Statistical properties of the optimal sensitivity matrix for compressed sensing with nonlinear sensors

Shanshan Qin[1],* Qianyi Li[2],* Chao Tang[1,3,4],† and Yuhai Tu[5]‡
[1] *Center for Quantitative Biology, Peking University,Beijing,100871,China*
[2] *Yuanpei College, Peking University, Beijing, 100871, China*
[3] *School of Physics, Peking University, Beijing 100871, China*
[4] *Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China*
[5] *IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA*
(Dated: November 6, 2018)

There are numerous different odorant molecules in nature. Typical odors are sparse mixtures of a few types of odorant molecules each with a wide range of concentrations. However, there are only a relatively small number of olfactory receptor neurons (ORNs), which respond to odorant concentration nonlinearly (sigmoidal) with a finite sensitivity range. Thus, how to encode a large number of sparse odor mixtures with a relatively small number of nonlinear ORNs – the nonlinear compressed sensing problem – remains a puzzle. Here, by using an information theory approach, we study the optimal coding strategies that enable nonlinear ORNs to best represent olfactory information (both the odorants' identities and their concentrations) in sparse odor mixtures. Our extensive numerical simulations and analytical analysis show that the optimal odor-receptor sensitivity matrix is sparse and the nonzero sensitivities follow roughly a log-normal distribution(matching the statistics of the odorants), both of which are consistent with existing experiments. For ORNs with a finite basal activity, our study shows that co-existence of both odor-evoked excitation and inhibition increases coding capacity, which provides a plausible explanation for such co-existence observed in the fly olfactory system. Furthermore, we show that coding the inputs with the optimal sensitivity matrix can enhance accuracy of the downstream decoding and learning tasks. These general statistical properties of the optimal sensitivity matrix for nonlinear compressed sensing may shed light on understanding the peripheral olfactory sensory system and improving performance of artificial neural networks.

I. INTRODUCTION

Animals rely on their olfactory systems to detect, discriminate and interpret external odor stimuli to guide their behavior. Natural odors often consist of many odorant molecule mixtures whose concentration vary within several orders of magnitude [1–3]. Yet, animals can distinguish a large number of odorants and their mixtures using relatively small number of olfactory receptors (ORs)[4, 5]. For example, humans have only$\sim$ 300 ORs [6, 7], and the often cited number of odors that can be distinguished is $\sim$ 10000[8], the real number might be even larger [9] (see also [10] and [11]). Humans can also distinguish odor mixture with up to 30 different compounds[12]. In comparison, the highly olfactory lifestyle and exquisite olfactory learning ability of fly is afforded by only $\sim$ 60 ORs [4, 13]. The olfactory system achieves such remarkable ability through combinatorial code, where each odorant is sensed by multiple receptors and each receptor can be activated by many odorants [14–16]. In both insects and mammals, odorants bind to the receptors in the cilia or dendrites of olfactory receptor neurons (ORNs), each of which expresses only one type of receptor. ORNs that express the same receptors then converge to the same glomerulus in olfactory bulb (mammals) or antennal lobe (insects), whose activity patterns carry the information about external odor stimuli [13, 17–19]. The key question here is how ORNs best represent external olfactory information that can be be interpreted by the brain to guide an animal's behavior [4, 13, 20].

It has long been hypothesized that the input-output response functions of sensory neurons are "selected" by statistics of the stimuli in an organism's natural environment to optimally transmit information of its environment [21–23]. For instance, the contrast-response function of interneurons in the fly's compound eye can be well approximated by the cumulative probability distribution of contrast in natural environment [22]. The receptive fields of neurons in early visual pathway are thought to exploit the statistics of natural scenes [24–27]. Similar result has also been observed in auditory system [28, 29]. In all these cases, to achieve maximum information transmission, an "ideal" neuron should transform the input distribution into a uniform output distribution as suggested first by Laughlin [22, 30] and a population of neurons should decorrelate their responses [27, 31, 32].

Unlike visual or auditory stimuli, which can be characterized by a single quantity such as wavelength or frequency, there are a huge number of odorants each with its own unique molecular structure and different physiochemical properties [33, 34]. The high dimensionality of the odor space poses a severe challenge for coding olfactory signals. Fortunately, typical olfactory stimuli are sparse with only a few types of odorant molecules in an

* These two authors contributed equally
† email address: tangc@pku.edu.cn; Corresponding author
‡ email address: yuhai@us.ibm.com; Corresponding author

odor mixture [1, 2]. The sparsity of the input signal immediately reminds us of the well-known compressed sensing (CS) theory, which states that sparse high dimensional signals can be encoded by a small number of sensors (measurements) through random projections that permit faithful reconstruction [35–37]. However, conventional CS theory assumes the sensors to have a linear response function with essentially an infinite dynamical range [38]. In contrast, ORN response is highly nonlinear [39, 40], with a dynamical range less than 2 orders of magnitude which is far less than the typical concentration range of odorants [40]. Indeed, though some insights have been gained by applying the CS theory to olfaction with either a linear response function [41] or an oversimplified binary ORNs model [42], the question of how ORNs with realistic nonlinearity encode the sparse high dimensional odor information remains unclear.

In this paper, we study how ORNs maximize information transmission from the external olfactory environment to the ORN level response by considering the nonlinearity of the input-output relationship and odor sparsity. We found that the optimal ORN sensitivity matrix is sparse, i.e. each receptor only responds to a fraction of the odorants in its environment, and the nonzero sensitivity values follow a log-normal distribution. Though the value of the optimal sparsity depends on the odor statistics, the sparsity itself is a robust (universal) feature of the optimal sensitivity matrix. We developed a simple mean-field theory to show that the sensitivity matrix sparsity is caused by the competition (trade-off) between enhancing odor detection and avoiding odor-odor interference. Next, we demonstrated the potential advantages of maximum entropy (information) coding in two 'decoding' (learning) tasks: reconstruction and classification of odor stimuli. Finally, we applied our approach to study neurons with a non-zero basal activity, where we found that co-existence of both odor-evoked inhibition and activation further enhances the coding capacity.

II. RESULTS

An odor mixture can be represented as a vector $\boldsymbol{c} = (c_1, \ldots, c_N)$, where c_j is the concentration of odorant (ligand) $j(= 1, 2, ..., N)$ and N is the number of all possible odorants in the environment. A typical odor mixture is sparse with only $n(\ll N)$ odorant molecules that have non-zero concentrations. As illustrated in Fig. 1A (dotted box), the odor mixture signal $\boldsymbol{c}$ is sensed by M sensors. For simplicity, we used a simple competitive binding model [43, 44] (and see Supporting Information (SI) Appendix), in which the normalized response of ORN i $(= 1, 2, \cdots, M)$ to odor $\boldsymbol{c}$ can be described by a nonlinear function

$$r_i = \frac{\sum_j^N W_{ij} c_j}{1 + \sum_j^N W_{ij} c_j} + \eta_i, \tag{1}$$

where the sensitivity of i-th sensor (ORN) to j-th odorant is given by W_{ij}, which is defined as the inverse of the dissociation constant K_{ij}^m. η_i represents the noise term that has various biochemical origins. For convenience, we assume η_i is a Gaussian noise with zero mean and standard deviation σ_0. Other forms of the nonlinear response function and noise can be used without affecting the general conclusions of our paper.

As shown in Fig. 1 B, the input-output response curve is highly nonlinear resulting in a finite response range for each sensor, which is less than the range of concentrations for a typical odorant molecule. Therefore, an odorant molecule needs to interact with multiple sensors with different sensitivities to encode its full range of concentrations. On the other hand, given the fact that $M < N$, each sensor has to sense multiple odorant molecules. These two considerations lead to the many-to-many odor-sensor interaction network characterized by the sensitivity matrix W_{ij}.

