Developmental and evolutionary constraints on olfactory circuit selection

Naoki Hiratani* and Peter E. Latham
Gatsby Computational Neuroscience Unit, University College London

Across species, neural circuits show remarkable regularity, suggesting that their structure has been driven by underlying optimality principles. Here, we ask whether we can predict the neural circuitry of diverse species by optimizing the neural architecture to make learning as efficient as possible. We focus on the olfactory system, primarily because it has a relatively simple evolutionarily conserved structure, and because its input and intermediate layer sizes exhibits a tight allometric scaling. In mammals, it has been shown that the number of neurons in layer 2 of piriform cortex scales as the number of glomeruli (the input units) to the 3/2 power; in invertebrates, we show that the number of mushroom body Kenyon cells scales as the number of glomeruli to the 7/2 power. To understand these scaling laws, we model the olfactory system as a three layered nonlinear neural network, and analytically optimize the intermediate layer size for efficient learning from a limited number of samples. We find that the 3/2 scaling observed in mammals emerges naturally, both in full batch optimization and under stochastic gradient learning. We extended the framework to the case where a fraction of the olfactory circuit is genetically specified, not learned. We show numerically that this makes the scaling law steeper when the number of glomeruli is small, and we are able to recover the 7/2 scaling law observed in invertebrates. This study paves the way for a deeper understanding of the organization of brain circuits from an evolutionary perspective.

I. INTRODUCTION

Brains exhibit a large range of cell types, connectivity patterns, and organizational structures, at both micro and macro scales. There is a rich history in neuroscience of explaining these structures from a normative point of view [1–3]. Most of that work focused on computation, in the sense that it asked what circuit, and connection strengths, lead to optimal performance on a particular task. However, the connection strengths have to be learned, and model selection theory tells us that the efficiency of learning depends crucially on architecture, especially when a limited number of trials is available [4–8]. In this study, we attempt to understand the organizational structure of the brain from a model selection perspective, hypothesizing that evolution optimized the brain for efficiency of learning.

Here we focus on the olfactory system, primarily because it has a relatively simple, evolutionarily conserved, predominantly feedforward structure [9–11]. In particular, odorants are first detected by olfactory sensory neurons; from there, olfactory information is transmitted to glomeruli. The number of glomeruli, however, varies widely across species, from between 10 and 100 in insects to ∼1000 in mammals. The question we address here is: how does the number of glomeruli affect downstream circuitry? And in particular, what downstream circuitry would best help the animal survive? The tradeoffs that go into answering this question are in principle straightforward: more complicated circuitry (i.e., more parameters) can do a better job accurately predicting reward and punishment, but, because there are more parameters, there is a danger of overfitting [4, 7, 8]. And even if learning is performed with sample-by-sample updates to avoid overfitting, learning tends to be slower in complicated circuitry, as typically more samples are required [12, 13]. Navigating these tradeoffs for a given architecture is reasonably straightforward. The architecture, though, must come from biology. For that we take inspiration from the olfactory system of both mammals and invertebrates.

In the mammalian olfactory system, information from the glomeruli is transmitted to mitral/tufted cells, then to layer 2 of piriform cortex among others, and then mainly to layer 3; after that, information is passed on to higher order cortical areas [9, 11]. Thus, although many studies suggest that reciprocal interactions between mitral/tufted cells and granule cells are also important for olfactory processing [14–16], as a first-order approximation the olfactory system can be modeled as a feedforward neural network. Moreover, because sister mitral cells receiving input from the same glomeruli are highly correlated, both with each other and with the glomeruli from which they receive input [17], the olfactory network essentially has three layers: an input layer corresponding to glomeruli, a hidden layer corresponding to layer 2 of piriform cortex, and an output layer corresponding to layer 3.

Based on this picture, in our analysis we use an architecture corresponding to a three layer feedforward network. The size of the input layer is the number of glomeruli, and we assume that each unit of the output layer is extracting a different feature of the olfactory input, such as expected reward or punishment, or a behaviorally relevant concept. Consequently, we focus on the hidden layer. For that we ask: how many units should the hidden layer have? That question was chosen partly because its answer provides insight into learning principles in general, and partly because it was recently addressed experimentally: Srinivasan and Stevens found, based on

* N.Hiratani@gmail.com

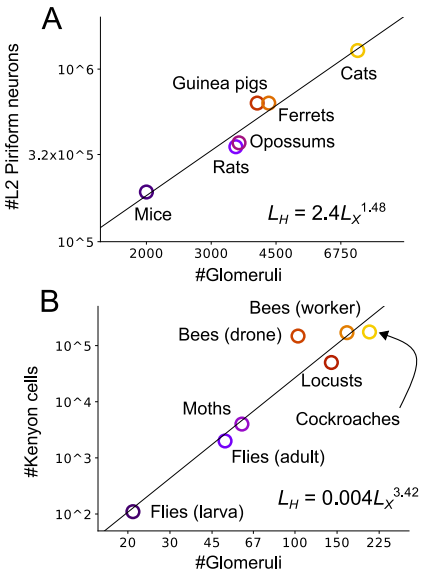

FIG. 1. **A)** Scaling law in mammalian olfactory circuits. Data was replicated from supplementary tables S2 and S3 of Srinivasan & Stevens, 2019. **B)** Scaling law in invertebrate olfactory circuits. See SI §1.1 for the details.

six mammalian species, a very tight relationship between the number of glomeruli and the number of neurons in layer two of piriform cortex (Fig. 1A; data taken from [18]). More precisely, using L_x to denote the input layer size (the number of glomeruli), and L_h to denote the hidden layer size (the number of neurons in layer 2 of piriform cortex), they found the approximate scaling law $L_h \sim L_x^{3/2}$.

Motivated by this result, we asked whether a similar scaling law holds for the invertebrate olfactory system. Like their mammalian counterparts, odors detected by olfactory sensory neurons converge to glomeruli. After that, though, the circuitry differs. Glomeruli send information to the projection neurons [11], which mainly extend synapses onto mushroom body Kenyon cells and lateral horn neurons [19]. The latter is mostly related to innate olfactory processing [20], so we focus on the mushroom body, which transmits information to higher-order regions through mushroom body output neurons. Thus, the invertebrate olfactory system can also be modeled as a three layer neural network: an input layer corresponding to glomeruli, a hidden layer corresponding to Kenyon cells, and an output layer corresponding to mushroom body output neurons [3, 21].

A literature survey of the number of glomeruli and Kenyon cells of various insects [22–34] (see SI §1.1 for details) yielded a scaling law, as in the mammalian olfactory system, but with an exponent of about 7/2 rather than 3/2 ($L_h \sim L_x^{7/2}$, as shown in Fig. 1B). Drone (male) bees are the clear outlier, but that is reasonable considering the caste system of honey bees that puts the drones under unique ecological pressure; for instance, the drones are the only ones among the seven insects listed that do not engage in foraging. It should be noted that the data was not properly controlled, as it was collected from different sources, and in some cases in different eras. Moreover, for the locust, we used the number of olfactory receptor genes instead of the number of glomeruli; that is because their micro-glomeruli structure makes direct comparison with other species difficult [35]. In addition, the mushroom body also takes part in visual processing in bees and cockroaches [36].

