the main text), we find the $N \times p$ embedding matrix such that the pairwise distances in the p-dimensional space minimize the mean squared error with respect to the dissimilarities specified by D. MDS is implemented using the scikit-learn PYTHON package [65].

G. Divergences between probability distributions

The Kullback-Leibler divergence between probability distributions P and Q is defined as $D_{\mathrm{KL}}(P;Q) = \sum_x P(x)\log\left(P(x)/Q(x)\right)$, which measures the distinguishability of distributions in bits and has multiple information theory and statistics motivations and interpretations [66]. The Jensen-Shannon divergence [67] is a symmetric and bounded extension of the Kullback-Leibler divergence, defined as $D_{\mathrm{JS}}(P;Q) = \frac{1}{2}D_{\mathrm{KL}}(P;M) + \frac{1}{2}D_{\mathrm{KL}}(Q;M)$, where $M = (P + Q/2)$. This is a measure of dissimilarity between probability distributions, which is 0 bits for identical distributions and 1 bit for nonoverlapping distributions.

ACKNOWLEDGMENTS

We thank Roy Harpaz, Udi Karpas, Tal Tamir, Yoni Mayzel, Omri Camus, Benny Brazowski, Gašper Tkačik, and D. Allan Drummond for discussions, comments, and ideas. E. S. was supported by the European Research Council (ERC 311238 NEURO-POPCODE), Simons Collaboration on the Global Brain (542997), the Israel

Science Foundation (Grant No. 1629/12), the Israel–U.S. Binational Science Foundation, research support from Martin Kushner Schnur and Mr. and Mrs. Lawrence Feis, and the Joseph and Bessie Feinberg Professorial Chair.

APPENDIX: EXPANDED METHODS AND IMPLEMENTATION DETAILS

1. Identifying stimulus strengths that elicit networks' responses

We repeat the analysis of network responses and similarity for a wide range of external input rates—from very weak stimuli that elicit almost no spiking responses to very strong stimuli that elicit firing rates that saturate our chosen binning resolution. The relation between the firing rates and the stimulus strength parameter η, for the ensemble of 15 neurons, is shown in Supplemental Fig. S2 [43] and is used to set the range of stimulus values that are studied in detail along the manuscript. The networks analyzed in Figs. 1–3 in the main text use $\eta = 1.5$.

2. Different choice for temporal bin size, synaptic dynamics, and single-neuron models

To verify our results do not depend on a specific choice of the simulation configuration, we repeat the analysis in the main text for different simulation and analysis parameters.

(1) *Temporal bin sizes.*—The analysis presented in main text uses a time bin of $\Delta t = 20$ ms. We repeat the analysis with $\Delta t = 10$ and 40 ms.

(2) *Synaptic dynamics.*—Throughout the main text, the synaptic dynamics of the simulated networks follows an alpha function, $I(t) = w_{\mathrm{syn}} \cdot (t/\tau) \cdot e^{1-(t/\tau)}$ [68], in which postsynaptic potentials have a finite rise time (here, w_{syn} is the synaptic weight and corresponds to the amplitude of the excitatory post synaptic potential). We repeat the analysis with delta-function activation function of synapses [69], in which postsynaptic potential jumps on each spike arrival.

(3) *Neuron model.*—The neuron model used in the main text is a current-based model ("iaf_psc_alpha" in NEST [61]), in which subthreshold dynamics are linear and threshold crossing is followed by an absolute refractory period. Conductance-based networks are simulated using Brian 2 [70], and simulation parameters are adopted from Ref. [71].

Similar to the results presented in the main text, the model based on the full Mahalanobis matrix, the complete feature-based model, and the model based on the total synaptic inputs all show significantly higher correlation with D_{func} compared to other structural measures. Euclidean distance is shown in Supplemental Fig. S3 [43] as an example, but the behavior for other structural metrics is similar.

3. Predicting PSTH-based functional similarity

The similarity between the spiking responses of networks can be evaluated using different metrics, and the specific choice of a metric highlights different aspects of the neural code. It is not immediately clear whether the model defined in Eq. (3), which accurately predicts D_{func} (as defined in the main text), generalizes to other functional metrics. We, therefore, ask whether our feature-based model can predict a PSTH-based similarity measure of the neural responses.

We use the binarized spiking response of network G to stimulus s (a matrix of the form $\{0,1\}^{N\times T}$ as in Fig. 1, where N is the number of neurons and T is the number of time bins), which we denote by $x_G(s)$, to get the time-dependent firing rate or PSTH: For a given stimulus s and for each network G, we convolve the ith row of $x_G(s)$ with a 200-ms sliding window to get the PSTH of the ith neuron in network G, $r_G^i(t)$ (which is a vector of real numbers of size $T - W$, where W is the number of time bins corresponding to the window—10 for 200-ms window and 20-ms time bins). We then measure the PSTH-based functional dissimilarity between two networks as the average over neurons of the PSTH difference between the corresponding neurons in each network:

$$D_{\mathrm{PSTH}}(G, G'|s) = \frac{1}{N}\sum_{i=1}^{N} \|r_G^i(t) - r_{G'}^i(t)\|_2.$$

We find that D_{func} and D_{PSTH} are highly correlated across different stimuli. Consequently, our feature-based model accurately predicts the PSTH-based functional dissimilarity, as shown in Supplemental Fig. S4 [43].

4. Low-dimensional complexity of the dissimilarity matrix between networks

To assess the lower-dimensional structure of the dissimilarity matrix, we compare its spectrum to that of randomly shuffled controls. In Supplemental Fig. S5 [43], we show, for three representative example networks of size 15 responding to three different stimuli, that the spectrum of the original dissimilarity matrix ("empirical" in blue) decays significantly faster than that of a randomly shuffled version of these matrices ("control" in black). Thus, the lower-dimensional structure of these matrices depends on the relations between all the pairwise distances and not merely their overall distribution.

5. Comparing the predictive power of different common structural metrics

We explore a wide range of structural metrics and ask how well they predict the functional dissimilarity of network responses. Supplemental Fig. S6 [43] shows an extended version of Fig. 2 in the main text: Accuracy is measured by the correlation between the distances computed by each of the methods and the functional dissimilarity D_{func}. As in Fig. 2, the stimulus to each neuron is an independent realization of the same Poisson inputs, with

$N = 15$, $\eta = 1.5$, and $\rho = 0$. Our learned model (black) outperforms all metrics by a large margin.