Eq. 1 maps the external odor stimulus $\boldsymbol{c}$ to the internal neuronal activity $\boldsymbol{r} = (r_1, r_2, \cdots, r_M)$. The downstream olfactory circuits then use this response pattern to infer $\boldsymbol{c}$ and evaluate odor quality and quantity to guide the animal's behavior. The encoding quality of sensory periphery sets the upper limit of how well the brain can decode the odor information (Fig. 1 C)[45]. Here, we focus on how the peripheral nonlinear ORNs, characterized by the sensitivity matrix W_{ij}, best represent external odor environment.

A given odor environment can be generally described by a probability distribution $P_{\text{env}}(\boldsymbol{c})$. To convey maximum information about external odor stimuli in their response patterns to the brain, ORs/ORNs can adjust their sensitivity matrix W_{ij} to "match" odor statistics $P_{\text{env}}(\boldsymbol{c})$. Without any assumption on what information the brain may need, the mutual information $I(\boldsymbol{c}, \boldsymbol{r})$ between stimuli and response pattern of ORNs sets the limit on how much odor information can be received by the peripheral ORNs and thus serves as a good "target function" to be maximized [23, 42, 46–48]. I is defined as

$$\begin{aligned} I(\boldsymbol{c}, \boldsymbol{r}) &= H(\boldsymbol{r}) - H(\boldsymbol{r}|\boldsymbol{c}) \\ &= -\int P(\boldsymbol{c}, \boldsymbol{r}) \log \frac{P(\boldsymbol{c}, \boldsymbol{r})}{P_r(\boldsymbol{r}) P_{\text{env}}(\boldsymbol{c})} d\boldsymbol{c} d\boldsymbol{r}, \end{aligned} \tag{2}$$

where $H(\boldsymbol{r})$ and $H(\boldsymbol{r}|\boldsymbol{c})$ are the entropy of output distribution $P_r(\boldsymbol{r})$ and conditional distribution $P(\boldsymbol{r}|\boldsymbol{c})$.

In the limit of relatively small noise $\sigma_0 \ll 1$ and $M \geq n$, Eq. 2 can be approximated as

$$\begin{aligned} I \approx &-\int P_{\text{env}}(\boldsymbol{c}) \ln P_{\text{env}}(\boldsymbol{c}) d\boldsymbol{c} + (M - n) \ln(\sqrt{2\pi}\sigma_0) \\ &+ \frac{1}{2} \int P_{\text{env}}(\boldsymbol{c}) \ln[\det(\chi)] \text{dc} \end{aligned} \tag{3}$$

where χ is an $N \times N$ Jacobian matrix with its matrix elements given as: $\chi_{kl} = \sum_{j=1}^M \frac{\partial r_j}{\partial C_k} \frac{\partial r_j}{\partial C_l}$. The first two terms of the left side of Eq. S10 are irrelevant to the

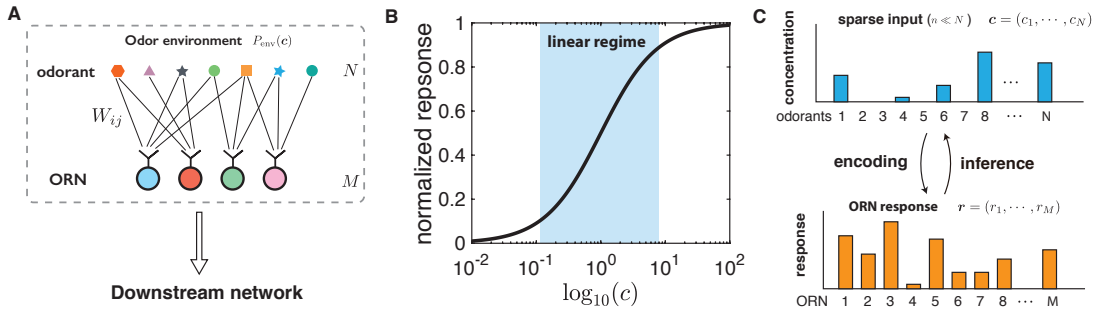

FIG. 1. Schematics of peripheral odor coding. (A) Illustration of interaction between odorants and odor receptors. Sparse odor mixture in a high dimensional space is mapped to a relatively low dimensional ORN-space, the downstream olfactory system then infers the external odor stimuli to guide animal's behavior. The dashed box highlights the peripheral coding problem. (B) Typical ORN-odorant dose response curve according to Eq. 1 with $W = 1$. The range of linear response is highlighted by the shaded area. (C) Odor coding and decoding. High dimensional sparse odor stimulus $\boldsymbol{c}$ is sensed by a small number of ORNs, whose response pattern $\boldsymbol{r}$ is relayed to the higher region (downstream network) of the brain where odor information is inferred and learning occurs.

maximization process and have clear physical meanings: the first one is just the entropy of odor distribution and the second one is due to dimension reduction. Only the third term, denoted as ΔH, depends on the sensitivity matrix. Hence we will focus on how W maximize ΔH throughout this study.

The mutual information as given in Eq. S10 quantifies the odor encoding capacity, which can only be computed analytically for very simple cases. For more general cases, we used the covariance matrix adaptation - evolution strategy (CMA-ES) algorithm to find the optimal sensitivity matrix[49, 50] (see SI Appendix for technical details).

A. Statistical properties of the optimal sensitivity matrix

The input structure is "sparse" and only a small number of odorants appear in each mixture. The number of ligands in the mixture n, i.e., the "sparsity" of input, reflects the complexity of odor composition. Odor concentration varies widely in natural environment [51]. To capture this property, we studied the case where the ligand concentrations in an odor mixture follows an i.i.d. log-normal distribution, i.e., $\ln(\boldsymbol{c}) \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 I_{n\times n})$, where $I_{n\times n}$ is a $n \times n$ identity matrix, other broad distributions such as power law distributions are also studied without changing the general conclusions. For simplicity, we consider the case where odorants appear independently in the mixture. This is a simplification of real odor environment, more realistic consideration such as correlation among odorants will be discussed later in the Discussion section.

For high dimensional sparse odor environment characterized by (N, n and σ_c), and given the number of nonlinear sensors M, we can compute and optimize the input-output mutual information $I(W|N, n, \sigma_c)$ with respect to the sensitivity matrix W of size $(M \times N)$. As shown in Fig. 2, the optimal sensitivity matrix W is "sparse", by which we mean that only a fraction ($\rho_w < 1$) of its elements have a non-zero value (sensitive, shown as the colored elements in Fig. 2 A), and the rest are insensitive (the black elements in Fig. 2 A), with an essentially zero value of W_{ij}. This can be seen more clearly from the histogram of $\ln(W_{ij})$ as shown in Fig. 2 B. It is clear that elements in the optimal sensitivity matrix fall into two distinctive populations: the insensitive population that has practically zero sensitivity (note the log-scale used in Fig. 2B), and a sensitive population with a finite sensitivity. These nonzero (NZ) sensitivities follow a log-normal distribution with mean value μ_w and standard deviation σ_w, i.e., $\ln(W_{NZ}) \sim \mathcal{N}(\mu_w, \sigma_w)$. The existence of a finite fraction of insensitive elements in the optimal sensitivity matrix is one of the main findings of our study.

