A Methods

A.1 Finding the stationary distribution $\pi_{W,\theta}(\mathbf{x}|\mathbf{s})$

Let M be the transition matrix defined by the network W and stimulus $\mathbf{s}$, as in eq. (1). From the theory of Markov chains, we know that since M is a regular transition matrix, then there exists a single vector w such that $w \cdot M = w$, up to multiplication by a constant factor (see Grinstead and Snell [37], Theorem 11.8a). Specifically, there exists a unique vector π such that $\pi \cdot M = \pi$ and the sum of its component is 1. Since π is in the left nullspace of the matrix $M - I$, and since this nullspace has dimension 1, we can find the unique stationary distribution by computing the nullspace (using SVD) and then normalizing a vector in the nullspace.

A.2 Ordering the rows and columns of the dissimilarity matrix

We sorted the rows (and columns) of the Daleian and of the non-Daleian parts of Figure 1d to reflect any structural organization. We used hierarchical clustering within each sub-matrix, namely the Dale vs. Dale part and the non-Dale vs. non-Dale part, and reordered the matrices using optimal leaf ordering, which minimizes the functional dissimilarity $D_{JS}\left(\pi_{W_k,\theta}\left(\mathbf{x}|\mathbf{s}\right)||\pi_{W_{k+1},\theta}\left(\mathbf{x}|\mathbf{s}\right)\right)$ between networks in successive matrix indices [38]. The full matrix was then ordered using the optimal ordering of the sub-matrices.

A.3 Optimization details

All optimizations presented in the main text were done using Adam optimizer [39], with a learning rate of 10^{-2} and 2500 optimization steps, except for optimizations of networks with more disperse initializations (see B.12) for which we used a learning rate of 10^{-1}. For the computation of the mutual information presented in eq. (4), optimization was done using batches of 100 different stimuli in each optimization step. All analysis was done using the `JAX` Python library [40].

A.4 Dissimilarity between probability distributions

The Kullback-Leibler divergence between probability distributions P, Q is defined as $D_{KL}\left(P||Q\right) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$, which measures the distinguishability of distributions in bits, and has multiple information theory and statistics motivations and interpretations [41]. The Jensen-Shannon divergence [22] is a symmetric and bounded measure, based on the Kullback-Leibler divergence, defined as $D_{JS}\left(P||Q\right) = \frac{1}{2}D_{KL}\left(P||M\right) + \frac{1}{2}D_{KL}\left(Q||M\right)$ where $M = \frac{P+Q}{2}$. This is a measure of dissimilarity between probability distributions, which is 0 bits for identical distributions, and 1 bit for non-overlapping distributions.

B Supplementary information

B.1 Number of Daleian networks of N neurons

To find the number of signed networks that abide by Dale's principle, we note that for each of the neurons in the network there are 2^{N-1} possible patterns of outgoing synapses to all the other neurons. Since each neuron is either excitatory or inhibitory, the number of possible combinations of synapses going out from a single neuron in a Daleian network is $2 \cdot 2^{N-1} - 1$, (the -1 term is to avoid double-counting of the zero-outdegree case). Therefore, there are $(2 \cdot 2^{N-1} - 1)^N$ Daleian networks of N neurons. We can further generalize and count the number of Daleian and non-Daleian whose synaptic weights (in absolute values) come from a finite set of possible values $\{0, w_1, \ldots, w_k\}$. By the same combinatorial argument, there are $(2k+1)^{N(N-1)}$ such non-Daleian networks, and $(2 \cdot (k+1)^{N-1} - 1)^N$ Daleian networks. This suggests that the ratio between the cardinality of the two sets of networks increase as we allow synapses to take more values.