6. Firing-rate-based dissimilarity measure

The firing rate r_i of the ith neuron in a network G of N neurons is defined as the ratio between the number of time bins in which the neuron spikes and the total number of time bins (equivalently, the marginal spike probability). The firing rate vector of each network is defined as the vector of firing rates of all N neurons, $\mathbf{r} = (r_1, \ldots, r_N)$. The firing-based dissimilarity between networks G and G' is taken to be the Euclidean distance between their firing rate vectors, $D_{\mathrm{FR}}(G, G') = \|\mathbf{r} - \mathbf{r}'\|$.

7. Constructing naturalistic stimuli from electrophysiological recordings

Spike trains are obtained from The Allen Brain Observatory, a database of the visually evoked functional responses of neurons in mouse visual cortex based on two-photon fluorescence imaging. We include 30-s data from 59 units with a signal-to-noise (SNR) ratio larger than 4, recorded using six neuropixel probes from a male mouse in response to a natural movie. To construct the stimulus to the network, we randomly sample a 59×15 feed-forward connectivity matrix, such that each column represents the subset of input units connected to each of the 15 neurons (different feed-forward connectivity values are considered; see Supplemental Fig. S11 [43]). Synaptic weights from the inputs to the networks are chosen to elicit nonpathological network response; different values are considered and give similar results (Supplemental Figure S11 [43]).

Details of the metrics used in the comparison are as follows.

(1) Metrics between continuous vectors:
   (i) L_1 (Manhattan), $\sum |x_i - y_i|$;
   (ii) Braycurtis, $\sum_i (|x_i - y_i|/|x_i + y_i|)$;
   (iii) Canberra, $\sum_i (|x_i - y_i|/|x_i| + |y_i|)$;
   (iv) L_∞ (Chebyshev), $\max |x_i - y_i|$.

(2) Metrics between binary vectors (applied to binarized connectivity matrix):
   (i) Dice, $\sum |x_i - y_i| / 2 \sum x_i \cdot y_i + \sum |x_i - y_i|$;
   (ii) Jaccard, $\sum |x_i - y_i| / \sum x_i \cdot y_i + \sum |x_i - y_i|$;
   (iii) Kulsinki, $\sum |x_i - y_i - | \sum x_i \cdot y_i + N(N-1) / \sum |x_i - y_i| + N(N-1)$;
   (iv) Rogers-Tanimoto, $2 \sum |x_i - y_i| / \sum x_i \cdot y_i + \sum (x_i - 1) \cdot (y_i - 1) + 2 \sum |x_i - y_i|$;
   (v) Russel-Rao, $N(N-1) - \sum x_i \cdot y_i / N(N-1)$;
   (vi) Sokal-Sneath, $2 \sum |x_i - y_i| / \sum x_i \cdot y_i + 2 \sum |x_i - y_i|$;
   (vii) Yule, $2 \sum |x_i - y_i| / \sum x_i \cdot y_i \cdot \sum (x_i - 1) \cdot (y_i - 1) + \sum |x_i - y_i|$.

All structural metrics are computed using the SciPy PYTHON package [72]. Spectral distances are computed as follows: For each connectivity matrix G, the directed graph Laplacian is computed using three different algorithms, as implemented by the networkx PYTHON package [73]. The real part of each Laplacian spectrum is computed and sorted, yielding a vector of size N. Finally, Euclidean pairwise distances are computed between the vectors [74].

8. Finding the optimal Mahalanobis matrix M^* under a regularized optimization

The optimal Mahalanobis matrix M^* is found by solving the optimization problem described in the main text [Eq. (2)]. For networks of size N, the number of parameters in M^* is $N^2 \cdot (N-1)^2$. To avoid overfitting, we use a regularization term that is weighted by a free parameter α. For networks of size 15, we solve the optimization problem with 50 different values of α, equally spaced on a logarithmic scale from 10^{-6} to 10^6. For each value of α, we estimate the accuracy of the resulting M^* on a held-out validation set (a quarter of the size of the train set). The value of α that gives the minimal loss for each stimulus is used to obtain M^* that are used in the main text; with some small variation across stimuli, $\alpha_{\mathrm{opt}} \approx 10^3$.

9. The optimal Mahalanobis M^* for shuffled data shows no structure

To verify that the success and accuracy of our learned Mahalanobis matrix M^* is not a result of overfitting or of overexpressive power of this model, we repeat the fitting procedure with shuffled functional dissimilarity values, by randomly permuting the entries of the dissimilarity matrix. For different values of the regularization parameter α, we solve the optimization problem and assess predictive performance on a validation set, as described above. The value of α that gives the best performance on the validation set is used to compute M^*, and its performance is measured on a test set, shown in Supplemental Fig. S8 [43]. Unlike the case for the real data, the optimal M^* is unable to model the shuffled data. This result, together with the cross-validation strategy described above, shows that the predictive accuracy of M^* stems from its ability to capture the geometry of the functional space of networks and not from its computational expressive power.

10. Correlation between dissimilarity matrices across the space of stimuli

The functional dissimilarity matrix between networks depends on the stimulus that is presented to the networks. It is not immediately clear that these dissimilarity matrices are related in a simple way. In particular, networks that respond very differently to one stimulus may have very similar responses to another stimulus. To investigate this, we calculate the Pearson correlation between all computed dissimilarity matrices for the different stimuli and the one presented in the main text (for which $\eta = 1.5$ and $\rho = 0$).

Supplemental Fig. S9 [43] shows that, over large parts of the space of stimuli, functional dissimilarity matrices are

highly correlated. This implies that the functional distances between networks are scaled on average by a multiplicative constant that is a function of the statistics of the stimulus. Up to this scaling behavior, the overall structure of functional networks space remains stable.

11. Relative importance of the sum of synaptic inputs and the sum of synaptic outputs for the feature-based models

The models based on the sum of synaptic inputs and the sum of synaptic outputs (IO-based models), described by Eq. (3) in the main text, are fitted by a linear regression that predicts D_{func} using a weighted sum of the squared differences between the sum of synaptic inputs and the sum of synaptic outputs of each of the neurons. To characterize the relative importance of the synaptic inputs and synaptic outputs in predicting functional dissimilarity of networks, we fit two additional models for each (η, ρ) stimulus: a model that relies only on the sum of synaptic inputs of each neuron and a model that relies only on the sum of synaptic outputs of each neuron. We compute the accuracy of each model, defined as Pearson's correlation with empirical D_{func}, denoting the accuracy of the input-based model as A_I and the accuracy of the output-based model as A_O. We then compare the accuracy of these two models by computing the ratio between correlation coefficients, A_I/A_O, as plotted in Supplemental Fig. S10 [43]. Interestingly, we find a transition from an output-dominated to an input-dominated regime as stimulus strength increased. We note that the model in the main text uses both the inputs and outputs as features and, therefore, outperforms both, by construction.

