Developmental and evolutionary constraints on olfactory circuit selection

Naoki Hiratani[a,1,2] and Peter E. Latham[a]

[a]Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, United Kingdom

Edited by Liqun Luo, Department of Biology, Stanford University, Stanford, CA; received January 14, 2021; accepted January 14, 2022

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 exhibit 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-layer nonlinear neural network and analytically optimize the intermediate-layer size for efficient learning from limited samples. We find, as observed, a power-law scaling, with the exponent depending strongly on the number of samples and thus on longevity. The 3/2 scaling seen in mammals is consistent with observed longevity, but the 7/2 scaling in invertebrates is not. However, when a fraction of the olfactory circuit is genetically specified, not learned, scaling becomes steeper for species with a small number of glomeruli and recovers consistency with the invertebrate scaling. This study provides analytic insight into the principles underlying both allometric scaling across species and optimal architectures in artificial networks.

olfaction | neural circuit | model selection | statistical learning theory

Brains exhibit a large range of cell types, connectivity patterns, and organizational structures, at both micro- and macroscales. 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, leads 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 are available (4–8). This is also true for deep networks, where the choice of neural architecture plays a critical role in both learning speed and performance (9). Here we attempt to understand the organizational structure of the brain from a model selection perspective, hypothesizing that evolution optimized the brain for efficient learning.

We build a model inspired by the olfactory circuitry and study its allometric scaling analytically. We focus on the olfactory system primarily because it has a relatively simple, evolutionarily conserved, predominantly feedforward structure (10–12). 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 ∼1,000 in mammals. The question we address 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 (13, 14). Navigating these tradeoffs requires that we choose an architecture, which 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 (10, 12). Thus, although many studies suggest that reciprocal interactions between mitral/tufted cells and granule cells (15, 16), as well as feedback from the cortex (17, 18), are also important for olfactory processing, 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 (19), 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

Significance

In this work, we explore the hypothesis that biological neural networks optimize their architecture, through evolution, for learning. We study early olfactory circuits of mammals and insects, which have relatively similar structure but a huge diversity in size. We approximate these circuits as three-layer networks and estimate, analytically, the scaling of the optimal hidden-layer size with input-layer size. We find that both longevity and information in the genome constrain the hidden-layer size, so a range of allometric scalings is possible. However, the experimentally observed allometric scalings in mammals and insects are consistent with biologically plausible values. This analysis should pave the way for a deeper understanding of both biological and artificial networks.

Author contributions: N.H. and P.E.L. designed research; N.H. performed research; and N.H. and P.E.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: N.Hiratani@gmail.com.

[2]Present address: Center for Brain Science, Harvard University, Cambridge, MA 02138.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2100600119/-/DCSupplemental.

Published March 9, 2022.

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 (20) found, based on six mammalian species, a very tight relationship between the number of glomeruli and the number of neurons in layer 2 of piriform cortex (Fig. 1*A*; data taken from ref. 20). 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 (12), which mainly extend synapses onto mushroom body Kenyon cells and lateral horn neurons (21). The latter is mostly related to innate olfactory processing (22), so we focus on the mushroom body, which transmits information to higher-order regions through mushroom body output neurons and is considered to be the learning center of the insect brain (23, 24). Insect olfactory circuits also contain various nonfeedfoward connections, such as lateral inhibition between the projection neurons (12). But, as with the mammalian olfactory system, as a first-order approximation we omit them from the model. 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, 25).

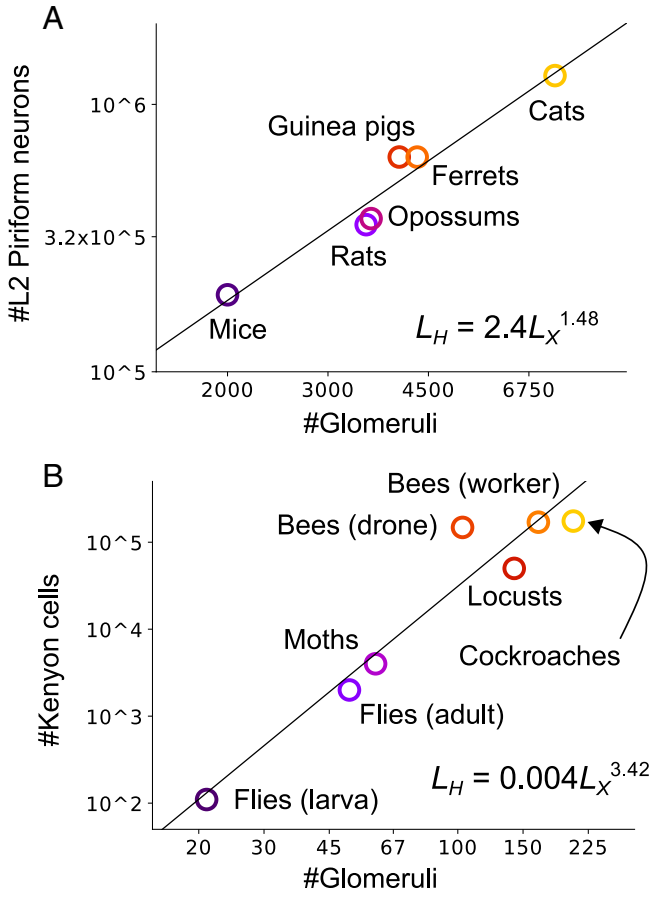

Fig. 1. (*A*) Scaling law in mammalian olfactory circuits. Data points were taken from supplementary tables S2 and S3 of Srinivasan and Stevens (20). (*B*) Scaling law in invertebrate olfactory circuits. See *SI Appendix*, section 1.1 for details.

A literature survey of the number of glomeruli and Kenyon cells of various insects (23, 24, 26–35) (see *SI Appendix*, section 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. 1*B*). Drone (male) bees are the clear outlier. That might be because the caste system of honey bees 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 were not properly controlled, as they were 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 microglomeruli structure makes direct comparison with other species difficult (36). In addition, the mushroom body also takes part in visual processing in bees and cockroaches (37).

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, 38, 39), 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 (40, 41) by external factors such as a constraint on energy (42). However, this energy constraint should be violated if increasing the number of neurons improves foraging ability, resulting in a better energy budget (43). Evaluation of the optimal population size was also attempted from other biological constraints, such as synaptic (44) and neuronal (45) 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 (20), on the other hand, offered a quantitative derivation of the 3/2 power law observed in mammals. Their derivation was based on the hypothesis that not much information is lost between areas. While this is a reasonable hypothesis, their derivation relied on several implicit assumptions; in particular, they assumed that the noise between different neurons is uncorrelated, and the olfactory signals are not mixed as they propagate across layers. However, both correlations and mixing are likely to exist, and that will affect the number of neurons required in downstream areas (46). In addition, their theory does not explain the 7/2 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. That explanation draws on model selection theory, in which the primary constraint is the poverty of teaching signals and resultant overfitting (4, 7, 8). Because the olfactory circuit has to tune its numerous synaptic weights from very sporadic, low-dimensional reward signals in the natural environment (47, 48), this constraint should be highly relevant. Therefore, we formulate the problem of olfactory circuit design as a model selection problem and then derive the optimal hidden-layer size under various learning rules and nonlinearities.