B.2 Network responses are shaped by both the stimulus and the recurrent activity in the network

To verify that we used stimulus values and synaptic connections in the network, such that the activity of the network is not dictated just by one of them, we computed the dissimilarity between the stationary distribution of a given network W responding to a random stimulus $\mathbf{s}$, and the stationary distributions that are the result of either setting all weights to zero, or setting the stimulus to zero. This amounts to computing $D_{JS}\left(\pi_{W,\theta}\left(\mathbf{x}|\mathbf{s}\right)||\pi_{[0],\theta}\left(\mathbf{x}|\mathbf{s}\right)\right)$ and $D_{JS}\left(\pi_{W,\theta}\left(\mathbf{x}|\mathbf{s}\right)||\pi_{W,\theta}\left(\mathbf{x}|\mathbf{0}\right)\right)$, respectively. We computed both quantities for 500 different networks and 500 different stimuli, and found that the average dissimilarity due to removing synaptic connections within the network was $\left\langle D_{JS}\left(\pi_{W,\theta}\left(\mathbf{x}|\mathbf{s}\right)||\pi_{[0],\theta}\left(\mathbf{x}|\mathbf{s}\right)\right)\right\rangle_{W,s} \sim 0.8\cdot 10^{-1}$, and the average dissimilarity due to removing the stimulus was $\left\langle D_{JS}\left(\pi_{W,\theta}\left(\mathbf{x}|\mathbf{s}\right)||\pi_{W,\theta}\left(\mathbf{x}|\mathbf{0}\right)\right)\right\rangle_{W,s} \sim 2.6\cdot 10^{-1}$ (Fig. S1). Since both effects have similar scale (compared to the distances considered in the main text), we concluded that functional dissimilarity is not driven solely by the stimulus or the structure of the networks, but by a combination of both.

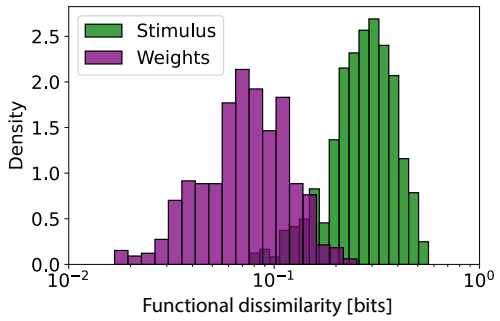

Figure S1: Functional dissimilarity between the stationary distribution $\pi_{W,\theta}\left(\mathbf{x}|\mathbf{s}\right)$ and the distributions resulting from zeroing the stimulus (green histogram) or the synaptic weights (purple histogram), for 500 networks and stimuli.

B.3 Convergence of the dynamics to the stationary distribution

To measure the rate of convergence of the Markov chains considered in the main text to their stationary distributions, we randomly sampled a network W and stimulus $\mathbf{s}$, computed the corresponding transition matrix M, as well as the stationary distribution $\pi_{W,\theta}\left(\mathbf{x}|\mathbf{s}\right)$. We then sampled a random initial distribution p_{init} from a uniform distribution over the space of probability distributions of binary activity patterns (i.e., Dirichlet distribution with $\alpha = 1$), and computed the Jensen-Shannon divergence between $\pi_{W,\theta}(\mathbf{x}|\mathbf{s})$ and the distribution after running the dynamics for k steps, starting from the initial distribution p_{init}:

$$D_{JS}\left[p_{\text{init}}\cdot M^{k}||\pi_{W,\theta}\right] \tag{5}$$

We repeated this for $k = 1\ldots 10$, using 100 different initial distributions p_{init}, and 30 different pairs of network W and stimulus $\mathbf{s}$ (Fig. S2). We found that ~ 5 transitions were enough to achieve convergence, measured as $D_{JS} < 10^{-6}$, corresponding to less than 1% of synaptic noise (see main text). Again, if we assume a time bin of $20\,ms$, the convergence to the stationary distribution happens within biologically-relevant timescales of $\sim 100\,ms$.

B.4 Distance profiles for all networks in the ensemble for a single stimulus

Figure 1g shows the distributions of functional dissimilarity values from a random non-Daleian network to all other Daleian and non-Daleian networks in the ensemble. To verify that the result is not sensitive to a particular choice of a non-Daleian network, we inspected the same distributions for all non-Daleian networks in the ensemble, and found similar results – i.e., the “left" tail of the distances’ distribution, corresponding to the closest network pairs in the ensemble, shows similar distance for Daleian and non-Daleian networks (Fig. S3).