[1] E. Marder, *Variability, Compensation, and Modulation in Neurons and Circuits*, Proc. Natl. Acad. Sci. U.S.A. **108**, 15542 (2011).

[2] N. Barkai and S. Leibler, *Robustness in Simple Biochemical Networks*, Nature (London) **387**, 913 (1997).

[3] T. Y.-C. Tsai, Y. S. Choi, W. Ma, J. R. Pomerening, C. Tang, and J. E. Ferrell, *Robust, Tunable Biological Oscillations from Interlinked Positive and Negative Feedback Loops*, Science **321**, 126 (2008).

[4] R. Albert, H. Jeong, and A.-L. Barabási, *Error and Attack Tolerance of Complex Networks*, Nature (London) **406**, 378 (2000).

[5] S. Gu, F. Pasqualetti, M. Cieslak, Q. K. Telesford, A. B. Yu, A. E. Kahn, J. D. Medaglia, J. M. Vettel, M. B. Miller, S. T. Grafton, and D. S. Bassett, *Controllability of Structural Brain Networks*, Nat. Commun. **6**, 8414 (2015).

[6] J. J. Atick, *Could Information Theory Provide an Ecological Theory of Sensory Processing?*, Network: Comput. Neural Syst. **3**, 213 (1992).

[7] G. Tkačik, J. S. Prentice, V. Balasubramanian, and E. Schneidman, *Optimal Population Coding by Noisy Spiking Neurons*, Proc. Natl. Acad. Sci. U.S.A. **107**, 14419 (2010).

[8] T. Mora and W. Bialek, *Are Biological Systems Poised at Criticality?*, J. Stat. Phys. **144**, 268 (2011).

[9] M. Rubinov, O. Sporns, J.-P. Thivierge, and M. Breakspear, *Neurobiologically Realistic Determinants of Self-Organized Criticality in Networks of Spiking Neurons*, PLoS Comput. Biol. **7**, e1002038 (2011).

[10] E. Ganmor, R. Segev, and E. Schneidman, *Sparse Low-Order Interaction Network Underlies a Highly Correlated, and Learnable Neural Population Code*, Proc. Natl. Acad. Sci. U.S.A. **108**, 9679 (2011).

[11] P. Zurn and D. S. Bassett, *Network Architectures Supporting Learnability*, Phil. Trans. R. Soc. B **375**, 20190323 (2020).

[12] M. E. J. Newman, *The Structure and Function of Complex Networks*, SIAM Rev. **45**, 167 (2003).

[13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network Motifs: Simple Building Blocks of Complex Networks*, Science **298**, 824 (2002).

[14] A.-L. Barabási and R. Albert, *Emergence of Scaling in Random Networks*, Science **286**, 509 (1999).

[15] E. T. Bullmore, O. Sporns, and S. A. Solla, *Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems*, Nat. Rev. Neurosci. **10**, 186 (2009).

[16] B. A. Olshausen and D. J. Field, *Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?*, Vis. Res. **37**, 3311 (1997).

[17] S. Ganguli and H. Sompolinsky, *Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis*, Annu. Rev. Neurosci. **35**, 485 (2012).

[18] O. Maoz, G. Tkačik, M. S. Esteki, R. Kiani, and E. Schneidman, *Learning Probabilistic Neural Representations with Randomly Connected Circuits*, Proc. Natl. Acad. Sci. U.S.A. **117**, 25066 (2020).

[19] J. Soriano, M. R. Martinez, T. Tlusty, and E. Moses, *Development of Input Connections in Neural Cultures*, Proc. Natl. Acad. Sci. U.S.A. **105**, 13758 (2008).

[20] A. Ponce-Alvarez, A. Jouary, M. Privat, G. Deco, and G. Sumbre, *Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics*, Neuron **100**, 1446 (2018).

[21] C. I. Bargmann and E. Marder, *From the Connectome to Brain Function*, Nat. Methods **10**, 483 (2013).

[22] J. W. Lichtman, H. Pfister, and N. Shavit, *The Big Data Challenges of Connectomics*, Nat. Neurosci. **17**, 1448 (2014).

[23] X. Jiang, S. Shen, C. R. Cadwell, P. Berens, F. Sinz, A. S. Ecker, S. Patel, and A. S. Tolias, *Principles of Connectivity among Morphologically Defined Cell Types in Adult Neocortex*, Science **350**, aac9462 (2015).

[24] S. W. Oh *et al.*, *A Mesoscale Connectome of the Mouse Brain*, Nature (London) **508**, 207 (2014).

[25] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, *Connectomic Reconstruction of the Inner Plexiform Layer in the Mouse Retina*, Nature (London) **500**, 168 (2013).

[26] L. K. Scheffer, C. S. Xu, M. Januszewski, Z. Lu, S.-y. Takemura, K. J. Hayworth, G. B. Huang, K. Shinomiya, J. Maitlin-Shepard, S. Berg *et al.*, *A Connectome and Analysis of the Adult Drosophila Central Brain*, eLife **9**, e57443 (2020).