Several normative hypotheses have been offered to explain the population size of sensory circuits. One line of theoretical work showed that expansion in the hidden layer is beneficial for sensory coding [3, 37, 38], but it remains elusive how much expansion is optimal, because in these studies, more expansion was in principle always better. Other studies estimated the optimal population size in multiple layers from a width-depth tradeoff, assuming that the total number of neurons is fixed [39, 40] by external factors such as a constraint on energy [41]. However, this energy constraint should be violated if increasing the number of neurons improves foraging ability, resulting in a better energy budget [42]. Evaluation of the optimal population size was also attempted from other biological constraints such as synaptic [43] and neuronal [44] noise. While these models provided insight into circuit structure, none were able to provide a quantitative explanation for the population sizes of circuits across different species. Srinivasan and Stevens, on the other hand, offered several explanations, based on coding efficiency and geometry, for the scaling in mammals [18]. While those explanations are reasonable candidate hypotheses, they are more abstract than mechanistic, and do not explain the scaling seen in invertebrates.

Here we develop a mechanistic explanation of the scaling laws, focusing on the fact that the transformation from glomeruli to piriform cortex (for mammals), or from glomeruli to mushroom body output neurons (for invertebrates), has to be learned from a limited number of samples. For that we apply model selection theory, in which the primary constraint on the circuit is the poverty of the teaching signals and resultant overfitting [4, 7, 8]. The olfactory circuit has to tune its numerous synaptic weights from very sporadic, low-dimensional reward signals in the natural environment [45, 46], so this constraint should be highly relevant. Therefore, we formulate the problem of olfactory circuit design as a model selection problem, then analytically derive the optimal hidden layer size under various learning rules and nonlinearities.

Not surprisingly (because learning takes time) we find that the optimal hidden layer size depends on the lifetime of the organism. Using observed lifetimes, we recover the 3/2 scaling found in mammals. However, our theory cannot capture the 7/2 power law found in invertebrates. That is because traditional model selection theory fails to take into account the fact that neural circuits are at least partially genetically specified. In particular, rich innate connectivity structure is known to exist in the inverte-

brate olfactory systems [20, 47]. Thus, we extend the framework to the case where a fixed genetic budget can be used to specify connections, and consider how that affects scaling. The budget we used – about 2000 bits – had little effect on the scaling of the mammalian circuit, primarily because mammals have a large number of glomeruli, for which a complicated downstream circuit is needed to achieve good performance – far more complicated than could be constructed by 2000 bits. However, it had a large effect on invertebrates, which contain far fewer glomeruli. Using this extended framework, we were able to recover the observed 7/2 power law without disrupting the 3/2 power law found in mammals. These results shed light on potential constraints on the development and evolution of neural circuitry.

II. RESULTS

To determine scaling in the olfactory system, we use a teacher-student framework [13, 48, 49]: we postulate a teacher network, which reflects the true mapping from odors to reward or punishment in the environment, and model the circuit in the animal's olfactory system using the same overall architecture, but with different nonlinearities and a different number of neurons in the hidden layer (see Fig. 2). We determine the optimal hidden layer size under several scenarios: batch learning and stochastic gradient learning, and with and without information about the weights supplied by the genome.

A. The model

Let us denote the olfactory input at the level of glomeruli as $\boldsymbol{x} = \{x_1, x_2, ..., x_{L_x}\}$, and the corresponding reward, or punishment, as y. We consider a student-teacher model, and define the true relationship between $\boldsymbol{x}$ and y in the environment by a three layer "teacher" network (Fig. 2A),

$$y = \boldsymbol{w}_t \cdot g_t(\mathbf{J}_t \boldsymbol{x}) + \sigma_t \zeta, \tag{1}$$

where g_t is a pointwise nonlinear activation function of the hidden neurons, and ζ is Gaussian noise, added because the relationship between input and reward is stochastic in real world situations. Throughout the text we use bold capital letters to denote matrices and bold small letter for vectors. Vectors are defined as column vectors, a superscript T denotes transpose (indicating a row vector), and for readability we use a dot product to denote the inner product between two vectors. We sampled $\mathbf{J}_t$, $\boldsymbol{w}_t$, and $\boldsymbol{x}$ from independent Gaussian distributions for analytical tractability.

As discussed above, we model the olfactory circuits of both vertebrates and invertebrates as a three layer neural network (Fig. 2B),

$$\hat{y} = \boldsymbol{w}_s \cdot g_s(\mathbf{J}_s \boldsymbol{x}). \tag{2}$$

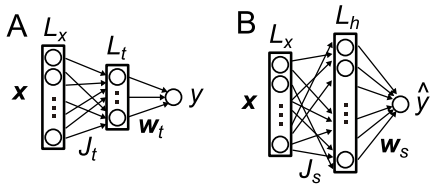

FIG. 2. Network models. **A**) Olfactory environment (teacher). **B**) Olfactory circuit that models the environment (student).

For simplicity, we assume that $\mathbf{J}_s$ is fixed and random, with elements drawn from an independent Gaussian distribution. Only the readout weights, $\boldsymbol{w}_s$, are learned from data. This is a good approximation for the invertebrate olfactory system, as the connection from the projection neurons to Kenyon cells are indeed mostly random [25] and fixed [50]. In the mammalian system, the connection from mitral/tufted cells to piriform cortex, which corresponds to $\mathbf{J}_s$, is suggested to be plastic [51]. However, it is thought that those connections are mainly shaped by unsupervised learning, but are seldom modulated by reward, as odor representation in layer 2 of piriform cortex is relatively stable under reward-based learning [52, 53].

The objective of learning is to predict the true reward signal, y, given the input, $\boldsymbol{x}$. Using the mean squared error as the loss, the generalization error is written

$$\epsilon_{gen} \equiv \left\langle (y - \hat{y})^2 \right\rangle \tag{3}$$

where here and in what follows angle brackets indicate an average over the input, $\boldsymbol{x}$, and the teacher noise, ζ. Under this problem setting, we ask what hidden layer size, L_h, minimizes the generalization error, ϵ_{gen}, when $\boldsymbol{w}_s$ is learned from N training samples. In particular, we investigate how the optimal hidden layer size scales with the input layer size, L_x. Intuitively, when the hidden layer size is too small, the neural network is not expressive enough, so the generalization error tends to be large even after an infinite number of training samples. On the other hand, if the hidden layer is too large relative to the number of training samples, the network becomes prone to over-fitting, again resulting in poor generalization. Here we solve this model selection problem analytically.