Our derivation proceeds in two steps: First, we expand the covariance matrix in the hidden layer in powers of the average hidden-layer correlation coefficient; then we use random matrix theory to compute the generalization error. That enabled us to determine, analytically, the factors that control the optimal hidden-layer size, thus bypassing the intensive numerical optimization typically used in deep-learning settings (9, 40, 49). Our analysis shows that the optimal hidden-layer size follows an allometric scaling with the input-layer size and reveals the factors that control the scaling exponent. Not surprisingly (because

learning takes time) we find that the optimal hidden-layer size, and thus the scaling exponent, depends on the lifetime of the organism. The 3/2 scaling found in mammals is, though, largely consistent with observed lifetimes. This scaling relationship is robust against the choice of nonlinearity, activity sparseness, and the noise level and also against the optimization method: It holds under both maximum-likelihood estimation and stochastic gradient descent with cumulative error minimization.

Our theory was not, however, able to 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 invertebrate olfactory systems (22, 50). 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 2,000 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 2,000 bits. However, it had a large effect on invertebrates, which contain far fewer glomeruli. In particular, the scaling became steeper, making it possible to replicate the observed 7/2 power law without disrupting the 3/2 power law in mammals. These results shed light on potential constraints on the development and evolution of neural circuitry.

Results

To determine scaling in the olfactory system, we use a teacher–student framework (14, 51, 52): We postulate a teacher network, which reflects the true mapping from odors to reward or punishment in the environment, and model the olfactory network using the same overall architecture, but with different nonlinearities and a different number of neurons in the hidden layer (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.

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

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

where g_t is a pointwise nonlinear activation function and ζ is Gaussian noise, added because the relationship between input and reward is stochastic in real-world situations. Throughout the text we use uppercase letters in boldface type to denote matrices and lowercase letters in boldface type 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. Note that we used a continuous, rather than binary, valence y because odor-driven animal behavior is sensitive to both the sign and the value of valence (53).

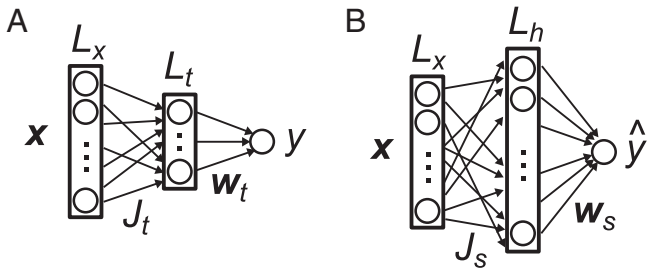

Fig. 2. Network models. (*A*) Olfactory environment (teacher). (*B*) Olfactory circuit that models the environment (student).

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

$$\hat{y} = \boldsymbol{w}_s \cdot g_s(\mathbf{J}_s \boldsymbol{x}), \qquad [2]$$

where $\mathbf{J}_s$ is an $L_h \times L_x$ matrix connecting the input neurons to the neurons in the hidden layer, and $\boldsymbol{w}_s$ is an L_h -dimensional vector connecting the hidden-layer neurons to the output neuron. 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 is indeed mostly random (23) and fixed (54). In the mammalian system, the connection from mitral/tufted cells to piriform cortex, which corresponds to $\mathbf{J}_s$, is suggested to be plastic (55). 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 (56, 57).

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

$$\epsilon_{gen} \equiv \left\langle (y - \hat{y})^2 \right\rangle, \qquad [3]$$

where 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 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 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 large relative to the number of training samples, the network becomes prone to overfitting, again resulting in poor generalization. Below we address this tradeoff quantitatively.

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 generalization error in terms of these two components,

$$\epsilon_{gen} = \sigma_t^2 + \epsilon_{apr} + \epsilon_{est}, \qquad [4]$$

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

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

and 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

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

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,

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

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,

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

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 factors of L_h more off-diagonal than diagonal elements). Working to second order in $1/L_x$, we show in *SI Appendix*, section 3 that

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

where

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

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 our expression for ϵ_{apr} does not explicitly depend on the teacher network size L_t. That holds as long as $L_t \gg 1$ (*SI Appendix*, Eqs. **S38–S41**).

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**: 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 effects of the off-diagonal elements of the covariance matrix in Eq. **7** increase, and the model becomes more expressive (and thus lowers the approximation error). Although a number of approximations were made in deriving Eq. **9**, the theoretical prediction (blue line in Fig. 3) matches well the numerical simulations (points) for a wide range of L_h.

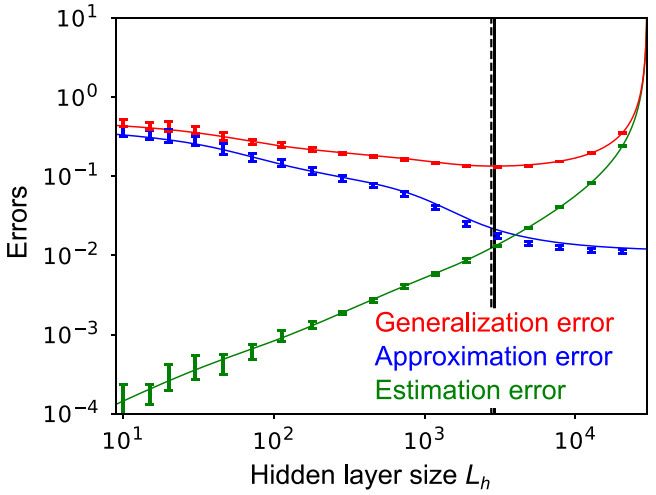

Fig. 3. Generalization error (red; Eq. **12**), approximation error (blue; Eq. **9**), and estimation error (green; Eq. **11**) at $L_x = 50$, $N = 30{,}000$, for various hidden-layer sizes L_h. Lines are analytical results; points are from numerical simulations (see *SI Appendix*, section 7.4 for details). Solid and dashed vertical lines are the minima of the generalization error from theory and simulations, respectively. Here, and in all figures except Fig. 4 *D* 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 SD over 10 simulations.

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 start with MLE. 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.

MLE Learning. In *SI Appendix*, section 4.1, we extend the analysis in ref. 58 to our maximum-likelihood setting and find that the estimation error from N samples is given by

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

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.

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, 13). Note that when $L_h \ll N$, the estimation error is small. That is because there are very few parameters, and so the network learns them almost perfectly. As L_h approaches N, the estimation error increases, and at $L_h = N$ it goes to infinity. The divergence at $L_h = N$ arises because the matrix on the right-hand side of Eq. **7** becomes singular.

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

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

The first term on the right-hand side is a decreasing function of L_h; the second term is an increasing function. Together, they produce a generalization error (red line in Fig. 3) that typically has a unique global minimum as a function of L_h. Moreover, the analytically estimated optimal hidden-layer size, L_h^*, closely matches its estimation from numerical simulations (solid vertical line vs. dashed vertical line in Fig. 3).

The generalization error has a concrete interpretation in terms of discriminability of valence: It tells us the resolution with which valence can be inferred and in particular how far apart odors need to be before we can confidently assign different valence to them. Because our learned network has low generalization error at the optimal hidden-layer size, its resolution at that point is relatively high. For instance, if odors are parallel to the learned direction, they can have a correlation coefficient of 0.99 and still be assigned distinct valences (blue line in *SI Appendix*, Fig. S1). If, on the other hand, the odors are randomly selected, the correlation coefficient must be below 0.85 (still relatively high) for the circuit to assign distinct valences (orange line in *SI Appendix*, Fig. S1).