[27] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, *Structural Properties of the Caenorhabditis Elegans Neuronal Network*, PLoS Comput. Biol. **7**, e1001066 (2011).
[28] A. A. Prinz, D. Bucher, and E. Marder, *Similar Network Activity from Disparate Circuit Parameters*, Nat. Neurosci. **7**, 1345 (2004).
[29] J. J. Jun *et al.*, *Fully Integrated Silicon Probes for High-Density Recording of Neural Activity*, Nature (London) **551**, 232 (2017).
[30] E. A. Susaki, K. Tainaka, D. Perrin, F. Kishino, T. Tawara, T. M. Watanabe, C. Yokoyama, H. Onoe, M. Eguchi, S. Yamaguchi, T. Abe, H. Kiyonari, Y. Shimizu, A. Miyawaki, H. Yokota, and H. R. Ueda, *Whole-Brain Imaging with Single-Cell Resolution Using Chemical Cocktails and Computational Analysis*, Cell **157**, 726 (2014).
[31] L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel, *Functional Organization of Excitatory Synaptic Strength in Primary Visual Cortex*, Nature (London) **518**, 399 (2015).
[32] F. Scala, D. Kobak, S. Shan, Y. Bernaerts, S. Laturnus, C. R. Cadwell, L. Hartmanis, E. Froudarakis, J. R. Castro, Z. H. Tan, S. Papadopoulos, S. S. Patel, R. Sandberg, P. Berens, X. Jiang, and A. S. Tolias, *Layer 4 of Mouse Neocortex Differs in Cell Types and Circuit Organization between Sensory Areas*, Nat. Commun. **10**, 4174 (2019).
[33] A. A. Wanner and R. W. Friedrich, *Whitening of Odor Representations by the Wiring Diagram of the Olfactory Bulb*, Nat. Neurosci. **23**, 433 (2020).
[34] A. Litwin-Kumar and S. C. Turaga, *Constraining Computational Models Using Electron Microscopy Wiring Diagrams*, Curr. Opin. Neurobiol. **58**, 94 (2019).
[35] A. Sizemore, C. Giusti, R. F. Betzel, and D. S. Bassett, *Closures and Cavities in the Human Connectome*, arXiv:1608.03520.
[36] M. Timme and J. Casadiego, *Revealing Networks from Dynamics: An Introduction*, J. Phys. A **47**, 343001 (2014).
[37] M. Tikhonov and W. Bialek, *Complexity of Generic Biochemical Circuits: Topology versus Strength of Interactions*, Phys. Biol. **13**, 066012 (2016).
[38] K. Morrison and C. Curto, in *Algebraic and Combinatorial Computational Biology* (Elsevier, New York, 2019), pp. 241–277.
[39] F. Mastrogiuseppe and S. Ostojic, *Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks*, Neuron **99**, 609 (2018).
[40] D. V. Raman, A. P. Rotondo, and T. O'Leary, *Fundamental Bounds on Learning Performance in Neural Circuits*, Proc. Natl. Acad. Sci. U.S.A. **116**, 10537 (2019).
[41] N. Brunel, *Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons*, J. Comput. Neurosci. **8**, 183 (2000).
[42] G. Buzsáki and K. Mizuseki, *The Log-Dynamic Brain: How Skewed Distributions Affect Network Operations*, Nat. Rev. Neurosci. **15**, 264 (2014).
[43] See Supplemental Material containing supplemental figures and explanations at http://link.aps.org/supplemental/10.1103/PhysRevX.12.021051.
[44] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, *Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population*, Nature (London) **440**, 1007 (2006).
[45] B. Kulis, *Metric Learning: A Survey*, Found. Trends Mach. Learn. **5**, 287 (2013).
[46] M. Perrot, A. Habrard, D. Muselet, and M. Sebban, *Modeling Perceptual Color Differences by Local Metric Learning*, in *European Conference on Computer Vision* (Springer, 2014), pp. 96–111, https://doi.org/10.1007/978-3-319-10602-1_7.
[47] J. Freeman and E. P. Simoncelli, *Metamers of the Ventral Stream*, Nat. Neurosci. **14**, 1195 (2011).
[48] E. Ganmor, R. Segev, and E. Schneidman, *A Thesaurus for a Neural Population Code*, eLife **4**, e06134 (2015).
[49] R. Chaudhuri, B. Gerçek, B. Pandey, A. Peyrache, and I. Fiete, *The Intrinsic Attractor Manifold and Population Dynamics of a Canonical Cognitive Circuit across Waking and Sleep*, Nat. Neurosci. **22**, 1512 (2019).
[50] A. E. Sizemore, C. Giusti, A. Kahn, J. M. Vettel, R. F. Betzel, and D. S. Bassett, *Cliques and Cavities in the Human Connectome*, J. Comput. Neurosci. **44**, 115 (2018).
[51] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds* (Princeton University Press, Princeton, NJ, 2007).
[52] J. Townsend, N. Koep, and S. Weichwald, *Pymaopt: A Python Toolbox for Optimization on Manifolds Using Automatic Differentiation*, J. Mach. Learn. Res. **17**, 1 (2016).
[53] H. J. Sommers, A. Crisanti, H. Sompolinsky, and Y. Stein, *Spectrum of Large Random Asymmetric Matrices*, Phys. Rev. Lett. **60**, 1895 (1988).
[54] C. Van Vreeswijk and H. Sompolinsky, *Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity*, Science **274**, 1724 (1996).
[55] C. v. Vreeswijk and H. Sompolinsky, *Chaotic Balanced State in a Model of Cortical Circuits*, Neural Comput. **10**, 1321 (1998).
[56] D. J. Amit and N. Brunel, *Model of Global Spontaneous Activity and Local Structured Activity during Delay Periods in the Cerebral Cortex*, Cereb. Cortex **7**, 237 (1997).
[57] C. Curto and K. Morrison, *Relating Network Connectivity to Dynamics: Opportunities and Challenges for Theoretical Neuroscience*, Curr. Opin. Neurobiol. **58**, 11 (2019).
[58] E. van Nimwegen, J. P. Crutchfield, and M. Huynen, *Neutral Evolution of Mutational Robustness*, Proc. Natl. Acad. Sci. U.S.A. **96**, 9716 (1999).
[59] G. G. Turrigiano and S. B. Nelson, *Homeostatic Plasticity in the Developing Nervous System*, Nat. Rev. Neurosci. **5**, 97 (2004).
[60] S. Royer and D. Paré, *Conservation of Total Synaptic Weight through Balanced Synaptic Depression and Potentiation*, Nature (London) **422**, 518 (2003).
[61] C. Linssen *et al.*, NEST 2.16.0, https://doi.org/10.5281/ZENODO.1400175.
[62] A. Kuhn, A. Aertsen, and S. Rotter, *Higher-Order Statistics of Input Ensembles and the Response of Simple Model Neurons*, Neural Comput. **15**, 67 (2003).
[63] O. Maoz and E. Schneidman, *maxent_toolbox: Maximum Entropy Toolbox for Matlab, Version 1.0.2* (2017), https://doi.org/10.5281/zenodo.191625.

[64] J. C. Nash, *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation* (Hilger, Bristol, 1979).

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine Learning in Python*, J. Mach. Learn. Res. **12**, 2825 (2011).

[66] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms* (Cambridge University Press, Cambridge, England, 2002).

[67] J. Lin, *Divergence Measures Based on the Shannon Entropy*, IEEE Trans. Inf. Theory **37**, 145 (1991).

[68] S. Rotter and M. Diesmann, *Exact Digital Simulation of Time-Invariant Linear Systems with Applications to Neuronal Modeling*, Biol. Cybern. **81**, 381 (1999).