Besides the distribution of the optimal sensitivity matrix elements, we also tested if there is any higher order statistical structure in W. We calculated the row (sensor)-wise and column (odorant)-wise rank-order correlation coefficients and compared them with those from the same matrix but with its elements shuffled randomly. As shown in Fig. 2C, the distributions of the sensor-wise correlation coefficients for the optimal matrix and the randomly shuffled matrix are all centered around a value close to zero. However, the distributions for the optimal sensitivity matrix is more narrowly distributed with a smaller variance in comparison with those from the randomly shuffled matrices, indicating a weak but significant statistical structure. Odorant-wise correlation coefficients also show similar result (see SI Appendix, fig. S1). This is further verified by several other statistical tests, such as the comparison of eigenvalues of correlation matrix, and the eigenvalues of inner product matrix (see SI Appendix, Fig.S1, and Fig. S2 A-C). Our results suggest that the optimal matrix achieves a higher level of orthogonality and thus independence among different sensors and among different odorants than those from the random matrices. This subtle difference en-

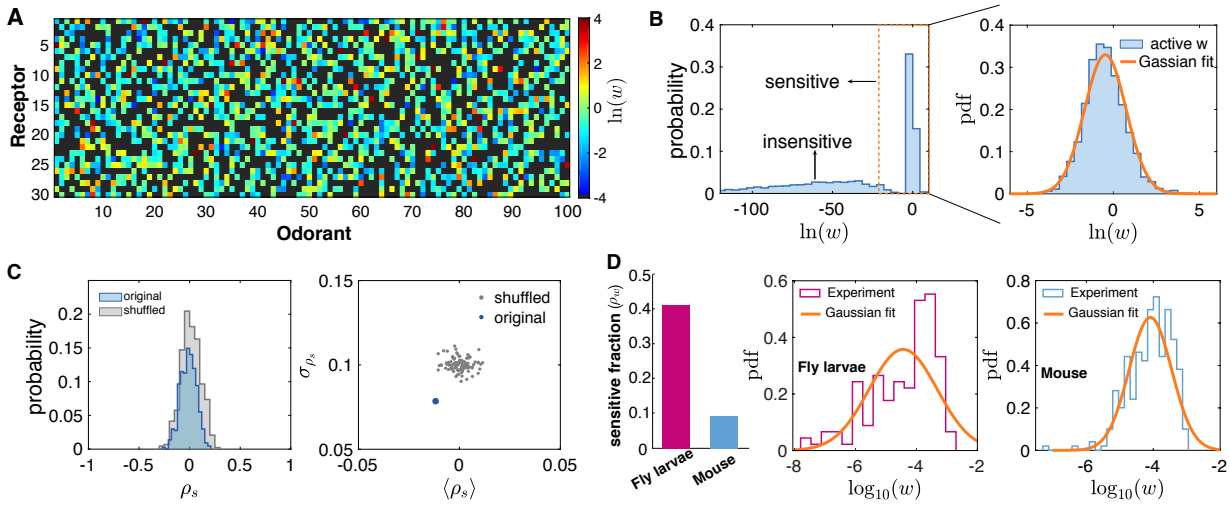

FIG. 2. The distribution and correlation of the optimal sensitivity matrix elements. (A) Heatmap of a typical optimal sensitivity matrix. Color indicates the value of $\ln(W_{ij})$, black indicates the "inactive" or negligible interactions. (B) Histogram of all the W_{ij} values. It shows a bimodal distribution: an insensitive part with near zero W_{ij} and a sensitive part with nonzero W_{ij}. The sensitive part is log-normal distributed. (C) Left: histogram of row-wise Spearman rank correlation coefficients ρ_s of the optimal matrix (blue) and a randomly shuffled matrix (gray). Right: comparison of the average and standard deviation of ρ_s with the grey points from 100 randomly shuffled matrices and blue dot representing the original optimal sensitivity matrix. (D) Left: the fraction of sensitive odorant-receptor interaction estimated in experiments for fly larvae [40] and mouse [52]. Right: the histogram of sensitive W_{ij} for fly larvae and mouse. Dashed lines are Gaussian fit. Parameters for A-C are $N = 100$, $M = 30$, $n = 2$, $\sigma_c = 2$, $\mu = 0$.

ables ORN response pattern with the optimal W to have a larger entropy than those from the shuffled matrices (SI Appendix, fig. S2 D). A close look at the distribution of output $P_r(\boldsymbol{r})$ shows that the optimal W_{ij} allows each ORN to use its dynamical range nearly uniformly for different odor mixtures (see SI Appendix, Fig. S3).

B. Sparsity of the optimal W and distribution of the non-zero elements depend on odor statistics

We then asked how the distribution of the optimal sensitivity matrix elements, which is described by the sparsity (ρ_w), the mean (μ_w) and standard deviation (σ_w) of the non-zero elements depends on the input statistics. We first fixed the dimension of input (N) and the number of types of ORNs (M), as well as the sparsity of odor mixtures (n), but systematically changed the concentration width σ_c. As the odor concentration becomes broader with increasing σ_c, ρ_w –the sparsity of W– increases, thus more receptors become sensitive to odorants in the mixture, i.e., more receptors are required to tile the broad range of input concentrations. For the distribution of the sensitive elements, their averaged sensitivity (μ_w) decreases, while the width of the sensitive distribution (σ_w) increases (Fig. 3 A). We then fixed N,M and σ_c, but changed the number of ligands n in odor mixture, the sparsity of W (ρ_w) goes down as odor mixture sparsity increases (Fig. 3B). Due to the limited dynamical range of receptors, to match the input statistics, a receptor tends to response to fewer ligands and use more moderate interactions to avoid saturation. Thus, the average value of sensitive elements (μ_w) decreases and the distribution of sensitive elements remain roughly unchanged with increasing n (Fig. 3 B). Finally, ρ_w decreases when N increases (with all the other parameters fixed), indicating that as the mapping from odor space to ORN space becomes more "compressed", the optimal strategy is to assign each receptor to respond to a smaller fraction of odorants with a roughly constant average sensitivity (μ_w) that matches the input concentration, and the distribution of active elements becomes slightly narrower (Fig. 3C).

To further test whether sparsity is a general feature in the optimal sensitivity matrix, we changed the odor concentration distribution to be power-law distributions, $P_{\text{env}}(c) \sim c^{-\beta}$, with different values of $\beta > 1$. For all the values of β we tested, there is always a finite sparsity $\rho_w < 1$ in the optimal sensitivity matrix and ρ_w decreases as odor concentration distribution becomes wider as β decreases. ρ_w also decreases as n increases. But the exact distribution of the sensitive elements differs slightly from the case with log-normal distribution (see SI Appendix, Fig. S4).

Taken together, these results suggest that the sparse optimal sensitivity matrix is a robust feature that when nonlinear receptors with a limited dynamical range try to best represent sparse high dimensional odor stimuli. The sparsity of sensitivity matrix of our model is also supported by a recent experimental study of *Drosophila* (fly) larvae [40], which indeed showed a large fraction

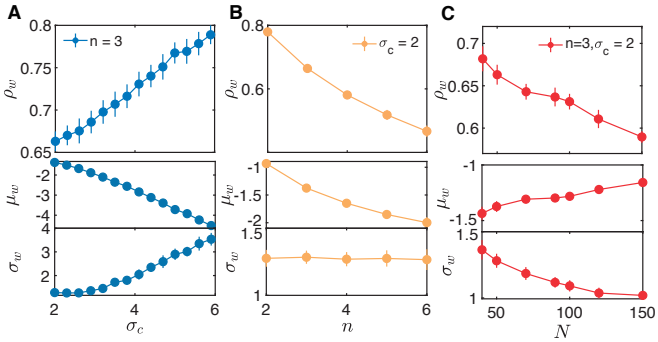

FIG. 3. Optimal sensitivity matrix depends on odor statistics. From top to bottom, dependence of the sparsity ρ_w, the average μ_w and standard deviation σ_w of the sensitive elements in the optimal W on the width of odor log concentration σ_c (A), the input sparsity n (B) and the number of total odorants N (C). In (A-B), $N = 50, M = 13$, in (C), $M = 13$. Error bars are standard deviation of 40 times simulation.