B. Generalization error

When the learning rule is unbiased, the generalization error consists of two components: the approximation error, which arises because we do not have a perfect model (we use $\mathbf{J}_s$ rather than the true weight, $\mathbf{J}_t$, to model the output, y, and we may have a different nonlinearity and hidden layer size), and the estimation error, which arises because we use a finite number of training samples [6–8]. Inserting Eqs. (1) and (2) into (3), we can write the

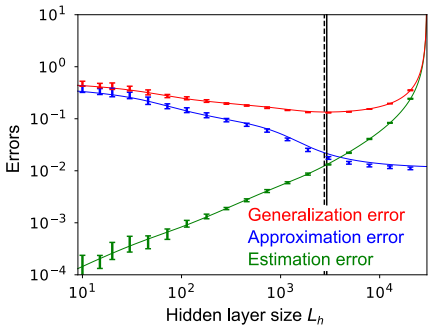

FIG. 3. Generalization, approximation, and estimation error at $L_x = 50$, $N = 30000$, under various hidden layer sizes L_h. Lines are analytical results (red, generalization error, Eq. (12); blue, approximation error, Eq. (9), and green, estimation error, Eq. (11)); points are from numerical simulations (see SI §7.3 for details). Here, and in all figures except Figs. 4D and E, both g_t and g_s are rectified linear functions ($g_t(u) = g_s(u) = \max(0, u)$). In all figures except Fig. 6 we use $\sigma_t^2 = 0.1$ for the noise in the teacher circuit, and in all figures the hidden layer size of the teacher network is fixed at $L_t = 500$. Error-bars represent the standard deviation over 10 simulation trials.

generalization error in terms of these two components,

$$\epsilon_{gen} = \sigma_t^2 + \epsilon_{apr} + \epsilon_{est}\,. \tag{4}$$

The approximation error, ϵ_{apr} (the error under the optimal weight $\boldsymbol{w}_s^*$), is given by

$$\epsilon_{apr} \equiv \left\langle \left(\boldsymbol{w}_t \cdot g_t(\mathbf{J}_t\boldsymbol{x}) - \boldsymbol{w}_s^* \cdot g_s(\mathbf{J}_s\boldsymbol{x})\right)^2 \right\rangle . \tag{5}$$

The estimation error, ϵ_{est} (the error induced by using the learned weight, $\boldsymbol{w}_s$, rather than the optimal one, $\boldsymbol{w}_s^*$), is

$$\epsilon_{est} \equiv \left\langle \left(\boldsymbol{w}_s^* \cdot g_s(\mathbf{J}_s\boldsymbol{x}) - \boldsymbol{w}_s \cdot g_s(\mathbf{J}_s\boldsymbol{x})\right)^2 \right\rangle . \tag{6}$$

Note that under an appropriate learning rule, ϵ_{est} converges to zero in the limit of an infinite number of training samples ($N \to \infty$).

We focus first on the approximation error, ϵ_{apr}, which depends on the optimal weight, $\boldsymbol{w}_s^*$. That weight is found by minimizing $\langle (y - \hat{y})^2 \rangle$ with respect to $\boldsymbol{w}_s$, with y and $\hat{y}$ given in Eqs. (1) and (2), respectively. This is a linear regression problem, and so $\boldsymbol{w}_s^*$ is given by the usual expression,

$$\boldsymbol{w}_s^* = \left\langle g_s(\mathbf{J}_s\boldsymbol{x}) g_s(\mathbf{J}_s\boldsymbol{x})^T \right\rangle^{-1} \left\langle g_s(\mathbf{J}_s\boldsymbol{x}) g_t(\mathbf{J}_t\boldsymbol{x})^T \right\rangle \boldsymbol{w}_t\,. \tag{7}$$

To compute $\boldsymbol{w}_s^*$, we need to invert a matrix. That is nontrivial because $g_s(\cdot)$ is a nonlinear function and the components of $\mathbf{J}_s\boldsymbol{x}$ are correlated,

$$\langle (\mathbf{J}_s\boldsymbol{x})_i (\mathbf{J}_s\boldsymbol{x})_j \rangle = \sum_{k=1}^{L_x} J_{ik}^s J_{jk}^s\,. \tag{8}$$

Because the J_{ik}^s are independent random variables, the off-diagonal elements are smaller than the diagonal elements by a factor of L_x. We can, therefore, compute $\boldsymbol{w}_s^*$ as an expansion in powers of $1/L_x$, multiplied by L_h (because there are factor of L_h more off-diagonal than diagonal elements). Working to second order in $1/L_x$, we show in SI §3 that

$$\epsilon_{apr} \approx \alpha + \frac{a_0}{L_h} + a_1 f\left(\frac{L_h}{L_x}, c_1\right) + a_2 f\left(\frac{2L_h}{L_x^2}, c_2\right) \tag{9}$$

where

$$f(z, c) \equiv \frac{\sqrt{(z + c - 1)^2 + 4c} - (z + c - 1)}{2} \tag{10}$$

is a monotonically decreasing function of z: $f(0, c) = 1$ and $f(z, c) \to c/z$ when $z \gg 1$. All constants are $\mathcal{O}(1)$; their values depend only on the nonlinearities $g_s(\cdot)$ and $g_t(\cdot)$. Note that ϵ_{apr} does not explicitly depend on the teacher network size L_t. That holds so long as $L_t \gg 1$ (see SI, Eqs. (36)-(39)).

As shown in Fig. 3 (blue line), ϵ_{apr} is a monotonically decreasing function of L_h. That function derives its shape from the three L_h-dependent terms in Eq. (9) (excluding α, which is a small constant): the second term, α_0/L_h, decays to zero when L_h is large compared to 1, the third decays to zero when L_h is large compared to L_x, and the last decays to zero when L_h is large compared to L_x^2. Essentially, as L_h increases, the effect of the off-diagonal elements of the covariance matrix in Eq. (7) increase, and the model becomes effectively more expressive (and thus lowers the approximation error). Although a number of approximations were made in deriving Eq. (9), the theoretical prediction matches well the numerical simulations (points in Fig. 3) for a wide range of L_h.

To complete the picture of the generalization error, we need the estimation error – the error associated with finite training data. For that it matters how we learn $\boldsymbol{w}_s$. There are two main choices: maximum likelihood estimation (MLE) and stochastic gradient descent (SGD). We consider MLE learning first. Although it is not biologically plausible (it requires the learner to compute, and invert, a covariance matrix after seeing all the data), we consider it first because it is reasonably straightforward. After that, we consider the more realistic case of SGD. Both exhibit the 3/2 scaling found in the mammalian olfactory circuit.

In SI §4.1, we extend the analysis in [54] to our maximum likelihood setting, and find that the estimation error from N samples is given by

$$\epsilon_{est} \approx (\epsilon_{apr} + \sigma_t^2)\frac{L_h}{N - L_h}\,. \tag{11}$$

This expression is intuitively sensible: in the limit of infinite data, $N \to \infty$, the estimation error vanishes, and in the opposite limit, $N \to L_h$, the estimation error blows up due to overfitting.