[69] A. N. Burkitt, *A Review of the Integrate-and-Fire Neuron Model: I. Homogeneous Synaptic Input*, Biol. Cybern. **95**, 1 (2006).

[70] M. Stimberg, R. Brette, and D. F. Goodman, *Brian 2, an Intuitive and Efficient Neural Simulator*, eLife **8**, e47314 (2019).

[71] M. Stimberg, D. F. M. Goodman, R. Brette, and M. D. Pittà, in *Computational Glioscience*, edited by M. De Pittà and H. Berry (Springer International, New York, 2019), pp. 471–505.

[72] P. Virtanen *et al.*, *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nat. Methods **17**, 261 (2020).

[73] A. A. Hagberg, D. A. Schult, and P. J. Swart, *Exploring Network Structure, Dynamics, and Function Using NetworkX*, Los Alamos National Lab.(LANL), Los Alamos, 2008, pp. 11–15, https://www.osti.gov/biblio/960616.

[74] P. Wills and F. G. Meyer, *Metrics for Graph Comparison: A Practitioner's Guide*, PLoS One **15**, e0228728 (2020).

SUPPLEMENTARY FIGURES

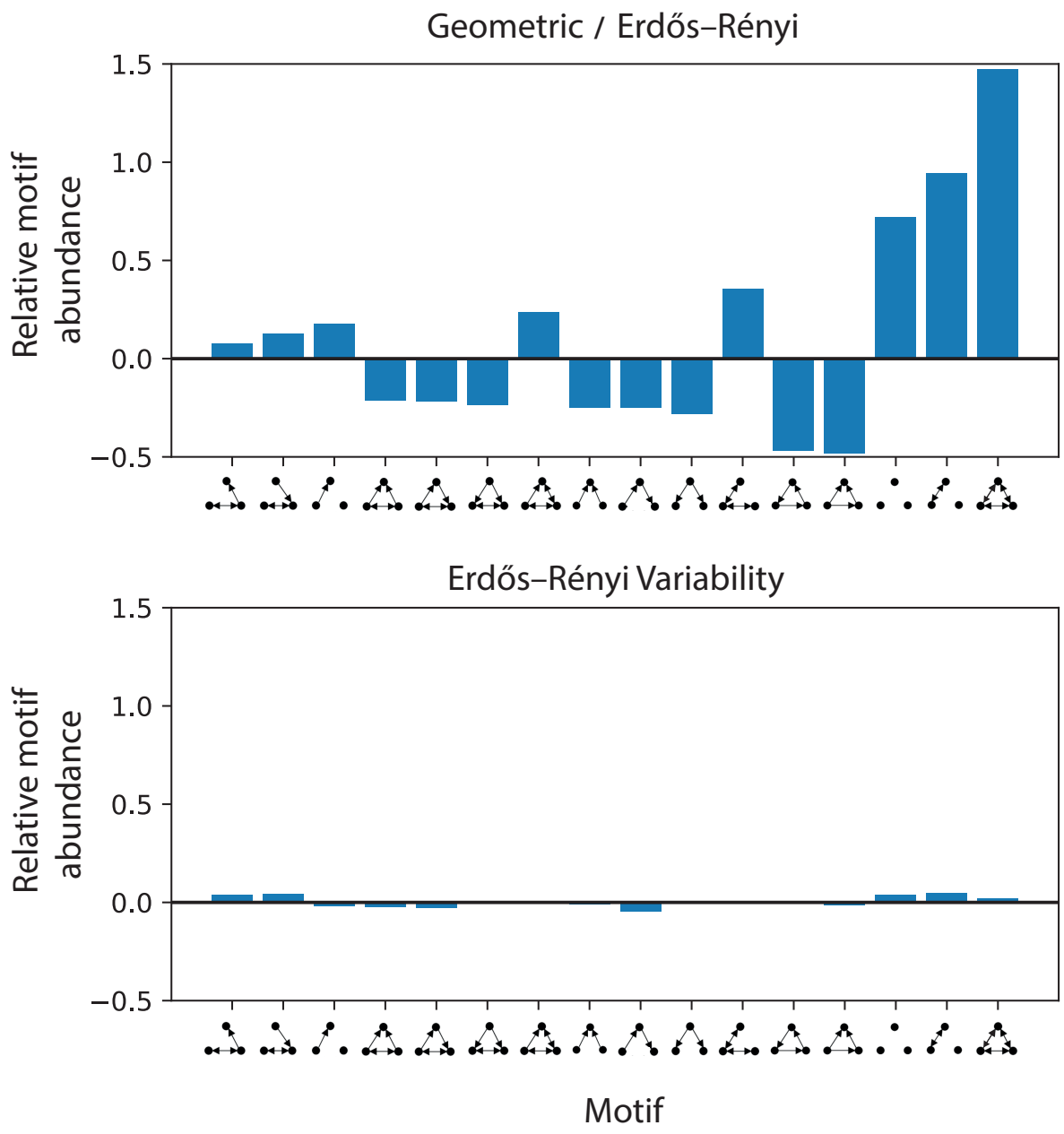

Figure S1. Comparing motifs distributions for Erdős–Rényi and Geometric ensembles. We computed the mean number of occurrences of each 3-nodes motif in 50 random ER networks and 50 random networks from the geometric ensemble, and computed the relative motif ensemble $\frac{\langle \text{Counts} \rangle_{\text{Geo}} - \langle \text{Counts} \rangle_{\text{ER}}}{\langle \text{Counts} \rangle_{\text{ER}}}$ (upper plot). We find a significant enrichment of reciprocal connections, 3-nodes cliques and empty triples in the geometric ensemble. We compared the original 50 ER networks to 50 additional networks from the same ensemble and computed $\frac{\langle \text{Counts} \rangle_{\hat{\text{ER}}} - \langle \text{Counts} \rangle_{\text{ER}}}{\langle \text{Counts} \rangle_{\text{ER}}}$ (bottom plot), suggesting the variability between ensembles is significantly higher than sampling noise.