($\sim$ 59%) of the odor-ORN sensitivities to be zero (or below experimental detection limit), similar result has been reported for mouse odor receptors (90% zero) [52] (Fig. 2 D, left). The non-zero sensitivities follow a log-normal distribution with its mean and variance dependent on odor statistics. The non-zero elements in the estimated odorant-receptor sensitivity matrices from existing experimental data for fly larvae and mouse can also be fitted approximately by log-normal distributions [40, 52] as shown in Fig. 2 D (right panels). Similar log-normal distribution is found in adult *Drosophila* [39] (see Fig. 7F later in this paper).

C. The origin of sparsity in the optimal W: trade-off between sensing multiple odors and avoiding interference

Given the constraint that the number of sensors is much smaller than the possible number of odors, i.e., $M < N$, each sensor needs to respond to (sense) multiple types of odorant molecule so that all odorant molecules can be sensed by at least one sensor. However, in an odor mixture with a few types of odorant molecules, two or more odorants in the mixture can bind with the same sensor and interfere with each other, e.g., by saturating the nonlinear sensor. The probability of interference increases with the sparsity of the sensitivity matrix. This trade-off between sensing multiple odors and the possible interference effect determines the sparsity in the optimal sensitivity matrix. To put this argument into a more rigorous mathematical framework, we developed a simple mean-field theory (MFT) described below.

We begin with the simplest case where many receptors sense only one ligand ($N = 1, M \gg 1$) where obviously there is no interference (see Fig. 4A). In this case, the optimal coding scheme is for the M receptors to distribute their sensitivities to tile the input concentration space, known as the Laughlin's principle [22]. For the case when the distribution of the odorant concentration is log-normal with a variance σ_c, the optimal sensitivity distribution $P_0(w)$ that maximizes ΔH is also approximately log-normally distributed

$$P_0(w) = \frac{1}{w\sqrt{2\pi\sigma_w^2}} \exp[-\frac{(\ln w - \mu_w)^2}{2\sigma_w^2}], \tag{4}$$

where the mean $\mu_w = 0$ and the variance σ_w increases with the variance σ_c of logarithmic concentration distribution. It is also easy to show that in general the coding capacity I increases logarithmically with the number of receptors M when $M \gg 1$ (see SI Appendix for details), which agrees with the simulation results presented in Fig. 4B:

$$I_1(M) = \frac{1}{2}\log(M) + \text{const.} \tag{5}$$

This means that sparsity $\rho_w = 1$, i.e., all sensitivities should be nonzero because there is no interference when only one type of odorant molecule ($n = 1$) is present in the mixture. However, even though the maximum mutual information increases with M, it only increases sublinearly (logarithmically) for large M.

We next consider the case where two odorants are sensed by many receptors ($N = 2, M \gg 1$) as shown in Fig. 4C. Let's denote the number of receptors that respond to each ligand as m ($m \leq M$) and the sparsity $\rho_w = m/M$. If each odorant is sensed by a dis-joint set of receptors, the total differential entropy will simply double the amount for a single odorant: $I_2(m) = 2I_1(m)$. However there is a finite probability $p = m/M = \rho_w$, that a given receptor in one set will also response to the other odorant. Therefore, on average there are $m \times p = m^2/M$ receptors whose output is "corrupted" due to interference between two different odorants in a given mixture (Fig. 4C). We can write down the differential entropy as

$$I_2(m) = 2I_1(m) - \frac{m^2}{M}\Delta I, \tag{6}$$

where $I_1(m)$ is the maximum differential entropy for one odor case (Eq. 5) and ΔI is the marginal loss of information (entropy loss), which can be approximated by $\Delta I \approx \alpha \partial I_1(m)/\partial m$ where $\alpha \leq 1$ is the average fraction of information loss for a "corrupted" sensor. We can get the optimal value of m by maximizing $I_2(m)$ with respect to m. For $m \ll M$, the interference effect is small, so $I_2(m) \approx 2I_1(m)$, which increases with m logarithmically according to Eq. (5). As m increases, the interference effect given by the second term on the RHS of Eq. (6) increases with m faster than the slow logarithmic growth of $2I_1(m)$, which leads to a peak of $I_2(m)$ at a optimal value of $m = m^* < M$ or a sparsity of the sensitivity matrix $\rho_w = m^*/M < 1$ (Fig. 4 D).

In the MFT, we can compute the olfactory coding and interference by ignoring the weak rank-order correlation

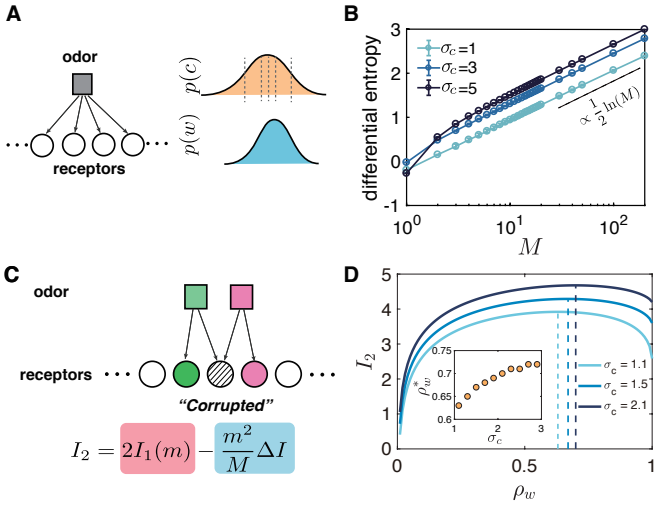

FIG. 4. A simple mean-field theory explains the sparsity of optimal W. (A)The simplest case of only one odor sensed by M receptors. The best strategy is to arrange the distribution $p(w)$ of the M receptor sensitivities to cover the concentration range weighted by their probability $p(c)$ uniformly. (B) The differential entropy in the single odor case (A) versus the number of receptors M for different width (σ_c) of odor log-concentration distribution. The differential entropy I_1 increases monotonically with M but it only grows logarithmically with M for large M. (C) In the presence of multiple odors (two odors shown here), there is a finite probability for the "corruption" of a receptor due to odor-odor interference, which result in an information loss ΔI. (D) Differential entropy I_2 for the case with two odors in the mixture each with a log-concentration width σ_c. I_2 depends non-monotonically on the fraction of sensitive receptors $\rho_w (= m/M)$ with a maximum (marked by the dashed lines) at ρ_w^* that depends on σ_c, which is shown in the inset.

in the optimal sensitivity matrix and assuming the distributions for the optimal sensitivity matrix elements are independent and identically distributed (i.i.d.). In particular, we used the following ansatz for the distribution of the sensitivity matrix W:

$$P(W) = \prod_{i=1}^{N}\prod_{j=1}^{M}[\rho_w \times P_0(w_{ij}) + (1-\rho_w)\times\delta(w_{ij})], \quad (7)$$

where ρ_w is the matrix sparsity, and $P_0(w_{ij})$ is a smooth distribution function, which is approximated here as a log-normal distribution with mean μ_w and variance σ_w as given in Eq. 4. The mean differential entropy

$$\langle \Delta H\rangle_W(\rho_w,\mu_w,\sigma_w) \equiv \int \Delta H(W)P(W)dW,$$

which is averaged over the distribution of the sensitivity matrix W, can be maximized with respect to the parameters ρ_w, μ_w, and σ_w (see SI for details). The resulting optimal parameters agree with our direct numerical simulations qualitatively with a sparsity $\rho_w < 1$ that increases with the width of the input distribution σ_c (see SI Appendix, Fig. S5).