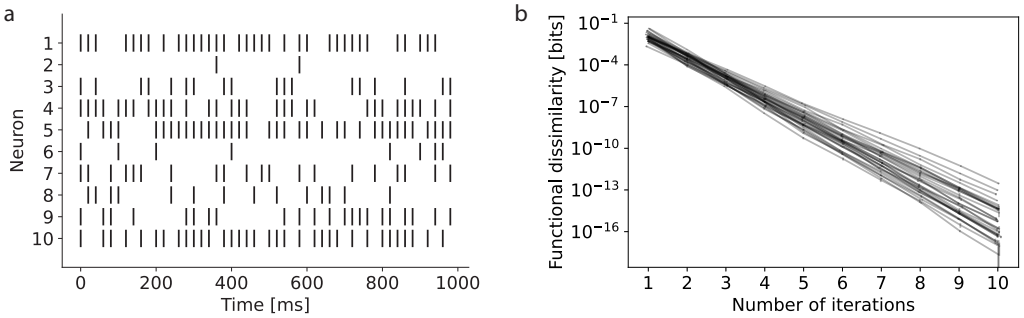

Figure S2: (**a**) An example of running the Markovian dynamics, for a random matrix W and stimulus **s**, for 50 steps (corresponding to 1 second) starting from a random binary state. (**b**) The functional dissimilarity $D_{JS}(p_{\text{init}} \cdot M^k || \pi_{W,\theta})$ for 30 different networks W and stimuli **s**, for different values of k. Error bars correspond to standard deviations over 100 different initial distributions p_{init}.

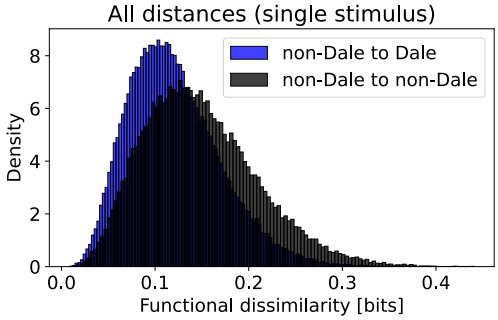

Figure S3: The distributions of the distances from all non-Daleian networks to all Daleian (blue) and non-Daleian (black) networks in the ensemble, for a single stimulus.

B.5 Average distance profiles for all networks in the ensemble.

The ability to find a close Daleian “neighbor" for a random non-Daleian network might depend on the specific stimulus that networks were presented with. To verify that this is not the case, we computed the mean functional dissimilarity matrix (averaged over 30 different stimuli), and found that, again, the closest network to a randomly chosen non-Daleian network is Daleian with a probability of $\sim 50\%$. The distributions of functional dissimilarity values were similar to the distributions computed for a single stimulus (Fig. S4).

B.6 Dissimilarity matrices for different stimuli are highly correlated

To measure the correlation between the dissimilarity values of pairs of networks across different stimuli, we computed the functional dissimilarity matrices between 300 Daleian networks and 300 non-Daleian networks, responding to 30 different stimuli. For each stimulus, we have computed the 600×600 functional dissimilarity matrix between all networks from the two ensembles. We then computed the Pearson correlation between all $\binom{30}{2}$ pairs of functional dissimilarity matrices (matrices were “flattened" as vectors for the computation of the correlation values), and found that the average correlation was 0.74 (Fig. S5).

B.7 Structural dissimilarity between a non-Daleian network and its Daleian approximation

The Daleian networks that learned to approximate the response of a non-Daleian network W^{nD} to a random stimulus **s** were characterized by very different connectivity structures compared to the original non-Daleian network. Figure S6 shows an example of one non-Daleian network, its Daleian approximation, and the corresponding learning curve (D_{JS} value at the end of optimization was

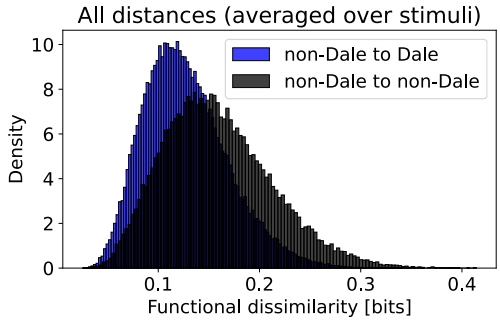

Figure S4: The distributions of the distances from all non-Daleian networks to all Daleian (blue) and non-Daleian (black) networks in the ensemble, averaged over 30 different stimuli. We note the very high similarity to the case of a single stimulus shown in Fig. S3

.