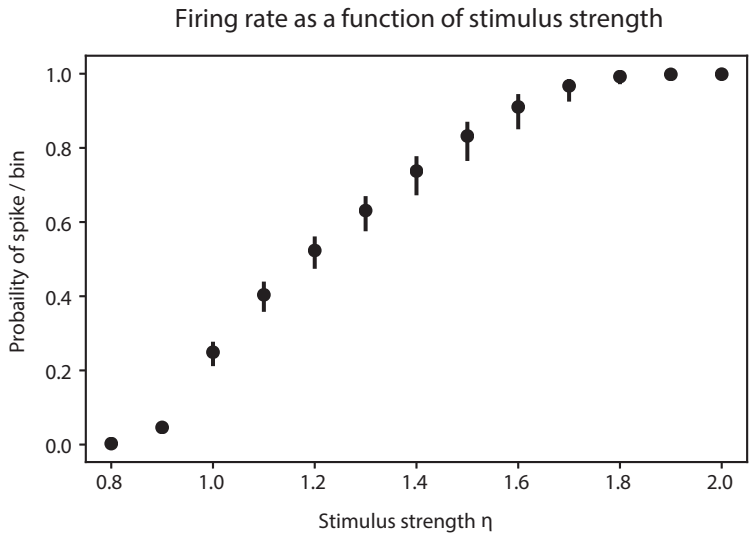

Figure S2. Firing rates as a function of stimulus strength. In the main text, results for the N=15 ensemble are for $\eta = 1.5$ unless stated otherwise.

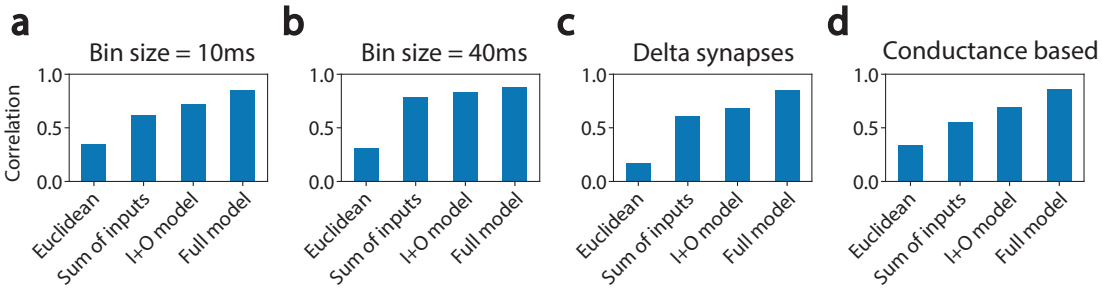

Figure S3. Comparison of the performance of models under different neuron model, synaptic dynamics, and time bins. Our model outperformed the Euclidean structural distance for different choice for temporal bin size (a,b), synaptic dynamics (c), and single neuron models (d).

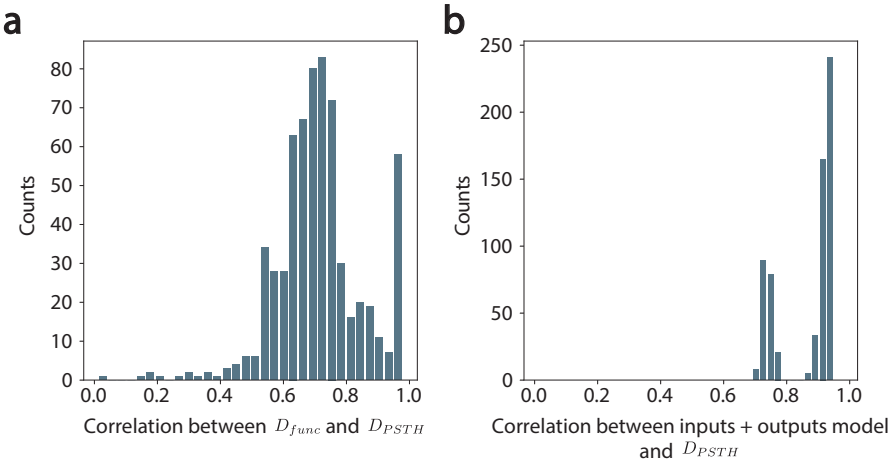

Figure S4. Feature based model accurately predicts other functional dissimilarity measures. (a) For different values of stimulus strength and correlation, the D_{func} measure and D_{PSTH} are highly correlated. **(b)** Similar to the results presented in the main text, a feature-based model that relies on the sum of synaptic inputs and sum of synaptic outputs of each neuron accurately predicts D_{PSTH}.

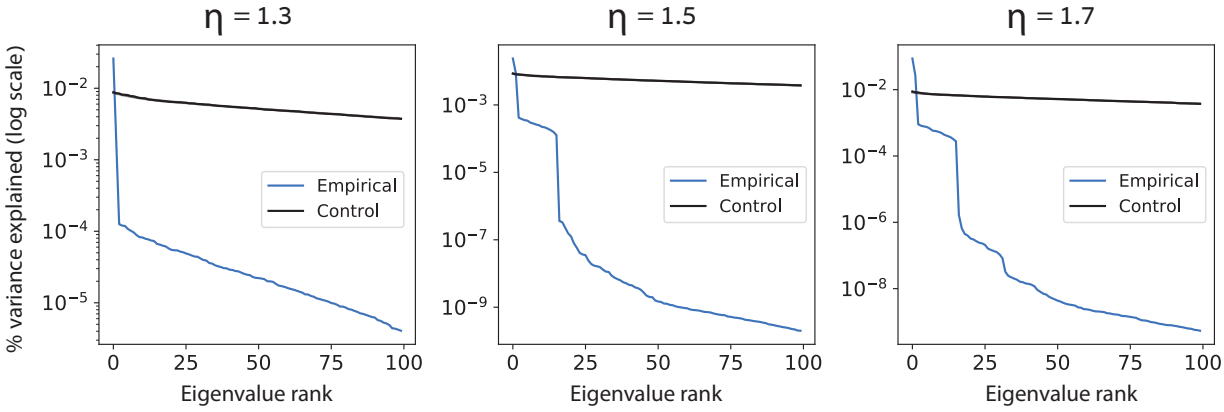

Figure S5. Low dimensional structure of the matrix of distances between networks. For 3 different stimuli, the spectrum of the empirical functional dissimilarity matrix decays faster than shuffled control.