D. Maximum entropy coding enhances downstream decoding performances

The response pattern of ORNs forms the internal representation of external odor stimuli that the higher region of brain can use to infer the odor information for controlling the organism's behavior. So far, we focused on the optimal "design" of sensitivity matrix W that maximizes the mutual information between odor statistics and ORN output. But what is this optimal coding design good for? From the theory of statistical inference, mutual information sets the upper limit of any unbiased statistical inference [45]. Below we test how the maximum entropy encoding scheme we found can affect the downstream decoding processes by examining two specific tasks: reconstruction and classification.

Task I: reconstruction. This task is directly related with the original compressed sensing problem in computer science where various downstream algorithms can be used to infer (decode) both the composition and the exact concentrations of all odorant components in an odor mixture from the sensor responses. Here, we vary the coding scheme by changing the distribution of the sensitivity matrix elements, in particular, we change the sparsity ρ_w without changing the distribution of the non-zero sensitivity matrix elements (e.g., the log-normal distribution with fixed mean and variance). As illustrated in Fig. 5A, the output of the coding process $\boldsymbol{r}(\boldsymbol{c}, W)$ serves as the input for the downstream "decoding" network. For reconstruction, we used a simple feed-forward artificial neural network (ANN) with a few (1-5) hidden layers and a output layer that has the same dimension N as the odor space. We trained the ANN with a training set of sparse odor mixtures drawn from the odor distribution $P_{\text{env}}(\boldsymbol{c})$, and tested its performance by using new odor mixtures randomly drawn from the same odor distribution. Denote the reconstructed odor vector as $\hat{\boldsymbol{c}}$ and a binary vector $\boldsymbol{\xi}$ associated with $\boldsymbol{c}$, i.e., $\xi_i = 1$ if $c_i \neq 0$, otherwise, $\xi_1 = 0$. Due to the sparse nature of $\boldsymbol{c}$ and its wide concentration range, the reconstruction error is defined as the sum of "identity error" $\mathcal{L}_1$ and "intensity error" $\mathcal{L}_2$, which is

$$\mathcal{L} \equiv \mathcal{L}_1 + \mathcal{L}_2 = -\frac{1}{N}\sum_i[\xi_i\log(z_i) + (1-\xi_i)\log(1-z_i)] + \frac{1}{N}\sum_i \xi_i[\ln(\hat{c}_i) - \ln(c_i)]^2, \quad (8)$$

where $z_i = 1/(1+\exp(-W_i^c\hat{c}_i - b_i^c)$. The reconstruction error depends on the coding matrix W, in particular its sparsity ρ_w, as shown in Fig. 5B. The best performance is achieved around $\rho_w = 0.6$ (see SI Appendix, Fig. S6 for the reconstruction performance), within the region where sparse W enable nearly maximum entropy coding (shaded region), this property is insensitive to the number of hidden layers in the reconstruction network (see SI Appendix, Fig. S7). Our results show that the opti-

mal entropy code provides an efficient representation of the high-dimensional sparse stimuli and is therefore critical (a necessary condition) for the downstream decoding algorithm to achieve a high reconstruction accuracy.

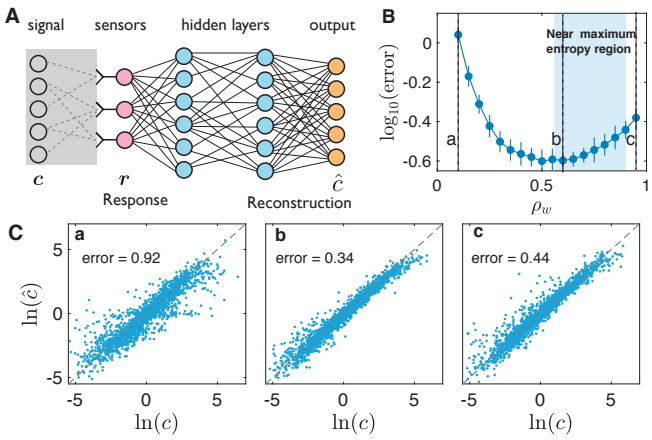

FIG. 5. Maximum entropy coding facilitates signal reconstruction. (A) Schematics of the reconstruction task. The decoding feed forward network with multiple hidden layers (blue nodes) receives the response $\boldsymbol{r}$ of the peripheral sensors (red nodes) as 'input' and produces the output $\hat{\boldsymbol{c}}$ to reconstruct (infer) the original signal $\boldsymbol{c}$. (B) The minimum reconstruction error on the testing set is achieve around sparsity level $\rho_w = 0.6$ of W, within the 95% maximum entropy region (shadowed area).(C) Comparison of the original c_i and reconstructed $\hat{c}_i$ for three different sparsity of W: (a) $\rho_w = 0.1$, (b) $\rho_w = 0.6$, (c) $\rho_w = 0.95$, which are labeled by the dashed lines in panel (B). Parameters used are: $N = 100$,$M = 20$,$n = 2$,$\sigma_c = 2$, $\sigma_0 = 0.05$, two hidden layers, each with 100 units. Error bars in (B) are standard deviation from 40 times simulation.

Task II: classification. The task for the downstream network is to infer the category of odor stimuli, such as the odor valence. We design the task and the network structure in close parallel with the *Drosophila* olfactory system, in which odor signals are sensed by the ~ 50 ORNs and are then relayed by the projection neurons (PNs) in antennal lobes to a much larger number of Kenyon cells (KCs) in the mushroom body (MB), as illustrated in Fig. 6A. Each of the ~ 2000 KCs in MB on average receives input from ~ 7 randomly selected PNs [53]. A single GABAergic neuron (APL) at each side of the brain can be activated by the KCs and inhibits KCs globally [54]. Such random expansion and global inhibition enables sparse and decorrelated representation of odor information at MB which is crucial for various tasks [53, 55, 56]. The large number of KCs in MB then converge to very few (only 34) mushroom output neurons (MBONs) [57], which project to the other brain regions and drives attractive or repulsive behavior [58]. Olfactory learning is mainly mediated by dopaminergic neurons (DANs) which controls the synaptic weight between KCs and MBONs [59]. To mimic the properties of MB, our "classifier", as illustrated in in Fig. 6B, contains a high dimensional mixed-layer (KCs). For simplicity, we consider a single readout neuron. Each KC unit in the mixed-layer pools the ORNs with a fixed random, sparse matrix. The synaptic weights from the KCs to the readout neuron are plastic .

To consider the variability of natural odors, we assumed that odor stimuli fall into clusters whose centers represent corresponding typical odor stimuli. Members in a given cluster are variations to the centroid [60]. The radius of a cluster ΔS characterizes the variability of a specific odor mixture. Centroids were drawn from $P_{\text{env}}(\boldsymbol{c})$ with randomly assigned valence (attractive or aversive) and members inside each cluster were generated by adding noise of size ΔS to the centroids, resulting in clouds of point in the odor space (Fig. 6C) with each cloud having a randomly assigned valence.