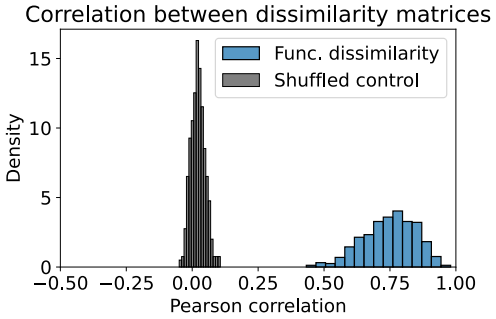

Figure S5: Correlation between functional dissimilarity matrices (blue histogram), computed for 30 different stimuli. The black histogram shows the correlation values between matrices whose rows and columns were shuffled (different shuffling for each matrix).

$2 \cdot 10^{-4}$). The correlation between the synaptic weights of the non-Daleian network and its Daleian approximation was 0.19, with a p-value of 0.19 (i.e., not statistically significant).

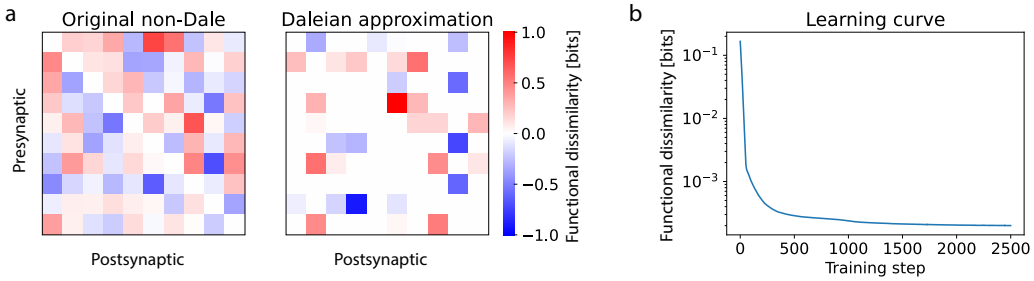

Figure S6: (**a**) A representative example of the connectivity matrix of the original non-Daleian network (left) and its Daleian approximation (right). Same color-bar for both networks. (**b**) The functional dissimilarity between the original non-Daleian network and its Daleian approximation during the learning process.

B.8 Effect of learning on synaptic weights distribution

As described in the main text (see Section 2.2), the distribution of synaptic weights prior to learning was normal for non-Daleian networks, and half-normal with the signs determined by the type of

the pre-synaptic neuron for Daleian networks. As the Daleian networks learned (using synaptic learning) to approximate the response distribution of a random non-Daleian network, the distribution of their synaptic weights became increasingly bi-modal, with some synapses becoming significantly stronger (in magnitude), and the rest of the synapses becoming practically indistinguishable from 0 (synaptic magnitude $< 10^{-6}$; Fig. S7). Interestingly, this "sparsification" of networks, as well as the deviance from normality towards a more skewed distribution, are known characteristics of real neuronal networks [30].

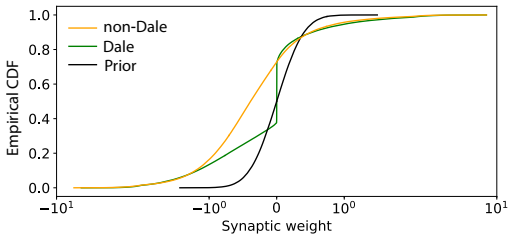

Figure S7: Empirical cumulative distribution function of synaptic weights before learning (black), and after learning for Daleian (green) and non-Daleian (orange) networks.

B.9 Measuring D_{func} for at different levels of synaptic noise

To measure the effects of synaptic noise on functional dissimilarity, and to give a more biological interpretation for the dissimilarity values, we have computed, for a given network W and stimulus $\mathbf{s}$, the functional dissimilarity between $\pi_{W,\theta}(\mathbf{x}|\mathbf{s})$ and $\pi_{W_\epsilon,\theta}(\mathbf{x}|\mathbf{s})$, where W_ϵ is a matrix in which every synapse W_{ij} is multiplied by either $1-\epsilon$ or $1+\epsilon$ at random. Other synaptic noise mechanisms (i.e., additive and not multiplicative noise, randomly distributed noise with increasing variance) were considered and gave similar results.