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^*,

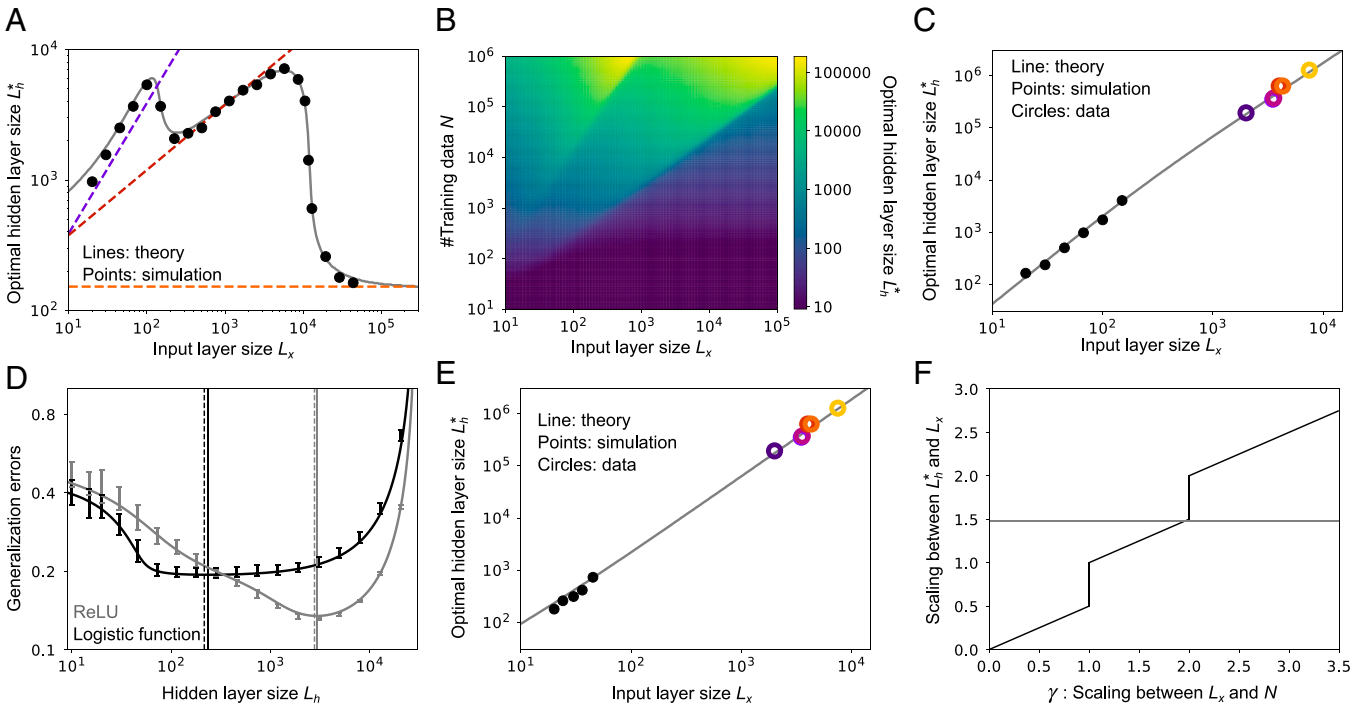

Fig. 4. Model behavior under maximum-likelihood estimation. (*A*) Relationship between the input-layer size, L_x, and the optimal hidden-layer size, L_h^*, at a fixed sample size ($N = 30{,}000$). Gray lines are found by optimizing Eq. **12** with respect to L_h; dashed lines are the asymptotic expression derived in *SI Appendix*, section 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 at $N = 1.65L_x^{1.96}$. Gray line is theory; black points are from simulations; colored circles are the experimental data from Fig. 1*A*. 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 $N = 30{,}000$. Lines are theory; bars are from simulations. Vertical lines mark the minima (solid, theory; dashed, simulations). Error bars are the SD over 10 simulations. (*E*) Scaling for the logistic activation function with $N = 240L_x^{1.96}$. Gray line is theory; black points are from simulations; colored circles are the experimental data from Fig. 1*A*. As in *C*, simulations were done only for low L_x, due to the computational cost of the simulations when L_x is large. (*F*) Analytical estimation of the L_h^* - L_x scaling versus the L_x-N scaling (y axis, coefficient β in the scaling $L_x^* \propto L_x^\beta$; x axis, coefficient γ in the scaling $L_x \propto N^\gamma$; see *SI Appendix*, section 5.1 for details). The gray horizontal line is the 3/2 scaling from Fig. 1*A*. As in Fig. 3, the teacher network had a hidden-layer size of 500, with a ReLU nonlinearity, and the noise was set to $\sigma_t^2 = 0.1$.

as a function of the input-layer size, L_x. As shown in Fig. 4*A*, 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 Appendix*, section 5.1; here we simply summarize the results.

The optimal hidden-layer size, L_h^*, roughly follows one of the three dashed lines in Fig. 4*A*, depending on the value of L_x relative to N. When the input layer size, L_x, is small compared to N, L_h^* is linear in L_x (purple dashed line in Fig. 4*A*); when L_x is comparable to N, L_h scales as the square root of L_x (red dashed line); and when L_x is larger than N, L_h stays constant as L_x changes (orange dashed line). This last scaling is reasonable because when the input layer is wide enough, expansion in the hidden layer is unnecessary. To further illustrate the dependence of L_h^* on L_x and N, in Fig. 4*B* we plot the optimal hidden-layer size versus these two quantities. This indeed shows three distinct phases separated by the lines $L_x \propto N$ and $L_x^2 \propto N$. This three-phase structure is robust to the choice of N and σ_t^2 (*SI Appendix*, Fig. S2*A*), although these parameters introduce an overall scale factor in the optimal hidden-layer size, L_h^*.

While L_h^* shows a relatively nontrivial dependence on L_x, its dependence on N (with L_x fixed) is simple: $L_h^* \propto \sqrt{N}$. That is because the effective approximation and estimation errors scale as $1/L_h$ and L_h/N, respectively (*SI Appendix*, section 5.1). Thus, to minimize the generalization error, which is the sum of these two terms, L_h^* needs to satisfy $L_h^* \propto \sqrt{N}$. This simple scaling is consistent with previous work, which suggests that the $1/L_h$ scaling of the MSE approximation error (59) and the L_h/N dependence of the estimation error (13) are robust to model settings.

In our analysis, we assumed that the activity of the glomeruli, $\boldsymbol{x}$, follows an independent Gaussian distribution. However, because the intrinsic dimensionality of the activity is bounded by the number of olfactory receptor genes, the glomeruli activity is not necessarily independent. This is especially relevant for mammalian olfactory circuits, where the number of glomeruli is typically larger than the number of olfactory receptor genes (20). To investigate this issue, we computed numerically the optimal hidden-layer size when the number of olfactory receptor genes was fixed. We found that if the activity at the glomeruli is whitened via a nonlinearity and lateral inhibition, the optimal hidden-layer size shows the same scaling with the number of glomeruli as it does when the number of glomeruli is equal to the number of olfactory receptor genes (*SI Appendix*, Fig. S3 and section 7.6). Thus, even though the intrinsic dimensionality of the activity is smaller than the number of glomeruli, because of the nonlinearities and whitening the effective dimensionality scales with the number of glomeruli.