The weights from the KCs to the readout neuron are trained by using a simple linear discriminant analysis (LDA) method although other linear classification algorithms such as support vector machine (SVM) would also work. After training, the performance of the "classifier" is quantified by the accuracy of classification on the testing dataset. As shown in Fig. 6 D, we find that the best performance is achieved near $\rho_w = 0.6$, which belongs to the range of ρ_w with large mutual information between odor input and the ORN/PN response (shaded region in Fig. 6 D). Changing parameters such as M, n and number of categories give similar results (see SI Appendix, Fig. S8). In line with recent studies that sparse high-dimensional representation facilitates downstream classification [60, 61], our results suggest that maximum entropy coding at ORNs/PNs level may enhance classification by enabling the representation at KCs to be of high effective dimensions through random expansion.

E. Odor-evoked inhibition enhances the coding capacity of ORNs

So far, we have only considered the case where the neuron activity is zero in the absence of stimulus and odorants only activate the neurons ORs/ORNs. It has been widely observed that some ORNs show substantial spontaneous activities, and some odorants can act as inhibitors to suppress the activities of neurons they bind to [39, 62, 63], as shown in Fig. 7 A. The presence of an inhibitory odorant can shift a receptor's dose-response curve to an excitatory odorant, diminishing the sensitivity of receptor to excitatory odorants[64]. It is then natural to ask what is the optimal design of the sensitivity matrix to maximize coding capacity if odorants can be either excitatory or inhibitory. To answer this question, we used a four-state model to characterize both odor-evoked excitation and inhibition [64]. The sensitivity matrix now can be divided into two parts: one for excitatory interaction W_{ij}^A and the other one for inhibitory interaction W_{ij}^I. The normalized response of i-th receptor to odor $\boldsymbol{c}$

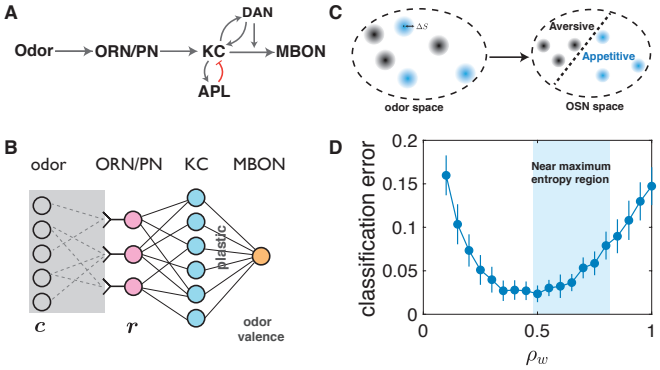

FIG. 6. Maximum entropy coding facilitates olfactory learning and classification. (A) Schematics of the neural circuitry for olfactory learning in fly, see text for detailed description. (B) A simplified model of the fly olfactory system shown in (A) for learning the valence of odor stimuli, where the effect of dopaminergic neurons (DAN) is replaced by simple plastic weights from KC to MBON. (C) Odors are organized as clusters of size ΔS and are randomly assigned with an odor valence. The decoding network receives the response pattern of ORNs and classifies them into the right categories. 100 clusters were drawn and each cluster contains 50 variations, resulting in 4000 odor stimuli among which 80% were used as training data and the rest were used as testing data. (D) Classification performance with respect to the sparsity of sensitivity matrix. Best performance appears at around $\rho_w = 0.5$, within the 95% maximum entropy region. 100 random cluster centroids and each has 50 variations to generate odor clouds. Parameters: $N = 100, M = 10, n = 3, \sigma_c = 2, \sigma_0 = 0.05, \Delta S = 0.1$, 500 KC units and two odor categories. Error bars are standard deviation from 50 simulations.

is

$$r_i = [1 + \gamma \frac{1 + \sum_{j=1}^{N} W_{ij}^{I} c_j}{1 + \sum_{j=1}^{N} W_{ij}^{A} c_j}]^{-1} + \eta_i, \tag{9}$$

where γ determines the basal activity by $r_0 = 1/(1+\gamma)$ and η_i is a Gaussian white noise.

Our numerical simulation shows that with spontaneous activity, the receptor array achieves maximum entropy coding by assigning certain number of inhibitory interactions in sensitivity matrix (Fig. 7B). The strength of both the excitatory and inhibitory elements follow log-normal distributions (Fig. 7C). As the basal activity increases, the coding capacity increases rapidly at first and quickly plateaus around $r_0 = 0.3$ (Fig. 7D). The increase of coding capacity can be understood intuitively by considering that the effective dynamical range of receptors increases in the presence of inhibition. Odor-evoked inhibition enables receptors to work bi-directionally and avoid saturation when responding to many odorants simultaneously.

The fraction of inhibitory interaction (f_i) in the optimal W is roughly proportional to the spontaneous activity of ORN r_0, with only a slight deviation when $r_0 \to 0$ and $r_0 \to 1$ (Fig. 7 E). Interestingly, as $r_0 \to 0$, f_i approaches a finite value that is related to the fraction of zero sensitivity elements $(1 - \rho_w)$ we studied in the previous sections for ORNs without a spontaneous activity.

From existing experimental data [39], we estimated the fraction of inhibitory interaction to be between $0.1 \sim 0.3$ (see SI Appendix), where the differential entropy rises sharply with r_0 as shown by the shaded region in Fig. 7 D. This seems to be a good strategy for increasing coding capacity without incurring too much metabolic energy cost associated with spontaneous firing. Although an even higher spontaneous activity towards $r_0 = 0.5$ can further increases the coding capacity, the gain is diminishing, while the metabolic cost increases drastically in maintaining the spontaneous activity [65]. Thus, an optimal basal activity would be expected in the shaded region of figure 7 D due to the trade-off between coding capacity and energy cost. We have also analyzed the statistics of the sensitivities for the excitatory and inhibitory interactions obtained from the experimental data [39]. As shown in Fig. 7 F, both the excitatory and inhibitory sensitivities follow log-normal distributions, which are good agreements with our model results shown in Fig. 7 C.

DISCUSSION

In this paper, we studied how a relatively small number of nonlinear sensors (ORNs) with a limited dynamical range can optimize transmission of high dimensional but sparse information in the environment. We found that the optimal sensitivity matrix is "sparse" – only a fraction $\rho_w (< 1)$ of sensors respond to a given odorant with their sensitivities following a broad log-normal distribution with mean μ_w and variance σ_w. The values of ρ_w, μ_w, and σ_w depend on the odor mixture statistics and the sensor characteristics. By investing the effects of different coding schemes on the downstream decoding/learning tasks, we showed that the maximum entropy code (representation) of the external signal enhances the performances of the downstream reconstruction and classification tasks. Finally, our study showed that introducing odor-evoked inhibition further enhances coding capacity for neurons with a finite spontaneous activity.