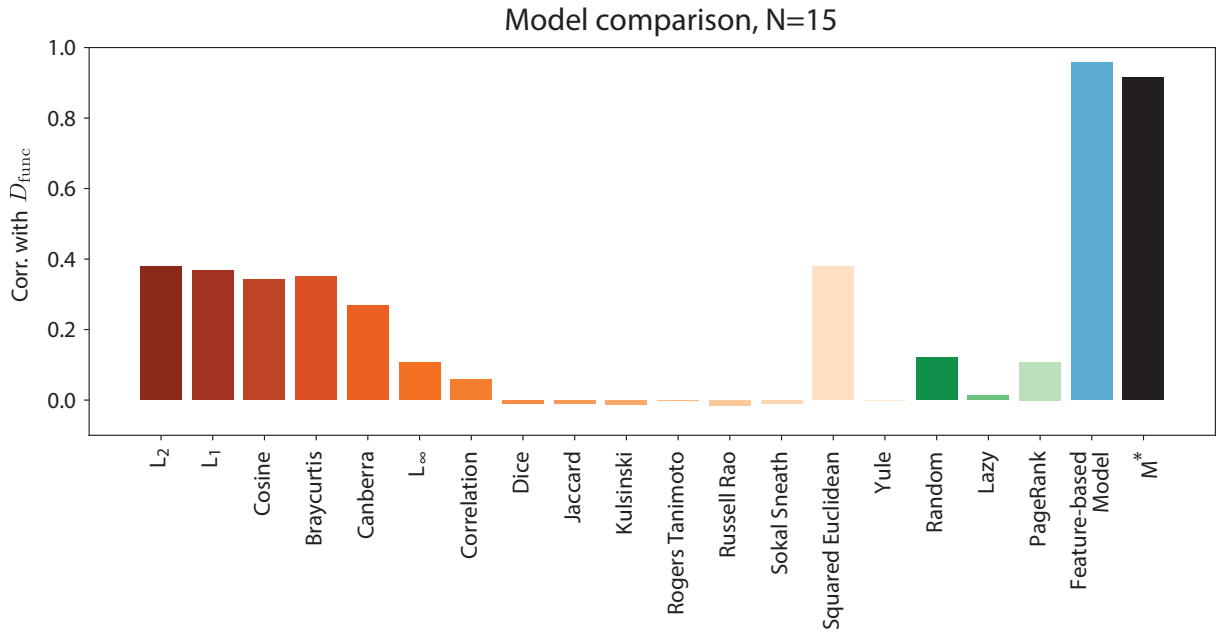

Figure S6. A wide range of structural metrics fail to predict functional dissimilarity. We compared a wide variety of structural metrics. The model based on M^* (black) significantly outperformed all metrics we have considered. Results for the $N = 15$ ensemble and $\eta = 1.5$.

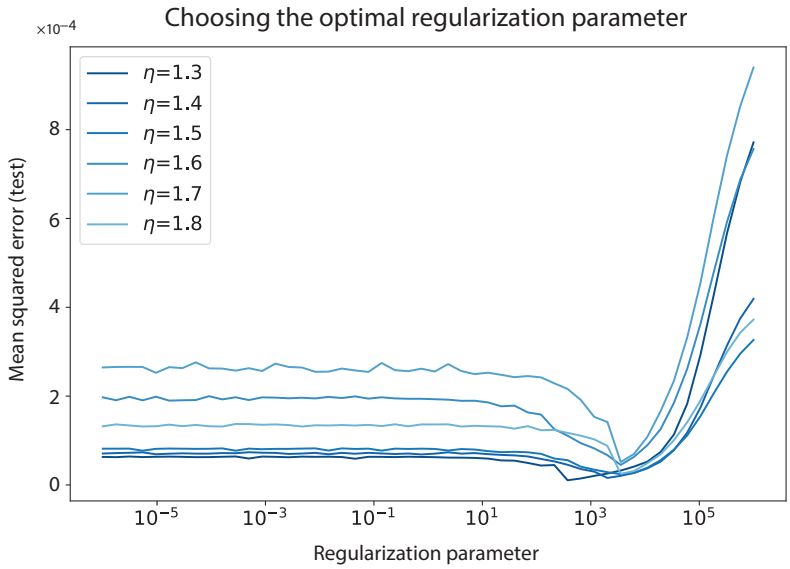

Figure S7. Choosing the optimal value of the regularization parameter α using train-test splitting. For the ensemble of networks with $N = 15$, M^* was found with regularized optimization for 50 different values of α.

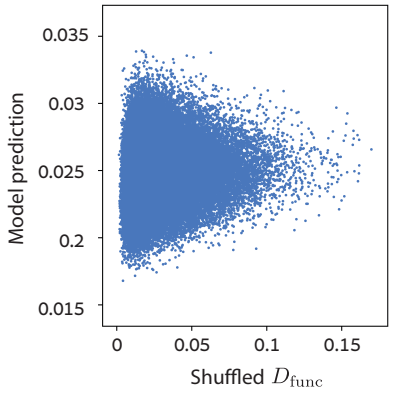

Figure S8. Fitting M^* to shuffled data. The optimal Mahalanobis M^* shows no structure.

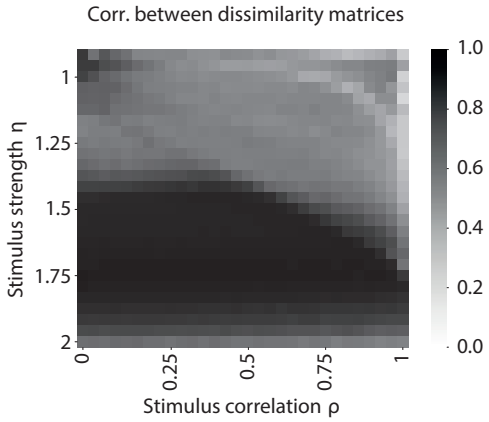

Figure S9. Correlation between dissimilarity matrices across the space of stimuli. Mean correlation across stimulus space is 0.8 with a standard deviation of 0.16 (median 0.81).

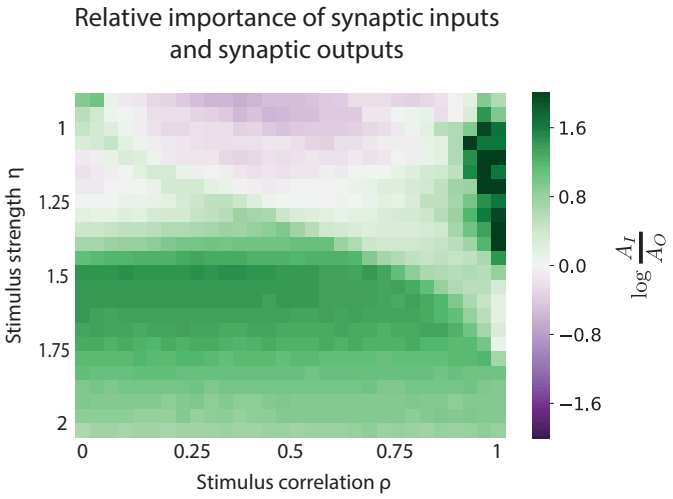

Figure S10. Relative importance of the sum of synaptic inputs and the sum of synaptic outputs for the feature based models varies across stimulus space. As stimuli strength increases, similarity of synaptic inputs dominates similarity of synaptic outputs.