Fig. 4*B* 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 number of rewards/teaching signals an animal experiences in its lifetime. However, we expect that N scales linearly with the duration of learning, so we use that as a proxy. Among the six mammalian

species, maximum longevity scales approximately as $L_x^{1.65\pm0.45}$ (*SI Appendix*, Fig. S4*A*; longevity data from AnAge database) (60). 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 $L_x^{1.97\pm0.58}$ (*SI Appendix*, Fig. S4*B*).

Given these observations, we assumed $N \propto L_x^\gamma$ with γ between 1.6 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. That is because, unlike Fig. 4*A* where an increase in L_x causes a phase transition, when N also increases with L_x the system stays in one of the phases (here, the second one). When we set N to $N = 1.65L_x^{1.96}$, the model reproduced the 3/2 scaling observed in the mammalian olfactory system (Fig. 4*C*). The coefficient, 1.65, and the exponent, $\gamma = 1.96$, were selected to match the data in Fig. 1*A* (*SI Appendix*, Fig. S5*A*); notably, though, the exponent fell into the expected range from *SI Appendix*, Fig. S4 (1.6 to 2.0). Other values of γ gave slightly different scaling (*SI Appendix*, Fig. S5*B*).

In the above examples, we used rectified linear units (ReLU) for both teacher (g_t) and student (g_s), but this matching ($g_t = g_s$) might be a strong assumption. To check the robustness of our results with respect to the choice of activation function, we used a logistic function for the student (g_s) while keeping a ReLU for the teacher (g_t). With this choice, the generalization error is minimized at a smaller hidden-layer size compared to the ReLU student networks (black line vs. gray line in Fig. 4*D*; see *SI Appendix*, section 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, as above, $N \propto L_x^{1.96}$, we obtain the experimentally observed 3/2 scaling law between L_h^* and L_x (Fig. 4*E* and *SI Appendix*, Fig. S5*C*). To achieve this scaling, however, the coefficient needs to be larger than when the teacher and student nonlinearity matched ($N = 240L_x^{1.96}$ versus $1.65L_x^{1.96}$ for the matching case).

So far, we have used parameters for which the activity in the intermediate layer is dense, meaning roughly half of the units are active for each odor. However, both Kenyon cells and layer 2 piriform neurons show sparse selectivity for olfactory stimuli (21, 61). To study the effect of sparse selectivity, we introduce a bias to the ReLU nonlinearity, $g_s(u) = \max(u - b, 0)$, with b sufficiently large that only a small fraction of neurons in the hidden layer show nonzero activity for each odor (*SI Appendix*, section 7.3). In this regime, the model prefers larger expansion in the hidden layer (*SI Appendix*, Fig. S6*A*), but we still observed a 3/2 scaling law between L_h^* and L_x (*SI Appendix*, Fig. S6*B*), indicating that the scaling law is robust with respect to the sparseness of activity.

Finally, in the large L_x limit the results simplify: The scaling, γ, of the hidden-layer size with the number of glomeruli follows the simple relationship shown in Fig. 4*F* (see *SI Appendix*, section 5 and Fig. S7*A* for details). The three lines in Fig. 4*F* correspond to the three phases we saw in Fig. 4 *A* and *B*. In particular, when γ is between 1.6 and 2.0—the range we found from our analysis of learning times (*SI Appendix*, Fig. S4)—the optimal hidden-layer size scales as $L_h^* \propto L_x^{1.3-1.5}$. Thus, our results are robust to the observed scaling of learning time with number of glomeruli.

SGD Learning. So far, we have considered learning by 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,

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

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 (14), the learning rate that enables the fastest decay of the error is (see *SI Appendix*, section 4.2 for details)

$$\eta^* = \frac{2}{L_h}. \qquad [14]$$

For this learning rate, the lifetime average estimation error after N training samples, denoted $\bar{\epsilon}_{est}^{(N)}$, is given approximately by (see *SI Appendix*, section 4.2, especially *SI Appendix*, Eq. S91)

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

where

$$L_1 = \min(L_x, L_h) \qquad [16a]$$

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

with $[\cdot]^+$, the rectified linear function (Fig. 5*A*). The coefficients b_0, b_1, b_2, and b_3 depend on L_h, but not on N, and α (which appears in the exponent of the last term) is the same constant that appeared in Eq. **9**.

The behavior of the estimation error under SGD is different from that 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 rate for the SGD update rule rather than letting it decay, as would be necessary to reduce the estimation error to zero (62). 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 guards against overfitting.

As can be seen in Eq. **15**, the lifetime average estimation error as a function of the number of training samples, N, exhibits three components, all decaying with different timescales (Fig. 5*B*). The timescales of these, L_1, L_2, and L_h, are nondecreasing functions of L_h (Fig. 5*A*). Thus, larger L_h means slower decay with N, as can be seen in Fig. 5*B*. If the coefficients b_q were constant, this would imply that larger L_h would lead to larger lifetime average error. And this is indeed what we see when L_h is larger than about 300 (green line in Fig. 5*C*). However, for smaller L_h, the dependence of the b_q on L_h becomes important, and the lifetime average error decreases with L_h. Because the approximation error decreases monotonically (blue line in Fig. 5*C*), the lifetime average generalization error (red line in Fig. 5*C*) typically has a global minimum at a finite hidden-layer size L_h.

As with MLE learning, under a fixed sample size N the optimal hidden-layer size, L_h^*, shows three different scalings (gray line and dashed lines in Fig. 5*D* and *SI Appendix*, Fig. S2*B*). That 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. 5*D* vs. Fig. 4*A*), although the theoretical prediction at large L_x under SGD does not match quite as well as under MLE. However, using the scaling $N \propto L_x^\gamma$ with the same γ as before ($\gamma = 1.96$), and fitting the coefficient in front of L_x, the experimentally observed scaling law in Fig. 1*A* is again reproduced (Fig. 5*E*). The effect of γ under SGD learning is about the same as it is under MLE (compare *SI Appendix*, Fig. S7*B* to *SI Appendix*, Fig. S7*A*, both of which are very similar to Fig. 4*F*), so the scaling of the optimal hidden-layer size versus that of the input-layer size can be read off Fig. 4*F*. Thus, as with MLE, the 3/2 scaling is relatively robust to decreases in γ, but less robust to increases.

In our model we initialized the readout weights to relatively large values, $\boldsymbol{w}_s^{(0)} \sim N(0, 9.0/L_h)$. If, however, the weights are