Although the exact natural odor environment is unknown, it is interesting to see the log-normal distribution of active W_{ij} in our model agrees with the sensitivity matrices estimated from experiments in fly larvae, mouse and adult fly(Fig. 2D and Fig. 7F). The relatively high level of spontaneous activity in ORNs has long been thought to only play a role in the formation of topographic map during development [66]. A recent study showed that odor-evoked inhibition can code olfactory information that drives the behavior of fly [64]. Our results provide a quantitative explanation for the advantage of having certain level of spontaneous basal activity and odor-evoked inhibitions in odor coding. By considering how the coding capacity of ORNs changes with basal activity and the associated extra energy cost[65], we were able to infer the existence of an "optimal" r_0 qualitatively

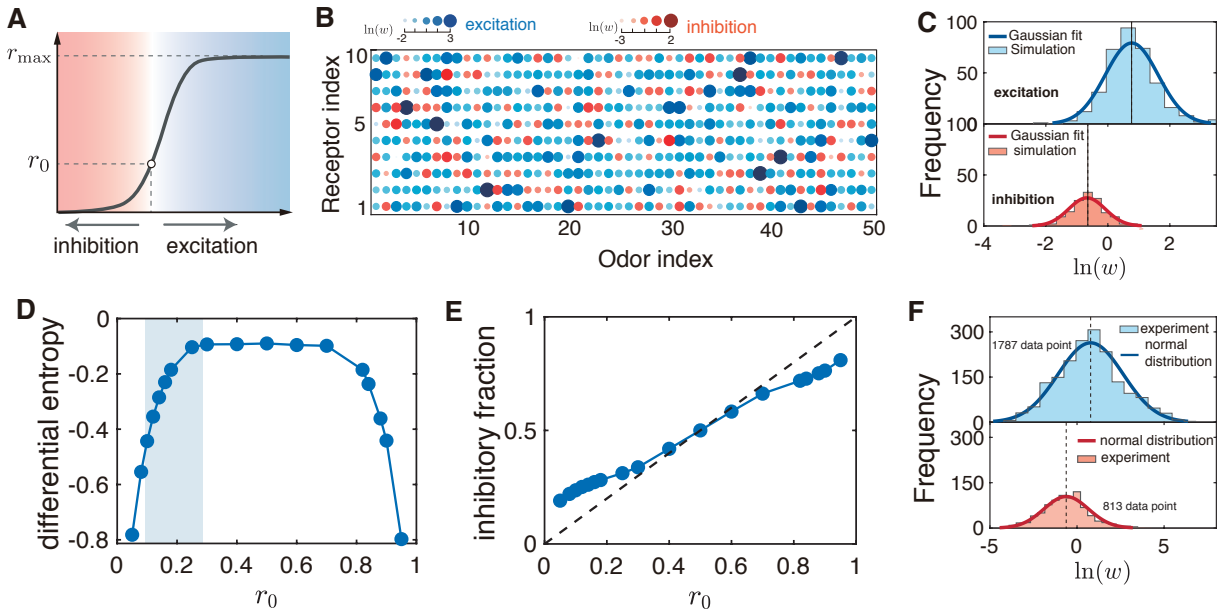

FIG. 7. Optimal sensitivity matrix when both odor-evoked excitation and inhibition are present. (A)Schematic of ORN response to excitatory (blue region) and inhibitory odorants (red region). Note that the neuron has a finite acitivity r_0 in the absence of any stimulus. (B) Heatmap of a typical optimal W, with the size of the elements indicating the strength of excitatory (blue) and inhibitory (red) interactions. (C) Both the excitatory and inhibitory interactions in optimal W can be well approximated by log-normal distribution (solids lines). (D)Differential entropy versus basal activity r_0. The shaded region shows the range of r_0 from experiments, which coincide with the range of r_0 where the differential entropy increases sharply with r_0. (E) Fraction of inhibitory interaction versus r_0. (F) The distributions of the estimated relative excitatory and inhibitory receptor-odor sensitivities from experimental data [39]. Both distributions can be well fitted by a log-normal distribution. Parameters used are $N = 50, M = 10, n = 2, \sigma_c = 2$. Error bars are small, comparable to the size of the symbols. In (A) and (B), $r_0 = 0.2$, in (C), $r_0 = 0.18$.

(Fig. 7D).

In this study, we assumed that odor information is contained in the instantaneous spiking rate of ORNs, and ignored dynamics of the neuronal activities, such as adaptation. Although adaptation plays an important role in all sensory systems [67], it happens in a relatively slower time scale than the time required for animals to detect and respond to odor stimuli [68, 69]. In general, sensory adaptation shifts the response function of the sensory neuron according to the background stimulus concentration and it leads to a larger yet still limited effective dynamical range without changing the qualitative characteristics of the input-output response curve [67, 70]. Thus, though it would be interesting to consider the effect of adaptation in optimal coding, we do not expect it to change our results qualitatively.

We have used reconstruction and classification as two learning tasks to demonstrate the advantage of having maximum entropy coding at the ORN level. While the classification task has clear biological relevance, it is unclear to what extent animals need to infer the concentrations of individual odorants in an odor mixture. The perception of odors has been thought as synthesis, i.e., odorant mixture is perceived as an unit odor [16]. Nevertheless, the performance of the reconstruction task indicates that most of the information about the odor mixture including the identities and concentrations of individual odors in a sparse mixture can potentially be extracted from the activity pattern of ORNs, which is consistent with the experimental finding that mice after training can detect target odorant in odor mixtures with up to 16 different odorants [71]. In this work, we focused only on the optimal coding strategy for the peripheral ORNs. In fly olfactory system, odorants that elicit very similar ORN response patterns can be represented by very distinct patterns of KCs [39, 56]. It remains an interesting open question whether and how architecture of the ORN/PN to KC network optimizes the odor information transmission to enhance precision of downstream learning and decision making.

In conventional compressed sensing theory with linear sensors, random measurement matrix enables accurate reconstruction of sparse high dimensional input signals [35, 38]. By using prior information about the input, better sensory matrix can be designed [72, 73]. In many cases, the optimal matrix maximizes the entropy of compressed representation [74]. Unlike the linear CS problem where the measurement matrix is known and can be used directly for reconstructing the sparse input signal (L1 minimization), reconstruction in the nonlinear CS problem studied here has to be done by learning without prior knowledge of the sensitivity matrix. Despite this differ-

ence, our results suggest that with nonlinear sensors, the sparse optimal sensory matrix that maximizes information transmission enables better learning and more accurate reconstruction. This general observation and the limit of reconstruction in nonlinear CS should be examined with more rigorous analysis and larger numerical simulations.

Finally, in our study, we considered the simplest case where odorants appear independently in odor mixture. However, even in this simplest case, we have found weak but statistically relevant "orthogonal" structure in the optimal sensitivity matrix (Fig. 2 2C and Fig. S1). In naturally occurring odor mixtures, co-occurrence of odorants in different odor sources are common. For example, odorants that are products in the same biochemical reaction pathway, i.e., fermentation, are likely to appear together [2, 75]. Although odorant-evoked ORN response patterns are not simply determined by molecular structure, some very similar odorants do trigger similar ORN response patterns [39]. On the other hand, ORNs and their responses to different odorants can be correlated due to structure similarities in their receptor proteins. It would be interesting to explore how such correlations among ORNs and odorant molecules as well as co-occurrences among different odors in odor mixtures can affect the optimal coding strategy at olfactory periphery in future study.

ACKNOWLEDGMENTS

We thank Xiaojing Yang, Jingxiang Shen, Guangwei Si and Louis Tao for helpful discussions and comments. The work was supported by the Chinese Ministry of Science and Technology (Grant No. 2015CB910300) and the National Natural Science Foundation of China (Grant No. 91430217). The work by YT is partially supported by a NIH grant (R01-GM081747).