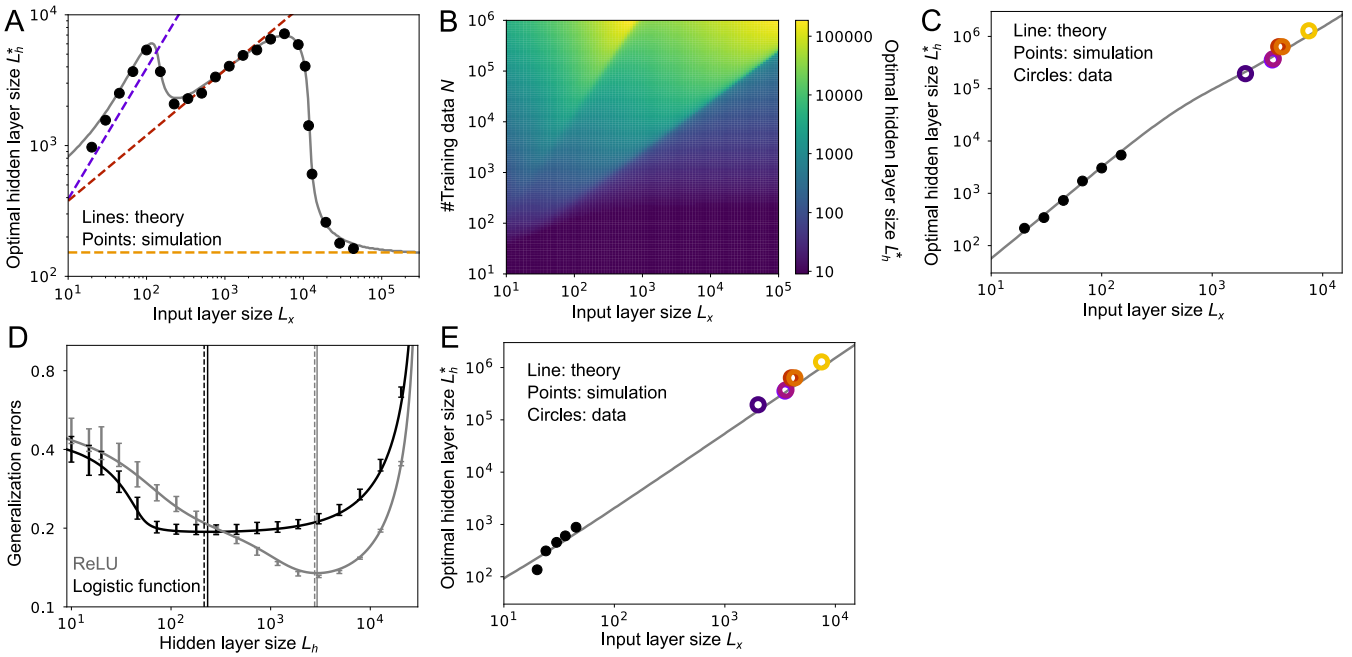

FIG. 4. Model behavior under maximum likelihood estimation. **A)** Relationship between the input layer size, L_x, and the optimal hidden layer size, L_h^*, with a fixed sample size ($N = 30000$). Gray lines are found by optimizing Eq. (12) with respect to L_h; dotted lines are the asymptotic expression derived in SI §5.1. **B)** Optimal hidden layer size, L_h^*, as a function of the input layer size, L_x, and the sample size, N, from Eq. (12). **C)** Scaling with $N = 2.7L_x^{1.9}$. Gray line is theory; black points are from simulations; colored circles are the experimental data from Fig. 1A. Simulations were done only for low L_x, due to the computational cost of the simulations when L_x is large. **D**) Relationship between the hidden layer size, L_h, and the generalization error, ϵ_{gen}, under the logistic activation function (black), and ReLU (gray), at $L_x = 50$ and $N = 30000$. Lines are theory; bars are from simulations. Error bars are the standard deviation over 10 simulations. **E**) Scaling for the logistic activation function with $N = 270L_x^{1.9}$. Gray line is theory; black points are from simulations; colored circles are the experimental data from Fig. 1A. As in panel C, simulations were done only for low L_x, due to the computational cost of the simulations when L_x is large. As in Fig. 3, the teacher network had a hidden layer size of 500, used a ReLU nonlinearity, and the noise was set to $\sigma_t^2 = 0.1$.

If σ_t^2 is not too small, ϵ_{est} is a monotonically increasing function of L_h, as shown in Fig. 3 (green line). In particular, when L_h is significantly smaller than the number of training samples, N, ϵ_{est} is a linearly increasing function of L_h, which is consistent with classical model selection theory [4, 12]. As L_h approaches N, the estimation error increases, and it goes to infinity when $L_h = N$, because the matrix on the right hand side of Eq. (7) becomes singular at that point.

Inserting ϵ_{est} from Eq. (11) into Eq. (4), the generalization error under MLE is

$$\epsilon_{gen} \approx (\epsilon_{apr} + \sigma_t^2)\frac{N}{N - L_h}\,. \tag{12}$$

The generalization error typically has a nontrivial global minimum as a function of L_h, as shown in Fig. 3 (red line). Moreover, the analytically estimated optimal hidden layer size, L_h^*, closely matches its estimation from numerical simulations (solid vs dotted vertical lines in Fig. 3).

C. Optimal hidden layer size

By minimizing the generalization error, Eq. (12) with respect to L_h (with the approximation error given by Eq. (9)), we can find the optimal hidden layer size, L_h^*, as a function of the input layer size, L_x. As shown in Fig. 4A, L_h^* has three different scalings. That is because only one term at a time in Eq. (9) is sensitive to L_h: the second term if $L_h \sim \mathcal{O}(1)$; the third term if $L_h \sim \mathcal{O}(L_x)$ and the fourth term if $L_h \sim \mathcal{O}(L_x^2)$. However, even considering one term at a time, minimizing Eq. (12) with respect to L_h is nontrivial, in large part because of the dependence on N. Details of the minimization are, therefore, left to SI, §5.1; here we simply summarize the results.

The optimal hidden layer size, L_h^*, roughly follows one of the three dotted lines in Fig. 4A, depending on the value of L_x. When the input layer size, L_x, is small compared to N, L_h^* is linear in L_x (purple line in Fig. 4A); when L_x is comparable to N, L_h scales as the square root of L_x (red line); and when L_x is larger than N, L_h stays constant as L_x changes (orange line). This last

scaling is reasonable because when the input layer is wide enough, expansion in the hidden layer is unnecessary. In all regions, L_h^* shows a square-root dependence on N, as suggested from previous studies [6, 8]. To further illustrate the dependence of L_h^* on L_x and N, in Fig. 4B we plotted the optimal hidden layer size versus these two quantities. This plot indeed shows three distinct phases separated by the lines $L_x \propto N$ and $L_x^2 \propto N$.

Figure 4B shows that the scaling relationship between L_h^* and L_x depends on N. Thus, to determine scaling across species, we need to know how N scales with L_x across species. We cannot directly measure N, which is the total amount of reward/teaching signal an animal experiences in its lifetime. However, we expect that N scales linearly with the duration of learning, so we can use that as a proxy for N. Among the six mammalian species, maximum longevity scales approximately as $L_x^{1.65}$ (SI, Fig. S1A; longevity data from AnAge database [55]). Alternatively, if we assume that learning happens mostly during the developmental period, here defined as the period from weaning to sexual maturation, a similar trend is observed, but with a slightly different exponent: duration from the time of weaning to sexual maturation scales approximately as $L_x^{1.97}$ (SI, Fig. S1B).

Given these observations, we assumed $N \propto L_x^\gamma$ with γ between 1.5 and 2. When we did that, we found a clear scaling law between L_x and L_h^* that spans more than three orders of magnitude. When we used $N = 2.7L_x^{1.9}$, the model reproduced the 3/2 scaling observed in the mammalian olfactory system (Fig. 4C). (Other values of γ gave slightly different scaling; see SI, Fig. S2A.)

In the above examples, we used ReLU for both teacher (g_t) and student (g_s), but this matching ($g_t = g_s$) is a rather strong assumption. To check the robustness of our results over the choice of the activation functions g, we used a logistic function g_s while keeping g_t as a ReLU. We found that the generalization error is minimized at a smaller hidden layer size compared to the ReLU student networks (black vs gray line in Fig. 4D; see SI §7.2 for details), primarily because large expansion is less helpful when the activation functions of the teacher and student networks are different. Nevertheless, assuming $N = 270L_x^{1.9}$, we obtain the experimentally observed 3/2 scaling law between L_h^* and L_x (Fig. 4E and Fig. S2B in SI).