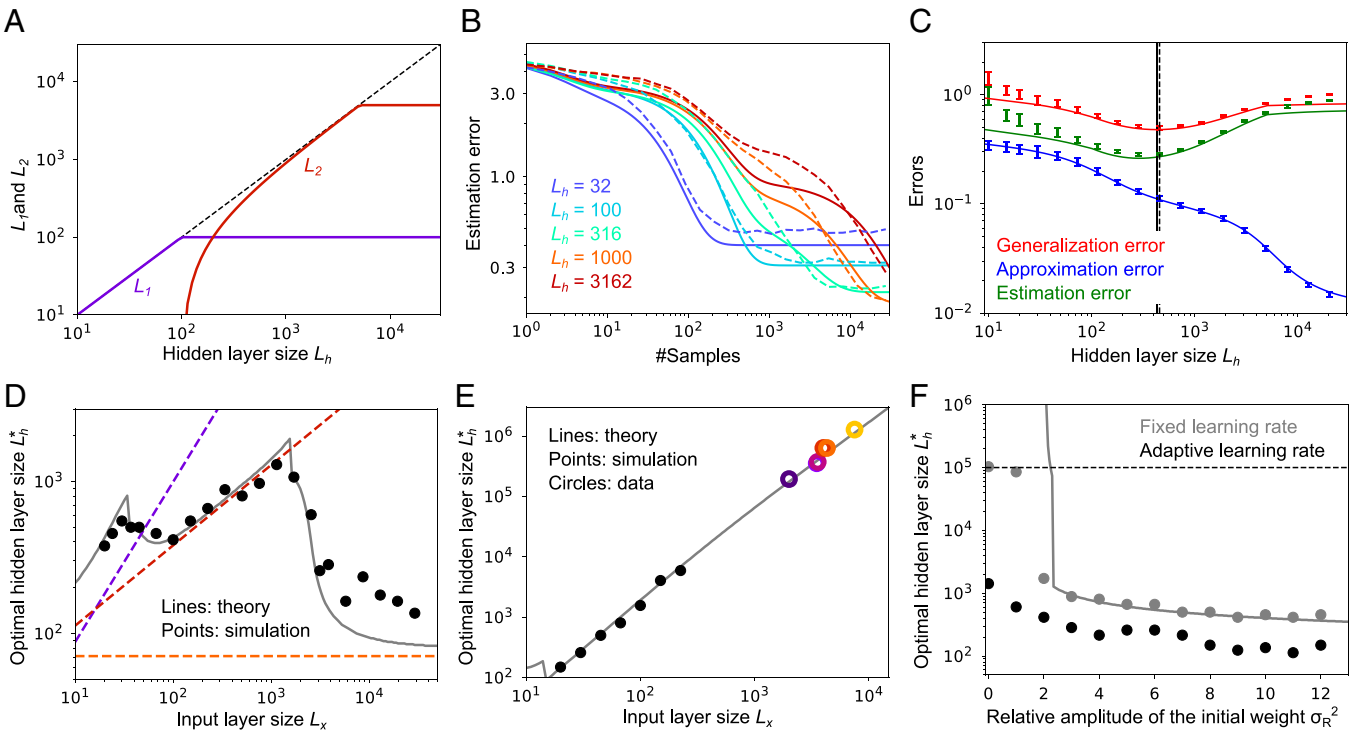

Fig. 5. Model behavior under stochastic gradient descent. (*A*) Hidden-layer size dependence of the decay time constant L_1 and L_2, with $L_x = 100$. (*B*) Dynamics of the estimation error under various hidden-layer sizes, L_h. Dashed lines, simulations; solid lines, theory. (*C*) The lifetime average generalization error, approximation error, and lifetime average estimation error under various hidden-layer sizes, L_h, at $N = 30{,}000$. (*D*) Optimal hidden-layer size, L_h^*, with $N = 30{,}000$. Dashed lines are asymptotic scaling (*SI Appendix*, section 5.2). See *SI Appendix*, Fig. S2*B* for curves with a range of N and σ_t^2. (*E*) Optimal hidden-layer size, L_h^*, with $N = 19L_x^{1.96}$. Gray line is theory; black points are from simulations; colored circles are the experimental data from Fig. 1*A*. 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. The discontinuity around $L_x \sim 10$ is originated from approximations that do not match perfectly around $L_h^* \sim L_x^2/2$ (*SI Appendix*, sections 4.2 and 8). (*F*) Optimal hidden-layer size, L_h^*, for various initial weight amplitudes, σ_R^2, and $N = 30{,}000$. Gray, fixed learning rate; black, adaptive learning rate. Lines are theory and dots are simulations. The initial readout weights were sampled from $w_s^{(0)} \sim N(0, \sigma_R^2/L_h)$. The horizontal dashed line represents the cutoff of L_h^* in the numerical simulations. When $\sigma_R^2 < 2$, under a fixed learning rate, L_h^* is larger than 10^5. In *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 and used a ReLU nonlinearity, and the noise was set to $\sigma_t^2 = 0.1$.

instead initialized to small values, the optimal hidden-layer size L_h^* diverges to infinity (gray line and points in Fig. 5*F*). 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), the cumulative generalization error is optimal at a finite hidden-layer size even when the initial readout weights are zero (black points in Fig. 5*F*). Although the optimal hidden-layer size, L_h^*, goes up as the initial weight amplitude σ_R^2 becomes smaller (Fig. 5*F*), the cumulative error becomes smaller under both fixed and adaptive learning rates (*SI Appendix*, Fig. S9), due to smaller initial error.

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 tend to be similar across species (63), we assume that the genetic budget for the specification of olfactory circuitry is similar among the insects listed in Fig. 1*B*.

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. 6*A*), and the output is

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

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 premotor area (24), where the output $\hat{y}$ could be represented. Because connections between projection neurons and lateral horn neurons tend to be stereotyped (22, 50), we assumed they were tuned over evolutionary timescales. The degree of tuning has a strong effect on the Kenyon cell pathway: If they are well tuned, the number of neurons in the hidden layer of the Kenyon cell pathway can be small; if they are poorly tuned, the number of neurons needs to be large. The lateral horn pathway cannot predict rewards and punishments in all environments, so we assumed this pathway has a limited tuning precision. Consequently, we set the weights $\mathbf{J}_p$ and $\boldsymbol{w}_p$ with low precision (*SI Appendix*, section 6). This is in contrast to $\mathbf{J}_s$, which was initialized randomly and fixed, and $\boldsymbol{w}_s$, which was learned with adaptive SGD. Using L_p to denote the number of lateral horn neurons, under a genetic information budget G, the amount of information encoded in $\mathbf{J}_p$ and $\boldsymbol{w}_p$ is bounded by

$$(L_p L_x + L_p) s_b < G, \qquad [18]$$

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 term is the number needed to specify $\boldsymbol{w}_p$ (see *SI Appendix*, section 6.1 for details).