B.10 Computing sensitivity to perturbations of synaptic weights for large networks.

To compute the sensitivity of large firing-rate networks (eq. 3) to perturbations of synaptic weights, we defined the following functional dissimilarity measure, using the squared Frobenius norm between two stimulus-to-steady-state mappings:

$$g_W(\Delta W) = \|(I - W^T)^{-1} - (I - (W + \Delta W)^T)^{-1}\|^2 \tag{6}$$

Unlike the case for the Markovian model, for the firing-rate model the mapping is independent of a particular stimulus, and the comparison of different networks is therefore not stimulus-specific. We then computed the trace of the Hessian of g_W at $\Delta W = 0$ for 100 Daleian networks and 100 non-Daleian networks, and found that the traces of Daleian networks are orders-of-magnitude smaller, indicating a significant robustness to weight perturbations of their input-output function (Fig. S8). Since the steady-state solution is a linear function of the stimulus and individual thresholds, the second derivative with respect to thresholds is always zero, and analyzing the robustness of networks to perturbations of $\boldsymbol{\theta}$ in a similar manner is not informative for this model.

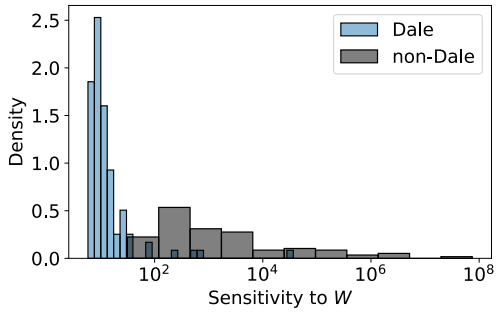

Figure S8: Traces of the Hessian of g_W evaluated at $\Delta W = 0$, for 100 Daleian and 100 non-Daleian networks.

B.11 Mutual information using other stimulus distributions.