D. Stochastic gradient descent (SGD) learning

So far we have considered learning by maximum likelihood estimation (MLE). However, that is not the best choice when the hidden layer size, L_h, is similar to the sample size N, as discussed above. In addition, batch learning is not particularly biologically plausible. Therefore, we consider online learning using stochastic gradient descent,

$$\boldsymbol{w}^{(n)} = \boldsymbol{w}^{(n-1)} + \eta(y_n - \hat{y}_n)g_s(\mathbf{J}_s\boldsymbol{x}_n) \tag{13}$$

where $\boldsymbol{w}^{(n)}$ is the readout weight after trial n and η is the learning rate. For online learning we consider minimization of the generalization error averaged over the lifetime of the organism, not the final error; that is because the fitness of an animal is much better characterized by the average proficiency during its lifetime than the proficiency at the end of its life.

Consistent with previous results [13], the learning rate that enables the fastest decay of the error is (see SI §4.2 for details)

$$\eta^* = \frac{2}{L_h}\,. \tag{14}$$

For this learning rate, the estimation error after N training samples is given approximately by (see SI §4.2, especially Eq. (91))

$$\bar{\epsilon}_{est}^{(N)} \approx \epsilon_{apr} + \sigma_t^2 + b_0 e^{-\frac{N}{\pi}} + b_1 e^{-\frac{N}{2L_1}} + b_2 e^{-\frac{N}{2\pi L_2}} + b_3 e^{-\frac{\alpha N}{2L_h}} \tag{15}$$

where

$$L_1 = \min(L_x, L_h) \tag{16a}$$

$$L_2 = \left[\min\left(\frac{L_x^2}{2}, L_h - L_x\right)\right]^+ \tag{16b}$$

(recall that $[\cdot]^+$ is the rectified linear function). The coefficients b_0, b_1, b_2 and b_3 depend on L_h, but not on N, and α (last term) is the same constant that appeared in Eq. (9).

The behavior of the estimation error under SGD is different than under MLE, Eq. (11), in two ways. First, for MLE, the estimation error goes to 0 as $N \to \infty$; for SGD, it asymptotes to a constant. That is because we used a fixed learning for the SGD update rule rather than letting it decay, as would be necessary to reduce the estimation error to zero [56]. Second, for MLE the estimation error diverges as L_h approaches N, whereas for SGD it remains finite. That is because of the online nature of SGD, which precludes overfitting.

As expected from Eq. (15), the estimation error as a function of the number of training samples, N, exhibits three components, all decaying with different timescales (Fig. 5B). The timescales of these, L_1, L_2, and L_h, are non-decreasing functions of L_h, as shown in Fig. 5A. Thus, larger L_h means slower decay, as can be seen in Fig. 5B. Therefore, unless L_h is small (where the estimation error decreases because the coefficients b_q depend on L_h), the lifetime average error increases with L_h, as can be seen in the green line in Fig. 5C. Notably, though, because the estimation error remains finite as $L_h \to N$, the lifetime average error does not diverge – in sharp contrast to maximum likelihood estimation, where it does diverge (compare the green line in Fig. 3 versus Fig. 5C). Because the approximation error decreases monotonically (blue line in Fig. 5C), the lifetime average generalization error (red line in Fig. 5C) tends to have a nontrivial global minimum.

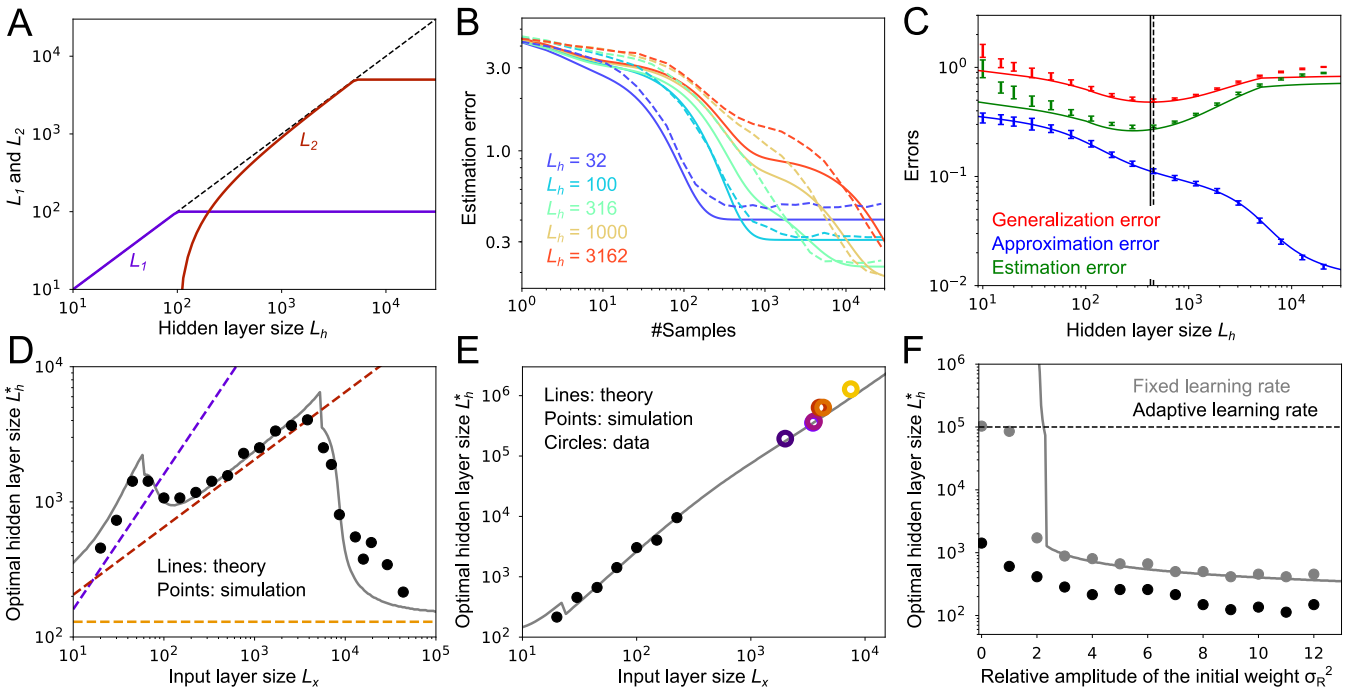

FIG. 5. Model behavior under stochastic gradient descent. **A)** Hidden layer size dependence of the decay time constant L_1 and L_2, with $L_x = 100$. **B)** Dynamics of the estimation error under various hidden layer sizes, L_h. Dashed lines: simulations; solid lines: theory. **C)** The cumulative generalization error, approximation error, and cumulative estimation error under various hidden layer sizes, L_h, at $N = 30000$. **D)** Optimal hidden layer size, L_h^*, with $N = 100000$. Dotted lines are asymptotic scaling (see SI §5.2). **E)** Optimal hidden layer size, L_h^*, with $N = 30L_x^{1.9}$. Gray line is theory; black points are from simulations; colored circles are the experimental data from Fig. 1A. As in Fig. 4, simulations were done only for low L_x, due to the computational cost of the simulations when L_x is large. **F)** Optimal hidden layer size, L_h^*, for various initial weight amplitudes, σ_R^2, and $N = 30000$. Gray: fixed learning rate; black: adaptive learning rate. Lines are theory and dots are simulations. The initial readout weights were sampled from $\boldsymbol{w}_s^{(0)} \sim N(0, \sigma_R^2/L_h)$. The horizontal dotted line represents the cutoff of L_h^* in the numerical simulations, meaning that at $\sigma_R^2 < 2$, under a fixed learning rate L_h^* is larger than 10^5. In panels A-C and F we set the input layer size to $L_x = 100$. As in Fig. 3, the teacher network had a hidden layer size of 500, used a ReLU nonlinearity, and the noise was set to $\sigma_t^2 = 0.1$.