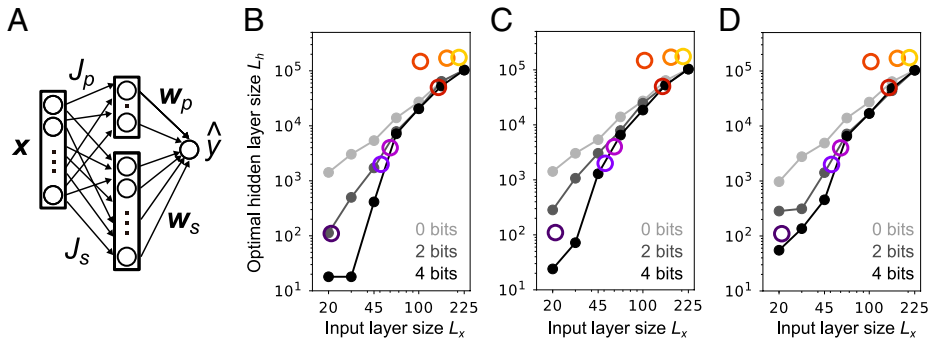

Fig. 6. Olfactory circuit augmented with a genetically specified pathway. (*A*) Schematic of the two-pathway model. For the top part of the circuit, the weights $\mathbf{J}_p$ and w_p are hard wired; for the bottom part, the weights $\mathbf{J}_s$ are randomly connected and w_s are learned with adaptive SGD. (*B–D*) Optimal layer size of the projection neurons-to-Kenyon cells pathway $w_s \cdot g(\mathbf{J}_s x)$ under different model settings. (*B*) Low-bit synapses were achieved by adding Gaussian noise to $\mathbf{J}_p$ and w_p. (*C*) Low-bit synapses were achieved by discretizing $\mathbf{J}_p$ and w_p. (*D*) Low-bit synapses were achieved by adding noise to $\mathbf{J}_p$ and w_p as in *B*, but w_p was additionally learned from training samples using SGD (*SI Appendix*, Eq. S145). In *B–D*, the teacher network had a hidden-layer size of 500 and a ReLU nonlinearity, and we used $\sigma_t^2 = 0.01$ and $N = 10L_x^2$ trials. For $s_b = 2$ bits we used $G = 2{,}000$, while for $s_b = 4$ bits we used $G = 4{,}000$. For $s_b = 0$ bits, we simply removed the hard-wired pathway. The width of the hard-wired intermediate layer, L_p, was found from Eq. **18**: $L_p = G/s_b(L_x + 1)$, rounded up to an integer. See *SI Appendix*, sections 6 and 7.5 for details.

The genetic budget, G, quantifies the accuracy with which the weights $\mathbf{J}_p$ and $\boldsymbol{w}_p$ can be encoded in the genome, with larger G corresponding to higher accuracy. Although we measured G using Shannon information (assuming that synapses are uncorrelated), that may not accurately characterize the minimum genome size required for hard wiring. In fact, the lower bound on the genetic complexity of the weight specification is given by the Kolmogorov complexity, which measures the minimum length of a computer program that generates the network (64). Here we make the assumption that the weights are not massively compressible. In this regime, the Shannon information provides a good estimate of the minimum genome size (*SI Appendix*, section 6). In our simulations we let s_b vary between 0 and 4 bits per synapse, which is broadly consistent in the variability seen in the fly connectome (65). We then keep the genetic budget fixed while we vary L_x; we do that by letting (from Eq. **18**) $L_p = G/s_b(L_x + 1)$.

Under a fixed budget, G, the number of bits per glomerulus, is bounded by G/L_x, suggesting that as the input-layer size, L_x, increases, tuning of $\mathbf{J}_p$ and $\boldsymbol{w}_p$ has 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 (66). 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**, the genetically specified pathway did a good job predicting the valence when L_x was small, but not when it was large. As a result, 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. 6 *B–D*), leading to steeper scaling. In particular, we found that by setting $s_b = 2$, the 7/2 scaling observed among insects is approximately reproduced (dark gray line in Fig. 6*B*). The predicted curve saturates at quadratic scaling around $L_x \approx 150$, resulting in underestimation of the Kenyon cell population in bees and cockroaches. This saturation also indicates that in the mammalian system, for which $L_x \sim 10^3$, the genetically specified pathway is unlikely to play much of a role. This trend was observed under a different implementation of low-bit synapses (Fig. 6*C* and *SI Appendix*, Fig. S10*A*), under sparse implementation of the lateral horn pathway (*SI Appendix*, Fig. S10*B* and section 6.2), and even when $\boldsymbol{w}_p$ was additionally trained with SGD from finely tuned initial weights (Fig. 6*D*). For additional details, see *SI Appendix*, section 6.3.

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 of neurons in the glomerular (i.e., input) layer. We hypothesized that the scarcity of labeled signals (reward and punishment) provides a crucial constraint on the hidden-layer size. This was indeed the case: We showed analytically, and confirmed with simulations, that the optimal hidden-layer size has a strong, nonmonotonic, dependence on the number of labeled signals. Assuming that the number of labeled signals an animal experiences is proportional to its lifetime, and using lifetimes in the range of those reported experimentally, we were able to recover the observed 3/2 scaling (the number of neurons in the hidden layer is proportional to the number of glomeruli to the 3/2 power) observed in mammals. This held under both maximum-likelihood (Fig. 4) and stochastic gradient descent (Fig. 5) learning and was robust to the choice of nonlinearity (Fig. 4 *D* and *E*), activity sparseness (*SI Appendix*, Fig. S6), and the noise level (*SI Appendix*, Fig. S2). Scarcity of labels alone does not, however, explain the 7/2 scaling found in the olfactory circuit of insects. But by considering the fact that genetic information is available for constructing hard-wired olfactory connections, and that it is limited, we recovered the 7/2 scaling law (Fig. 6), without disrupting the 3/2 scaling law in mammals. Note, though, that a certain amount of fine-tuning was required to recover this scaling: The genetic budget (for which we used 2,000 bits) had to be specified to within ∼20% (*SI Appendix*, Fig. S10).

To derive these results, we assumed that the input is white, the connectivity in the hidden layer is fixed and random, and the teacher network is random. If the activity is not whitened at the glomeruli, it is likely that the size of the input layer would be replaced by its effective linear dimensionality. In addition, the actual olfactory environment might have hidden structure that the genetically specified pathway can exploit, even with a low genetic budget. If that is the case, genetic encoding could play a significant role even in the mammalian system. We leave these cases for future work.

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. 4*A* and 5*D*). Second, the optimal population size of the piriform cortex also increases with the number of training samples (Fig. 4*B*). Because species with more glomeruli tend to live longer and experience more samples (*SI Appendix*, Fig. S4), 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 supralinearly on the number of glomeruli (Figs. 4 *C* and *E* and 5*E*). 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 a wide range of scaling is possible (Fig. 4*F* and *SI Appendix*, section 5). The 3/2 scaling we found was because in mammals, lifetime scales approximately quadratically with the number of glomeruli (*SI Appendix*, section 1.2 and Table 2).

Our analysis predicts that $L_h^* \propto \sqrt{N}$ under both MLE and SGD learning. Assuming that this relationship holds at microscopic level as well, we predict that the number of presynaptic connections received by a neuron in the output layer should scale with the square root of the frequency of feedback signals it receives. It might be possible to test this prediction in the mushroom body output neurons of flies, where compartmentalized units receive diverse neuromodulatory inputs (24), and its detailed connectivity structure is known (28). Although we also need to estimate the frequency of the feedback signal at each compartment in the natural environment, in the near future it should be technically feasible to test this prediction.

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, 38, 39, 67) and engineering (68, 69) perspectives. Under batch learning, the upper bound on the approximation error for this network structure is known for a large class of the target functions (59, 70), but these bounds are often too loose to be practical. Here, we instead focused on the average approximation error (*SI Appendix*, section 3). This allowed us to derive, analytically, accurate estimates of the optimal hidden-layer size. We found a nontrivial three-phase structure, which has not been reported before in the context of model selection (but see ref. 71). The behavior of the estimation error is also well characterized in the large sample size limit ($N \to \infty$ while $L_x, L_h < \infty$) (13, 72), but this limit is not a good approximation of an overparameterized neural network. On the other hand, the characteristics of the error in the large parameter limit (number of synapses proportional to N as $N \to \infty$) remain mainly elusive, except for linear regression (58) (*SI Appendix*, section 4.1). Similarly, model selection in neural networks has been studied mostly in the large sample size limit (7, 73). The upper bound on the network size was also studied from Vapnik–Chervonenkis theory (5) and the minimum description length principle (6).