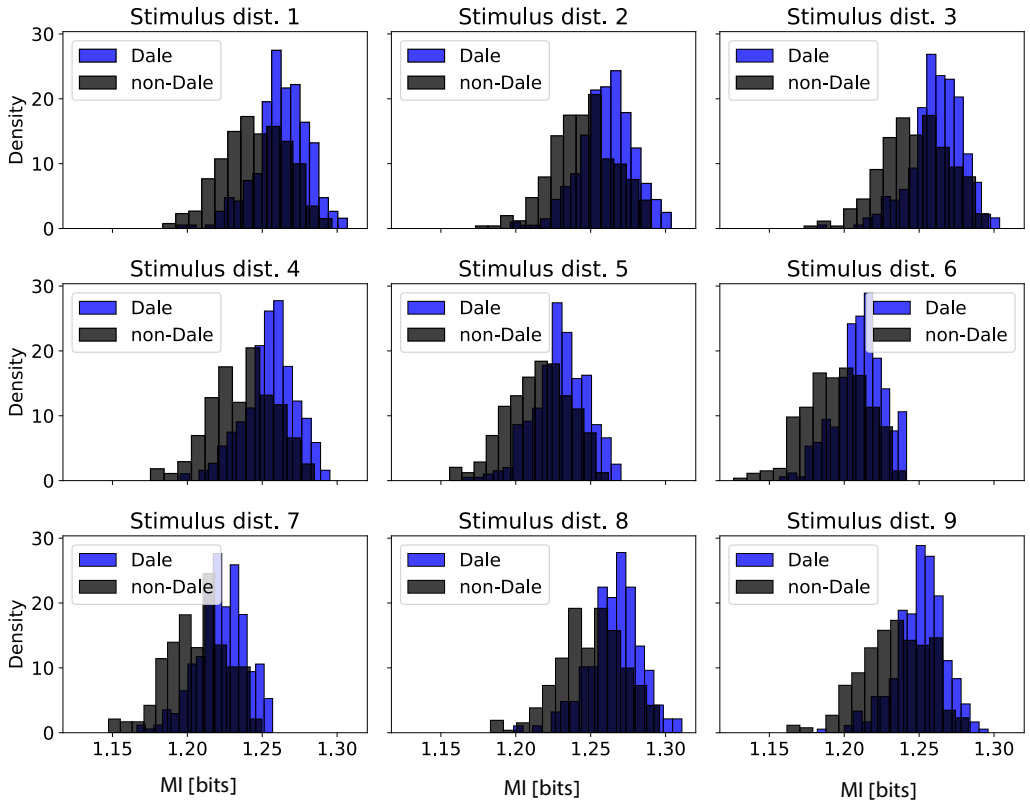

Figure S9: Distribution of mutual information scores of 500 non-Daleian (black) and 500 Daleian networks (blue), computed for a set of 500 random stimuli with different stimuli distributions $P(s)$. Each figure corresponds to one stimulus distribution. All distributions were sampled from a uniform distribution over the space of probability distributions of 500 stimuli – i.e., a Dirichlet distribution with $\alpha = 1$.

B.12 Distributions of functional similarity and spiking properties at different scales of synaptic weights distributions.

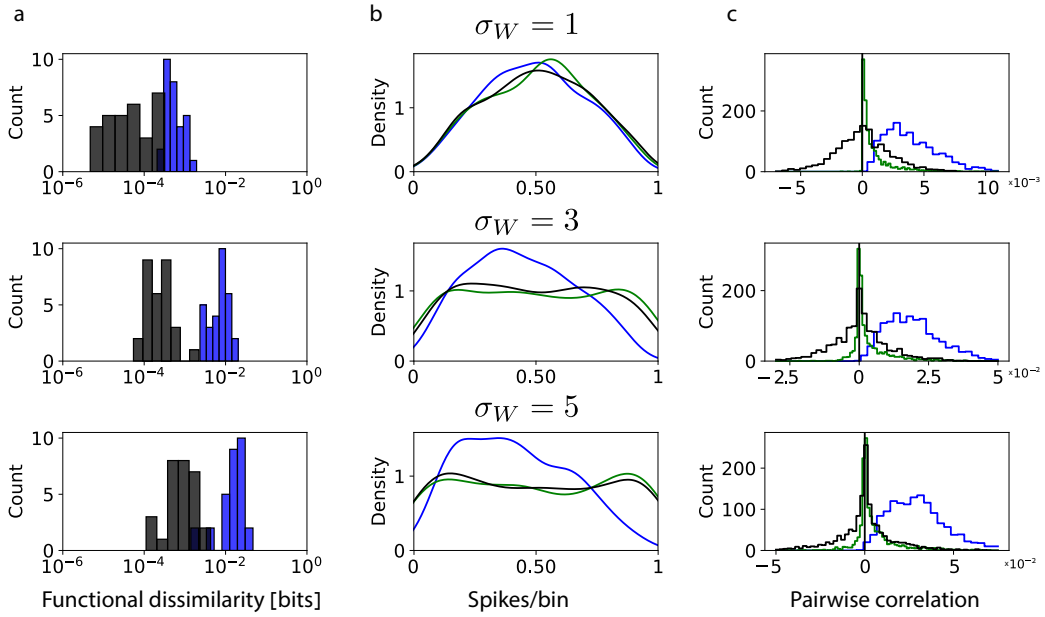

Figure S10: **(a)** Functional dissimilarity between the Daleian (blue) and non-Daleian (black) approximations of the response distributions of 30 random non-Daleian networks responding to 30 random stimuli. Synaptic weights of the target networks and optimization initial networks were sampled from a normal distribution centered around 0, and standard deviations $\sigma_W = 1, 3, 5$. **(b)** Firing rates distributions (smoothed using a Gaussian kernel) of all 300 neurons of 30 random Daleian networks (blue), non-Daleian networks (black) and 30 Daleian approximations (green) responding to randomly sampled stimuli. Same variance of synaptic weights as (a). **(c)** Pairwise correlations distributions of all 1350 pairs of neuron of 30 random Daleian networks (blue), non-Daleian networks (black) and 30 Daleian approximations (green) responding to randomly sampled stimuli. Same variance of synaptic weights as (a).

B.13 Code availability

The code used for performing all the analysis presented in the paper is in an accompanying GitHub repository `https://github.com/adamhaber/daleian_networks`. Network simulations, sensitivity computations, and optimizations were conducted on an internal cluster using the code above.