As with MLE learning, under a fixed sample size N the optimal hidden layer size, L_h^*, shows three different scalings (dotted lines in Fig. 5D). This is because the approximation error decreases with three distinct phases (Eq. (9)). As a result, we observe effectively the same structure in SGD that we saw in MLE (Fig. 5D vs Fig. 4A), although the theoretical prediction at large L_x under SGD does not match quite as well as under MLE. But by introducing the scaling $N = 30L_x^{1.9}$, the experimentally observed scaling law in Fig. 1A is again reproduced (Fig. 5E).

In the model, we initialized the readout weights at relatively large values, $\boldsymbol{w}_s^{(0)} \sim N(0, 9.0/L_h)$. If the weights are instead initialized to small values, the optimal hidden layer size L_h^* diverges to infinity (gray line and points in Fig. 5F). This is partially because the fixed learning rate (Eq. (14)), employed for analytical tractability, causes poor convergence at small L_h. If an adaptive learning rate, $\eta_n = 2/\max(L_h, n)$, is used instead [16, 57], the cumulative generalization error is optimal at a finite hidden layer size even when the initial readout weights are zero (black points in Fig. 5F). Although the optimal hidden layer size, L_h^*, goes up as the initial weight amplitude σ_R^2 becomes smaller (Fig. 5F), the cumulative error becomes smaller under both fixed and adaptive learning rate (Fig. S3 in SI), due to smaller initial error.

E. Evolutionary constraints

The results so far indicate that developmental constraints explain the scaling law observed in the mammalian olfactory system. However, our analysis also revealed that developmental constraints alone do not explain the 7/2 power law scaling observed in the invertebrate olfactory circuit, suggesting the presence of additional principles. The primary candidate is a constraint on the genetic budget an animal can use to specify the olfactory circuit. We refer to this as an evolutionary constraint. Because both the number of protein-encoding genes and the total size of the genome tends to be similar across species [58], we assume that the genetic budget for the specification of

olfactory circuitry is similar among the insects listed on Fig. 1B.

Inspired by the insect olfactory circuitry, we consider a two-pathway model, in which projection neurons extend connections to both lateral horn neurons and Kenyon cells (Fig. 6A), and the output is

$$\hat{y} = \boldsymbol{w}_p \cdot g(\mathbf{J}_p \boldsymbol{x}) + \boldsymbol{w}_s \cdot g(\mathbf{J}_s \boldsymbol{x}) \tag{17}$$

where $\boldsymbol{w}_p \cdot g(\mathbf{J}_p \boldsymbol{x})$ is the pathway through lateral horn neurons. Although lateral horn neurons do not directly project to mushroom body output neurons, the two pathways eventually converge in the pre-motor area [26], where the output $\hat{y}$ might be represented. Because connections between projection neurons and lateral horn neurons tend to be stereotypical [20, 47], we assumed they were optimized on evolutionary timescales. We thus tuned the values of $\mathbf{J}_p$ and $\boldsymbol{w}_p$ while constraining the total information required for specifying the weights (see SI §6). In contrast, $\mathbf{J}_s$ was initialized randomly and fixed, and $\boldsymbol{w}_s$ was learned with adaptive SGD. Note that the initialization of $\mathbf{J}_s$ and $\boldsymbol{w}_s$ should require very little genetic information, compared to the hard-wired projection neuron-to-lateral horn neuron pathway. Using L_p to denote the number of lateral horn neurons, under a genetic information budget G, the amount of information encoded in the initial condition of $\mathbf{J}_p$ and $\boldsymbol{w}_p$ is bounded by

$$(L_p L_x + L_p) s_b < G, \tag{18}$$

where s_b is the number of bits per synapse. The first term is the number of bits needed to specify $\mathbf{J}_p$; the second is the number needed to specify $\boldsymbol{w}_p$. When only the presence/absence of connections is determined genetically, s_b is at most 1 bit; additional bits are needed if the weights are specified as well. Under a fixed budget, G, the number of bits per synapse, s_b, is bounded by $G/L_x L_p$, suggesting that as the input layer size, L_x, increases, tuning of $\mathbf{J}_p$ and $\boldsymbol{w}_p$ have to be more coarse-grained. In particular, in the mammalian olfactory system where $L_x \sim 10^3$, the hard-wired pathway should play a minor role unless $G > 10^4$. Indeed, except for encoding of pheromone signals, evidence of hard-wired connections in the mammalian olfactory circuits is limited [59]. For invertebrates, which have far fewer glomeruli, hard-wired pathways should be far more important. As the effect of the genetic budget, G, is difficult to characterize analytically, we numerically investigate its effect.

When we allowed information about the weights to be transmitted genetically, subject to the constraint given in Eq. (18), we found that the optimal Kenyon cell population size, L_h, was much smaller than the circuit without the projection neuron-to-lateral horn neuron pathway (compare 0 bit lines to 2 and 4 bit lines in Fig. 6B-D), leading to steeper scaling. Note that the two estimates became close at large L_x (as predicted above). Thus, for mammals, which have $L_x \sim 10^3$, genetic information has a negligible effect on scaling. In particular, we found that by setting $s_b = 2$, and $G = 2000$, the 7/2 scaling observed among insects is approximately reproduced (dark gray line in Fig. 6B). The predicted curve saturates at square scaling around $L_x \approx 150$, resulting in under-estimation of the Kenyon cell population in bees and cockroaches. This trend was observed under a different implementation of low-bit synapses (Fig. 6C), and even when $\boldsymbol{w}_p$ was additionally trained with SGD from finely-tuned initial weights (Fig. 6D).

III. DISCUSSION

In this work, we modeled the olfactory circuit of both mammals and insects as a three layer feedforward network, and asked how the number of neurons in the hidden layer scales with the number neurons in the glomerular (i.e., input) layer. We hypothesized that the scarcity of labeled signals (e.g., reward and punishment) provides a crucial constraint on the hidden layer size. We showed analytically, and confirmed with simulations, that this hypothesis robustly explains the experimentally observed 3/2 scaling found in the mammalian olfactory circuit, both under maximum likelihood (Fig. 4) and stochastic gradient descent (Fig. 5) learning. (Here "3/2 scaling" means the number of neurons in the hidden layer is proportional to the number of glomeruli to the 3/2 power.) This hypothesis alone does not, however, explain the 7/2 scaling found in the olfactory circuit of insects. But by considering the fact that genetic information used for constructing hard-wired olfactory connections is limited, we recovered the 7/2 scaling law (Fig. 6), without disrupting the 3/2 scaling law in mammals.