Learning dynamics in neural networks under SGD has also been widely studied (14, 52, 74). In particular, recent results suggest that overparameterization of a neural network does not harm the generalization error under both full-batch and stochastic gradient descent learning (71, 74–76). 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, overparameterization does tend to harm performance, because learning becomes slower (Fig. 5*B*), even under an adaptive learning rate (Fig. 5*F*).

Slow learning is consistent with previous observations that deep reinforcement learning, in which the model often needs to be trained online, requires a large number of iterations for successful learning (77). Neural architecture optimization for the minimization of cumulative loss may help build efficient and generalizable deep reinforcement learning. However, we also found that if the learning rate is fixed and the initial weights are set to very small values, having infinitely many neurons in the hidden layer minimizes the cumulative error (Fig. 5*F*), suggesting that overparameterization is not always harmful, even when the cumulative error is the relevant cost function. Our analysis of neural architecture selection may provide insight into scaling laws observed in artificial neural networks, although those models face different constraints than their biological counterparts (49).

Scaling laws are also observed in other regions of the brain. For instance, the number of neurons in the primary visual cortex scales with the 3/2 power relative to the population size of the lateral geniculate nucleus (78), and the number of neurons in the cerebral cortex is linear in the total number of neurons in the cerebellum (79). 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.

Data Availability. The source codes of the simulations and the data analysis are deposited in GitHub (https://github.com/nhiratani/olfactory_design) (80).

ACKNOWLEDGMENTS. This work was supported by the Gatsby Charitable Foundation and the Wellcome Trust (110114/Z/15/Z). N.H. was partially supported by the Swartz Foundation.

1. A. Mathis, A. V. Herz, M. Stemmler, Optimal population codes for space: Grid cells outperform place cells. *Neural Comput.* **24**, 2280–2317 (2012).
2. J. Gjorgjieva, H. Sompolinsky, M. Meister, Benefits of pathway splitting in sensory coding. *J. Neurosci.* **34**, 12127–12144 (2014).
3. A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, L. F. Abbott, Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164.e7 (2017).
4. H. Akaike, A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
5. E. B. Baum, D. Haussler, "What size net gives valid generalization?" in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed. (NIPS, 1988), vol. **1**, pp. 81–90.
6. A. R. Barron, Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **14**, 115–133 (1994).
7. N. Murata, S. Yoshizawa, S. Amari, Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Trans. Neural Netw.* **5**, 865–872 (1994).
8. M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of Machine Learning* (MIT Press, 2018).
9. T. Elsken, J. H. Metzen, F. Hutter, Neural architecture search: A survey. *J. Mach. Learn. Res.* **20**, 1997–2017 (2019).
10. R. L. Davis, Olfactory learning. *Neuron* **44**, 31–48 (2004).
11. B. W. Ache, J. M. Young, Olfaction: Diverse species, conserved principles. *Neuron* **48**, 417–430 (2005).
12. R. I. Wilson, Z. F. Mainen, Early events in olfactory processing. *Annu. Rev. Neurosci.* **29**, 163–201 (2006).
13. S. I. Amari, N. Murata, Statistical theory of learning curves under entropic loss criterion. *Neural Comput.* **5**, 140–153 (1993).
14. J. Werfel, X. Xie, H. S. Seung, "Learning curves for stochastic gradient descent in linear feedforward networks" in *Advances in Neural Information Processing Systems*, S. Thrun *et al.*, Eds. (NIPS, 2003), vol. **16**, pp. 1197–1204.
15. O. Gschwend *et al.*, Neuronal pattern separation in the olfactory bulb improves odor discrimination learning. *Nat. Neurosci.* **18**, 1474–1482 (2015).
16. N. Hiratani, P. E. Latham, Rapid Bayesian learning in the mammalian olfactory system. *Nat. Commun.* **11**, 3845 (2020).
17. G. H. Otazu, H. Chae, M. B. Davis, D. F. Albeanu, Cortical feedback decorrelates olfactory bulb output in awake mice. *Neuron* **86**, 1461–1477 (2015).
18. A. Grabska-Barwińska *et al.*, A probabilistic approach to demixing odors. *Nat. Neurosci.* **20**, 98–106 (2017).
19. A. K. Dhawale, A. Hagiwara, U. S. Bhalla, V. N. Murthy, D. F. Albeanu, Non-redundant odor coding by sister mitral cells revealed by light addressable glomeruli in the mouse. *Nat. Neurosci.* **13**, 1404–1412 (2010).
20. S. Srinivasan, C. F. Stevens, Scaling principles of distributed circuits. *Curr. Biol.* **29**, 2533–2540.e7 (2019).
21. J. P. Martin *et al.*, The neurobiology of insect olfaction: Sensory processing in a comparative context. *Prog. Neurobiol.* **95**, 427–447 (2011).
22. M. Fişek, R. I. Wilson, Stereotyped connectivity and computations in higher-order olfactory neurons. *Nat. Neurosci.* **17**, 280–288 (2014).