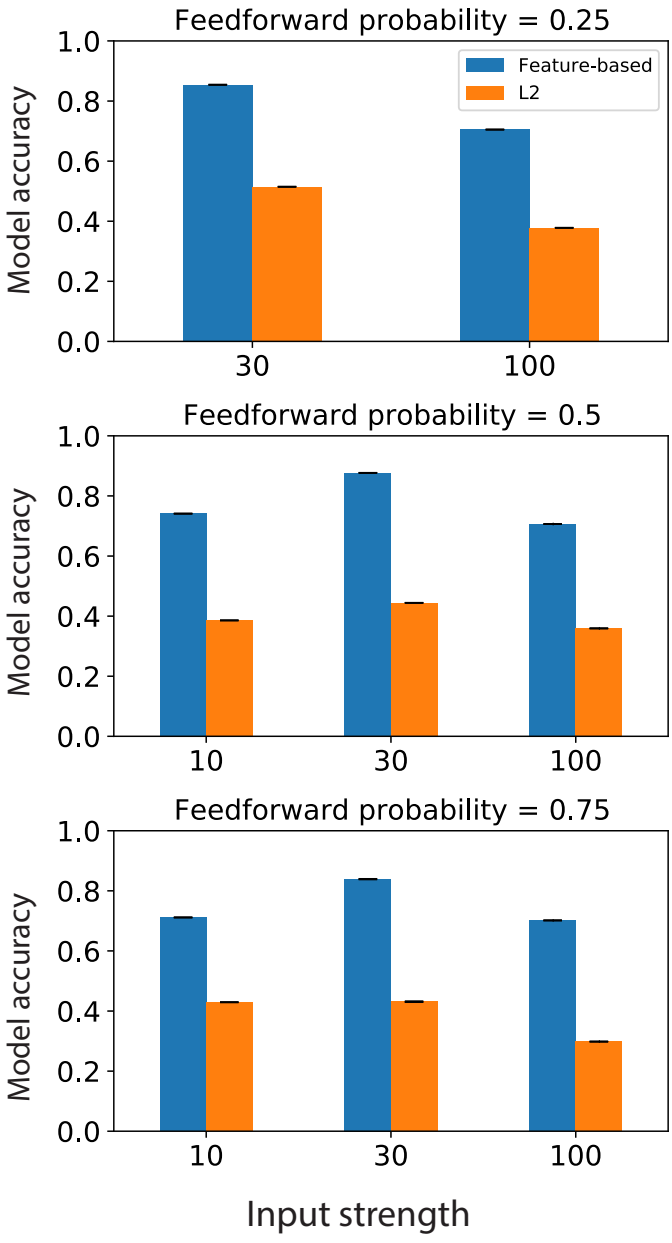

Figure S11. Model performance for different parameters of the natural stimulus. Feature based model remains accurate for different values of feed-forward connectivity and input strength. Blue - correlation between I+O model and D_{func}, orange - correlation between Euclidean distances and D_{func}. Error bars represent standard deviations over 10 initial conditions.

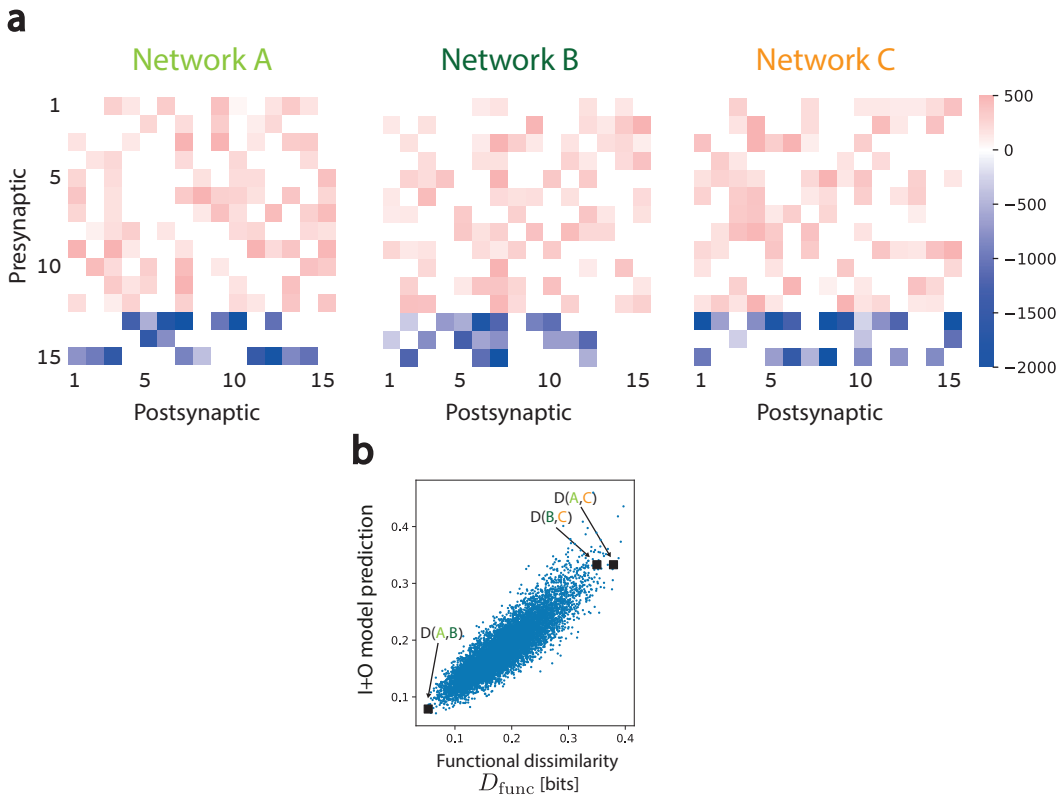

Figure S12. Feature-based model predicts functional similarity under natural stimuli for ensemble of fixed neuron types. Networks of N=15 neurons were sampled from the same distributions of weights and topologies as in the main text, but with neuronal types fixed across all networks in the ensemble (neurons 1-12 were excitatory, neurons 13-15 were inhibitory). The responses of the networks in this ensemble were simulated and analyzed as in Figure 6. **(a)** Three examples of network with fixed neuronal types. **(b)** Feature-based model remains highly predictive of functional similarity for the fixed types ensemble.