The 3/2 power in the scaling law we derived for mammals comes from two factors. First, when the number of training samples is fixed, the optimal population size of the piriform cortex increases as the number of glomeruli increases, unless the number of glomeruli is very large (Figs. 4A and 5D). Second, the optimal population size of the piriform cortex also increases with the number of training samples (Fig. 4B). Because species with more glomeruli tend to live longer and experience more samples (Fig. S1), this sample size dependence causes an additional scaling between the number of glomeruli and the piriform population size. From these two factors, the optimal intermediate layer size scales supra-linearly on number of the glomeruli (Figs. 4C, E, and 5E). Because of the dependence on the number of training samples, N, the power in the scaling law is not fixed at 3/2. In fact, depending on how N scales with the input layer size, L_x, theoretically any scaling is possible (SI §5). The 3/2 scaling we found was because in mammals, lifetime scales approximately quadratically with the number of glomeruli (SI §1.2, Table 2).

The three layer feedforward neural network with random fixed hidden weights is a class of neural networks that is widely studied from both biological [3, 37, 38, 60] and engineering [61, 62] perspectives. Under batch learn-

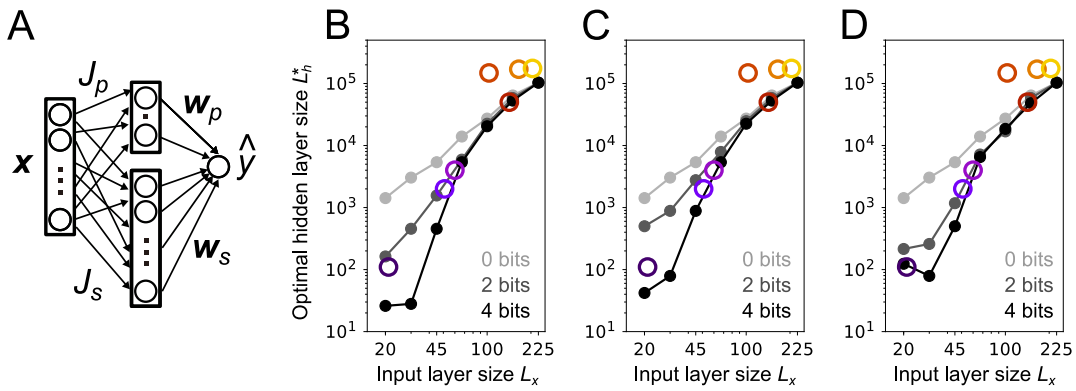

FIG. 6. **A)** Schematic of the two pathway model. **B-D)** Optimal layer size of the randomly initialized pathway $\boldsymbol{w}_s \cdot g(\mathbf{J}_s\boldsymbol{x})$ under different model settings. **B)** Low bit synapses were achieved by adding Gaussian noise to $\mathbf{J}_p$ and $\boldsymbol{w}_p$. **C)** Low bit synapses were achieved by discretizing $\mathbf{J}_p$ and $\boldsymbol{w}_p$. **D)** Low bit synapses were achieved by adding noise on $\mathbf{J}_p$ and $\boldsymbol{w}_p$ as in B, but $\boldsymbol{w}_p$ was additionally learned from training samples using SGD (see SI, Eq. (145)). In panels B-D, for the teacher network we used we had a hidden layer size of 500, used a ReLU nonlinearity, set the to $\sigma_t^2 = 0.01$, and used $N = 10L_x^2$ trials. For $s_b = 2$ bits we used $G = 2000$, while for $s_b = 4$ bits we used $G = 4000$, and for $s_b = 0$ bit, we simply removed the hard-wired pathway. The width of the hardwired intermediate layer, L_p, was found from Eq. (18): $L_p = G/s_b(L_x + 1)$, rounded up to an integer. See SI §6 and §7.4 for details.

ing, the upper bound on the approximation error for this network structure is known for a large class of the target functions [63, 64], yet these bounds are often too loose in practice. Here, we instead derived the average approximation error (SI §3). This allowed us to derive, analytically, accurate estimates of the optimal hidden layer size. The behavior of the estimation error is also well characterized in the large sample size limit ($N \to \infty$ while $L_x, L_h < \infty$) [12, 65], but this limit is not a good approximation of an over-parameterized neural network. On the other hand, the characteristics of the error at the large parameter space limit (number of synapses $\propto N$ as $N \to \infty$) remains mostly elusive, except for linear regression [54] (see also §4.1). Similarly, model selection in neural networks was previously studied mostly in the large sample size limit [7, 66]. The upper bound on the network size was also studied from VC (Vapnik-Chervonenkis) theory [5], and the minimum description length principle [6].

Learning dynamics in neural networks under SGD has also been widely studied [13, 49, 57]. In particular, recent results suggest that over-parameterization of a neural network does not harm the generalization error under both full-batch and stochastic gradient descent learning [67–70]. Here, though, we focused on the cumulative error, not the error at the end of training, as the former is more relevant to the fitness of the species. Under this objective function, over-parameterization does tend to harm performance, because learning becomes slower (Fig. 5B), even under an adaptive learning rate (Fig. 5F). Nevertheless, we found that if the initial weights are set to very small values and the learning rate is fixed, having infinitely many neurons in the hidden neuron minimizes the cumulative error (Fig. 5F), suggesting that over-parameterization is not always harmful, even when the cumulative error is the relevant cost function.

Scaling laws are ubiquitous in the brain. For instance, the number of neurons in the primary visual cortex scales with the 3/2 power against the population size of the LGN [71], while the number of neurons in the cerebral cortex is linear in the total number of neurons in the cerebellum [72]. Given the anatomical similarity between the olfactory circuit and cerebellum [3], our methodology should be directly applicable to understanding the latter scaling. But it is not limited to olfactory-like structures; it could be applied, possibly with some modifications, anywhere in the brain, and has the potential to provide insight into circuit structure in general.

Code availability The source codes of the simulations and the data analysis are deposited at Github (*https://github.com/nhiratani/olfactory_design*).

Acknowledgement This work was supported by the Gatsby Charitable Foundation and the Wellcome Trust (110114/Z/15/Z).

Competing interests The authors declare no competing interests.