23. S. J. Caron, V. Ruta, L. F. Abbott, R. Axel, Random convergence of olfactory inputs in the Drosophila mushroom body. *Nature* **497**, 113–117 (2013).
24. Y. Aso *et al.*, The neuronal architecture of the mushroom body provides a logic for associative learning. *eLife* **3**, e04577 (2014).
25. R. Huerta, T. Nowotny, M. García-Sanchez, H. D. Abarbanel, M. I. Rabinovich, Learning classification in the olfactory system of insects. *Neural Comput.* **16**, 1601–1640 (2004).
26. A. Ramaekers *et al.*, Glomerular maps without cellular redundancy at successive levels of the Drosophila larval olfactory circuit. *Curr. Biol.* **15**, 982–992 (2005).
27. L. M. Masuda-Nakagawa, N. Gendre, C. J. O'Kane, R. F. Stocker, Localized olfactory representation in mushroom bodies of *Drosophila* larvae. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10314–10319 (2009).
28. K. Eichler *et al.*, The complete connectome of a learning and memory centre in an insect brain. *Nature* **548**, 175–182 (2017).
29. S. Anton, B. S. Hansson, Central processing of sex pheromone, host odour, and oviposition deterrent information by interneurons in the antennal lobe of female Spodoptera littoralis (Lepidoptera: Noctuidae). *J. Comp. Neurol.* **350**, 199–214 (1994).
30. M. Sjöholm, I. Sinakevitch, R. Ignell, N. J. Strausfeld, B. S. Hansson, Organization of Kenyon cells in subdivisions of the mushroom bodies of a lepidopteran insect. *J. Comp. Neurol.* **491**, 290–304 (2005).
31. Z. Wang *et al.*, Identification and functional analysis of olfactory receptor family reveal unusual characteristics of the olfactory system in the migratory locust. *Cell. Mol. Life Sci.* **72**, 4429–4443 (2015).
32. B. Leitch, G. Laurent, GABAergic synapses in the antennal lobe and mushroom body of the locust olfactory system. *J. Comp. Neurol.* **372**, 487–514 (1996).
33. G. Arnold, C. Masson, S. Budharugsa, Comparative study of the antennal lobes and their afferent pathway in the worker bee and the drone (Apis mellifera). *Cell Tissue Res.* **242**, 593–605 (1985).
34. H. Watanabe, H. Nishino, M. Nishikawa, M. Mizunami, F. Yokohari, Complete mapping of glomeruli based on sensory nerve branching pattern in the primary olfactory center of the cockroach *Periplaneta americana*. *J. Comp. Neurol.* **518**, 3907–3930 (2010).
35. S. M. Farris, N. J. Strausfeld, Development of laminar organization in the mushroom bodies of the cockroach: Kenyon cell proliferation, outgrowth, and maturation. *J. Comp. Neurol.* **439**, 331–351 (2001).
36. K. D. Ernst, J. Boeckh, V. Boeckh, A neuroanatomical study on the organization of the central antennal pathways in insects. *Cell Tissue Res.* **176**, 285–306 (1977).
37. E. A. Capaldi, G. E. Robinson, S. E. Fahrback, Neuroethology of spatial learning: The birds and the bees. *Annu. Rev. Psychol.* **50**, 651–682 (1999).
38. O. Barak, M. Rigotti, S. Fusi, The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *J. Neurosci.* **33**, 3844–3856 (2013).
39. B. Babadi, H. Sompolinsky, Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).
40. A. M. Hermundstad, K. S. Brown, D. S. Bassett, J. M. Carlson, Learning, memory, and the role of neural network architecture. *PLOS Comput. Biol.* **7**, e1002063 (2011).
41. J. Kadmon, H. Sompolinsky, "Optimal architectures in a solvable model of deep networks" in *Advances in Neural Information Processing Systems*, D. Lee *et al.*, Eds. (NIPS, 2016), vol. **29**, pp. 4781–4789.
42. L. C. Aiello, P. Wheeler, The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Curr. Anthropol.* **36**, 199–221 (1995).
43. A. Navarrete, C. P. van Schaik, K. Isler, Energetics and the evolution of human brain size. *Nature* **480**, 91–93 (2011).
44. D. V. Raman, A. P. Rotondo, T. O'Leary, Fundamental bounds on learning performance in neural circuits. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10537–10546 (2019).
45. L. Yu, C. Zhang, L. Liu, Y. Yu, Energy-efficient population coding constrains network size of a neuronal array system. *Sci. Rep.* **6**, 19369 (2016).
46. J. Zylberberg, A. Pouget, P. E. Latham, E. Shea-Brown, Robust information propagation through noisy neural circuits. *PLOS Comput. Biol.* **13**, e1005497 (2017).
47. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
48. N. E. Raine, L. Chittka, The correlation of learning speed and natural foraging success in bumble-bees. *Proc. Biol. Sci.* **275**, 803–808 (2008).
49. Y. Bahri, E. Dyer, J. Kaplan, J. Lee, U. Sharma, Explaining neural scaling laws. arXiv [Preprint] (2021). https://arxiv.org/abs/2102.06701 (Accessed 18 February 2022).
50. S. Frechter *et al.*, Functional and anatomical specificity in a higher olfactory centre. *eLife* **8**, e44590 (2019).
51. H. Sompolinsky, N. Tishby, H. S. Seung, Learning from examples in large neural networks. *Phys. Rev. Lett.* **65**, 1683–1686 (1990).
52. A. M. Saxe, J. L. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv [Preprint] (2013). https://arxiv.org/abs/1312.6120 (Accessed 18 February 2022).
53. L. Badel, K. Ohta, Y. Tsuchimoto, H. Kazama, Decoding of context-dependent olfactory behavior in Drosophila. *Neuron* **91**, 155–167 (2016).
54. T. Hige, Y. Aso, M. N. Modi, G. M. Rubin, G. C. Turner, Heterosynaptic plasticity underlies aversive olfactory learning in Drosophila. *Neuron* **88**, 985–998 (2015).
55. J. Chapuis, D. A. Wilson, Bidirectional plasticity of cortical pattern recognition and behavioral sensory acuity. *Nat. Neurosci.* **15**, 155–161 (2011).
56. D. J. Millman, V. N. Murthy, Rapid learning of odor–value association in the olfactory striatum. *J. Neurosci.* **40**, 4335–4347 (2020).
57. P. Y. Wang *et al.*, Transient and persistent representations of odor value in prefrontal cortex. *Neuron* **108**, 209–224.e6 (2020).
58. M. Advani, S. Ganguli, Statistical mechanics of optimal convex inference in high dimensions. *Phys. Rev. X* **6**, 031034 (2016).
59. A. Rahimi, B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning" in *Advances in Neural Information Processing Systems*, D. Koller *et al.*, Eds. (NIPS, 2008), vol. **21**, pp. 1313–1320.
60. R. Tacutu *et al.*, Human ageing genomic resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41**, D1027–D1033 (2013).
61. S. L. Pashkovski *et al.*, Structure and flexibility in cortical representations of odour space. *Nature* **583**, 253–258 (2020).
62. L. Bottou, "Online learning and stochastic approximations" in *On-Line Learning in Neural Networks*, D. Saad, Ed. (Cambridge University Press, Cambridge, UK, 1998), pp. 142–177.
63. L. Pray, Eukaryotic genome complexity. *Nature Education* **1**, 96 (2008).
64. P. Grunwald, P. Vitányi, Shannon information and kolmogorov complexity. arXiv [Preprint] (2004). https://arxiv.org/abs/cs/0410002 (Accessed 18 February 2022).
65. S. Y. Takemura *et al.*, Synaptic circuits and their variations within different columns in the visual system of *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13711–13716 (2015).
66. K. K. Ishii *et al.*, A labeled-line neural circuit for pheromone-mediated sexual behaviors in mice. *Neuron* **95**, 123–137.e8 (2017).
67. S. Ganguli, H. Sompolinsky, Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* **35**, 485–508 (2012).
68. G. B. Huang, Q. Y. Zhu, C. K. Siew, Extreme learning machine: Theory and applications. *Neurocomputing* **70**, 489–501 (2006).
69. S. Dasgupta, C. F. Stevens, S. Navlakha, A neural algorithm for a fundamental computing problem. *Science* **358**, 793–796 (2017).
70. X. Liu, S. Lin, J. Fang, Z. Xu, Is extreme learning machine feasible? A theoretical assessment (part I). *IEEE Trans. Neural Netw. Learn. Syst.* **26**, 7–20 (2015).
71. S. d'Ascoli, L. Sagun, G. Biroli, "Triple descent and the two kinds of overfitting: Where & why do they appear?" in *Advances in Neural Information Processing Systems*, H. Larochelle *et al.*, Eds. (NeurIPS, 2020), vol. **33**, pp. 3058–3069.
72. A. W. Van der Vaart, *Asymptotic Statistics* (Cambridge University Press, 2000), vol. **3**.
73. N. Barkai, H. S. Seung, H. Sompolinsky, Scaling laws in learning of classification tasks. *Phys. Rev. Lett.* **70**, 3167–3170 (1993).
74. S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, L. Zdeborová, "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (NeurIPS, 2019), vol. **32**, pp. 6981–6991.
75. M. S. Advani, A. M. Saxe, H. Sompolinsky, High-dimensional dynamics of generalization error in neural networks. *Neural Netw.* **132**, 428–446 (2020).
76. M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15849–15854 (2019).
77. M. Botvinick *et al.*, Reinforcement learning, fast and slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
78. C. F. Stevens, An evolutionary scaling law for the primate visual system and its basis in cortical function. *Nature* **411**, 193–195 (2001).
79. S. Herculano-Houzel, Coordinated scaling of cortical and cerebellar numbers of neurons. *Front. Neuroanat.* **4**, 12 (2010).
80. N. Hiratani, Simulation and analysis code. GitHub. https://github.com/nhiratani/olfactory_design. Deposited 9 July 2021.