[1] A. Mathis, A. V. Herz, and M. Stemmler, Neural computation **24**, 2280 (2012).
[2] J. Gjorgjieva, H. Sompolinsky, and M. Meister, Journal of Neuroscience **34**, 12127 (2014).
[3] A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. Abbott, Neuron **93**, 1153 (2017).
[4] H. Akaike, IEEE transactions on automatic control **19**, 716 (1974).
[5] E. B. Baum and D. Haussler, in *Advances in neural information processing systems* (1989) pp. 81–90.
[6] A. R. Barron, Machine learning **14**, 115 (1994).
[7] N. Murata, S. Yoshizawa, and S.-i. Amari, IEEE Transactions on Neural Networks **5**, 865 (1994).
[8] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning* (MIT press, 2018).
[9] R. L. Davis, Neuron **44**, 31 (2004).
[10] B. W. Ache and J. M. Young, Neuron **48**, 417 (2005).
[11] R. I. Wilson and Z. F. Mainen, Annu. Rev. Neurosci. **29**, 163 (2006).
[12] S.-I. Amari and N. Murata, Neural Computation **5**, 140 (1993).
[13] J. Werfel, X. Xie, and H. S. Seung, in *Advances in neural information processing systems* (2004) pp. 1197–1204.
[14] O. Gschwend, N. M. Abraham, S. Lagier, F. Begnaud, I. Rodriguez, and A. Carleton, Nature neuroscience **18**, 1474 (2015).
[15] A. Grabska-Barwińska, S. Barthelmé, J. Beck, Z. F. Mainen, A. Pouget, and P. E. Latham, Nature neuroscience **20**, 98 (2017).
[16] N. Hiratani and P. E. Latham, Nature communications **11**, 1 (2020).
[17] A. K. Dhawale, A. Hagiwara, U. S. Bhalla, V. N. Murthy, and D. F. Albeanu, Nature neuroscience **13**, 1404 (2010).
[18] S. Srinivasan and C. F. Stevens, Current Biology **29**, 2533 (2019).
[19] J. P. Martin, A. Beyerlein, A. M. Dacks, C. E. Reisenman, J. A. Riffell, H. Lei, and J. G. Hildebrand, Progress in neurobiology **95**, 427 (2011).
[20] M. Fişek and R. I. Wilson, Nature neuroscience **17**, 280 (2014).
[21] R. Huerta, T. Nowotny, M. García-Sanchez, H. D. Abarbanel, and M. I. Rabinovich, Neural computation **16**, 1601 (2004).
[22] A. Ramaekers, E. Magnenat, E. C. Marin, N. Gendre, G. S. Jefferis, L. Luo, and R. F. Stocker, Current biology **15**, 982 (2005).
[23] L. M. Masuda-Nakagawa, N. Gendre, C. J. O'Kane, and R. F. Stocker, Proceedings of the National Academy of Sciences **106**, 10314 (2009).
[24] K. Eichler, F. Li, A. Litwin-Kumar, Y. Park, I. Andrade, C. M. Schneider-Mizell, T. Saumweber, A. Huser, C. Eschbach, B. Gerber, *et al.*, Nature **548**, 175 (2017).
[25] S. J. Caron, V. Ruta, L. Abbott, and R. Axel, Nature **497**, 113 (2013).
[26] Y. Aso, D. Hattori, Y. Yu, R. M. Johnston, N. A. Iyer, T.-T. Ngo, H. Dionne, L. Abbott, R. Axel, H. Tanimoto, *et al.*, Elife **3**, e04577 (2014).
[27] S. Anton and B. S. Hansson, Journal of comparative neurology **350**, 199 (1994).
[28] M. Sjöholm, I. Sinakevitch, R. Ignell, N. J. Strausfeld, and B. S. Hansson, Journal of Comparative Neurology **491**, 290 (2005).
[29] Z. Wang, P. Yang, D. Chen, F. Jiang, Y. Li, X. Wang, and L. Kang, Cellular and Molecular Life Sciences **72**, 4429 (2015).
[30] B. Leitch and G. Laurent, Journal of comparative Neurology **372**, 487 (1996).
[31] G. Arnold, C. Masson, and S. Budharugsa, Cell and tissue research **242**, 593 (1985).
[32] W. Witthöft, Zeitschrift für Morphologie der Tiere **61**, 160 (1967).
[33] H. Watanabe, H. Nishino, M. Nishikawa, M. Mizunami, and F. Yokohari, Journal of Comparative Neurology **518**, 3907 (2010).
[34] S. M. Farris and N. J. Strausfeld, Journal of Comparative Neurology **439**, 331 (2001).
[35] K. Ernst, J. Boeckh, and V. Boeckh, Cell and tissue research **176**, 285 (1977).
[36] E. Capaldi, G. Robinson, and S. Fahrbach, Annual review of psychology **50**, 651 (1999).
[37] O. Barak, M. Rigotti, and S. Fusi, Journal of Neuroscience **33**, 3844 (2013).
[38] B. Babadi and H. Sompolinsky, Neuron **83**, 1213 (2014).
[39] A. M. Hermundstad, K. S. Brown, D. S. Bassett, and J. M. Carlson, PLoS computational biology **7**, e1002063 (2011).
[40] J. Kadmon and H. Sompolinsky, in *Advances in Neural Information Processing Systems* (2016) pp. 4781–4789.
[41] L. C. Aiello and P. Wheeler, Current anthropology **36**, 199 (1995).
[42] A. Navarrete, C. P. van Schaik, and K. Isler, Nature **480**, 91 (2011).
[43] D. V. Raman, A. P. Rotondo, and T. O'Leary, Proceedings of the National Academy of Sciences **116**, 10537 (2019).
[44] L. Yu, C. Zhang, L. Liu, and Y. Yu, Scientific reports **6**, 19369 (2016).
[45] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, Behavioral and Brain Sciences **40** (2017).
[46] N. E. Raine and L. Chittka, Proceedings of the Royal Society B: Biological Sciences **275**, 803 (2008).
[47] S. Frechter, A. S. Bates, S. Tootoonian, M.-J. Dolan, J. D. Manton, A. R. Jamasb, J. Kohl, D. Bock, and G. S. Jefferis, eLife **8**, e44590 (2019).
[48] H. Sompolinsky, N. Tishby, and H. S. Seung, Physical Review Letters **65**, 1683 (1990).
[49] A. M. Saxe, J. L. McClelland, and S. Ganguli, arXiv preprint arXiv:1312.6120 (2013).
[50] T. Hige, Y. Aso, M. N. Modi, G. M. Rubin, and G. C. Turner, Neuron **88**, 985 (2015).
[51] J. Chapuis and D. A. Wilson, Nature neuroscience **15**, 155 (2012).
[52] D. J. Millman and V. N. Murthy, Journal of Neuroscience (2020).
[53] P. Y. Wang, C. Boboila, M. Chin, A. Higashi-Howard, P. Shamash, Z. Wu, N. P. Stein, L. Abbott, and R. Axel, Neuron (2020).
[54] M. Advani and S. Ganguli, Physical Review X **6**, 031034 (2016).
[55] R. Tacutu, T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha, J. Costa, V. E. Fraifeld, and J. a. P. De Magalhães, Nucleic acids research **41**,

D1027 (2012).
[56] L. Bottou, On-line learning in neural networks **17**, 142 (1998).
[57] S. Amari, IEEE Transactions on Electronic Computers , 299 (1967).
[58] L. Pray, Nature Education **1**, 96 (2008).
[59] K. K. Ishii, T. Osakada, H. Mori, N. Miyasaka, Y. Yoshihara, K. Miyamichi, and K. Touhara, Neuron **95**, 123 (2017).
[60] S. Ganguli and H. Sompolinsky, Annual review of neuroscience **35**, 485 (2012).
[61] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, Neurocomputing **70**, 489 (2006).
[62] S. Dasgupta, C. F. Stevens, and S. Navlakha, Science **358**, 793 (2017).
[63] A. Rahimi and B. Recht, in *Advances in neural information processing systems* (2009) pp. 1313–1320.
[64] X. Liu, S. Lin, J. Fang, and Z. Xu, IEEE Trans. Neural Netw. Learning Syst. **26**, 7 (2015).
[65] A. W. Van der Vaart, *Asymptotic statistics*, Vol. 3 (Cambridge university press, 2000).
[66] N. Barkai, H. Seung, and H. Sompolinsky, Physical review letters **70**, 3167 (1993).
[67] M. S. Advani, A. M. Saxe, and H. Sompolinsky, Neural Networks (2020).
[68] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Proceedings of the National Academy of Sciences **116**, 15849 (2019).
[69] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, in *Advances in Neural Information Processing Systems* (2019) pp. 6981–6991.
[70] S. d'Ascoli, L. Sagun, and G. Biroli, arXiv preprint arXiv:2006.03509 (2020).
[71] C. F. Stevens, Nature **411**, 193 (2001).
[72] S. Herculano-Houzel, Frontiers in neuroanatomy **4**, 12 (2010).