[1] R. A. Raguso, Annual Review of Ecology, Evolution, and Systematics **39**, 549 (2008).
[2] A. Dunkel, M. Steinhaus, M. Kotthoff, B. Nowak, D. Krautwurst, P. Schieberle, and T. Hofmann, Angewandte Chemie International Edition **53**, 7124 (2014).
[3] I. Beyaert and M. Hilker, Biological Reviews **89**, 68 (2014).
[4] C.-Y. Su, K. Menuz, and J. R. Carlson, Cell **139**, 45 (2009).
[5] K. Touhara and L. B. Vosshall, Annual review of physiology **71**, 307 (2009).
[6] G. Glusman, I. Yanai, I. Rubin, and D. Lancet, Genome research **11**, 685 (2001).
[7] C. Verbeurgt, F. Wilkin, M. Tarabichi, F. Gregoire, J. E. Dumont, and P. Chatelain, PloS one **9**, e96333 (2014).
[8] C. I. Bargmann, Nature **444**, 295 (2006).
[9] J. P. McGann, Science **356**, eaam7263 (2017).
[10] C. Bushdid, M. O. Magnasco, L. B. Vosshall, and A. Keller, Science **343**, 1370 (2014).
[11] R. C. Gerkin and J. B. Castro, Elife **4**, e08127 (2015).
[12] T. Weiss, K. Snitz, A. Yablonka, R. M. Khan, D. Gafsou, E. Schneidman, and N. Sobel, Proceedings of the National Academy of Sciences **109**, 19959 (2012).
[13] N. Y. Masse, G. C. Turner, and G. S. Jefferis, Current Biology **19**, R700 (2009).
[14] B. Malnic, J. Hirono, T. Sato, and L. B. Buck, Cell **96**, 713 (1999).
[15] J. Hopfield, Proceedings of the National Academy of Sciences **96**, 12506 (1999).
[16] G. Laurent, Science **286**, 723 (1999).
[17] V. N. Murthy, Annual review of neuroscience **34**, 233 (2011).
[18] V. Grabe, A. Baschwitz, H. K. Dweck, S. Lavista-Llanos, B. S. Hansson, and S. Sachse, Cell reports **16**, 3401 (2016).
[19] O. Gschwend, N. M. Abraham, S. Lagier, F. Begnaud, I. Rodriguez, and A. Carleton, Nature neuroscience **18**, 1474 (2015).
[20] R. L. Davis, Annu. Rev. Neurosci. **28**, 275 (2005).
[21] H. B. Barlow, (1961).
[22] S. Laughlin, Zeitschrift für Naturforschung c **36**, 910 (1981).
[23] J. J. Atick, Network: Computation in neural systems **3**, 213 (1992).
[24] D. J. Field, Journal of the Optical Society of America A **4**, 2379 (1987).
[25] J. J. Atick and A. N. Redlich, Neural Computation **2**, 308 (1990).
[26] Y. Dan, J. J. Atick, and R. C. Reid, Journal of Neuroscience **16**, 3351 (1996).
[27] A. J. Bell and T. J. Sejnowski, Vision research **37**, 3327 (1997).
[28] D. L. Ruderman and W. Bialek, in *Advances in neural information processing systems* (1994) pp. 551–558.
[29] M. S. Lewicki, Nature neuroscience **5**, 356 (2002).
[30] N. Brenner, W. Bialek, and R. d. R. Van Steveninck, Neuron **26**, 695 (2000).
[31] O. Schwartz and E. P. Simoncelli, Nature neuroscience **4**, 819 (2001).
[32] X. Pitkow and M. Meister, Nature neuroscience **15**, 628 (2012).
[33] R. Haddad, R. Khan, Y. K. Takahashi, K. Mori, D. Harel, and N. Sobel, Nature methods **5**, 425 (2008).
[34] A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, *et al.*, Science **355**, 820 (2017).
[35] E. J. Candes and T. Tao, IEEE transactions on information theory **51**, 4203 (2005).
[36] E. J. Candes, J. K. Romberg, and T. Tao, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences **59**, 1207 (2006).
[37] S. Ganguli and H. Sompolinsky, Annual review of neuroscience **35**, 485 (2012).
[38] E. J. Candès and M. B. Wakin, IEEE signal processing magazine **25**, 21 (2008).

[39] E. a. Hallem and J. R. Carlson, Cell **125**, 143 (2006).
[40] G. Si, J. K. Kanwal, Y. Hu, C. J. Tabone, J. Baron, M. E. Berck, G. Vignoud, and A. D. Samuel, bioRxiv , 208538 (2017).
[41] Y. Zhang and T. O. Sharpee, PLoS computational biology **12**, e1004850 (2016).
[42] D. Zwicker, A. Murugan, and M. P. Brenner, Proceedings of the National Academy of Sciences **113**, 5570 (2016).
[43] V. Singh, N. R. Murphy, V. Balasubramanian, and J. D. Mainland, arXiv preprint arXiv:1805.00563 (2018).
[44] G. Reddy, J. D. Zak, M. Vergassola, and V. N. Murthy, eLife **7**, e34958 (2018).
[45] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).
[46] X.-X. Wei and A. A. Stocker, Neural computation **28**, 305 (2016).
[47] C. R. Sims, Science **360**, 652 (2018).
[48] G. Tkačik and W. Bialek, Annual Review of Condensed Matter Physics **7**, 89 (2016).
[49] N. Hansen and A. Ostermeier, Evolutionary computation **9**, 159 (2001).
[50] N. Hansen, in *Towards a new evolutionary computation* (Springer, 2006) pp. 75–102.
[51] A. Celani, E. Villermaux, and M. Vergassola, Physical Review X **4**, 041015 (2014).
[52] H. Saito, Q. Chi, H. Zhuang, H. Matsunami, and J. D. Mainland, Sci. Signal. **2**, ra9 (2009).
[53] S. J. Caron, V. Ruta, L. Abbott, and R. Axel, Nature **497**, 113 (2013).
[54] A. C. Lin, A. M. Bygrave, A. De Calignon, T. Lee, and G. Miesenböck, Nature neuroscience **17**, 559 (2014).
[55] G. C. Turner, M. Bazhenov, and G. Laurent, Journal of neurophysiology **99**, 734 (2008).
[56] R. A. Campbell, K. S. Honegger, H. Qin, W. Li, E. Demir, and G. C. Turner, Journal of Neuroscience **33**, 10568 (2013).
[57] 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).
[58] D. Owald, J. Felsenberg, C. B. Talbot, G. Das, E. Perisse, W. Huetteroth, and S. Waddell, Neuron **86**, 417 (2015).
[59] P. Cognigni, J. Felsenberg, and S. Waddell, Current opinion in neurobiology **49**, 51 (2018).
[60] B. Babadi and H. Sompolinsky, Neuron **83**, 1213 (2014).
[61] A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. Abbott, Neuron **93**, 1153 (2017).
[62] J. Joseph, F. A. Dunn, and M. Stopfer, Journal of Neuroscience **32**, 2900 (2012).
[63] T. Connelly, A. Savigner, and M. Ma, Journal of neurophysiology **110**, 55 (2013).
[64] L.-H. Cao, D. Yang, W. Wu, X. Zeng, B.-Y. Jing, M.-T. Li, S. Qin, C. Tang, Y. Tu, and D.-G. Luo, Nature communications **8**, 1357 (2017).
[65] J. E. Niven and S. B. Laughlin, Journal of Experimental Biology **211**, 1792 (2008).
[66] A. Nakashima, H. Takeuchi, T. Imai, H. Saito, H. Kiyonari, T. Abe, M. Chen, L. S. Weinstein, C. R. Yu, D. R. Storm, *et al.*, Cell **154**, 1314 (2013).
[67] B. Wark, B. N. Lundstrom, and A. Fairhall, Current opinion in neurobiology **17**, 423 (2007).
[68] A. Egea-Weiss, C. J. Kleineidam, P. Szyszka, *et al.*, iScience **4**, 76 (2018).
[69] Q. Gaudry, E. J. Hong, J. Kain, B. L. de Bivort, and R. I. Wilson, Nature **493**, 424 (2013).
[70] C. Martelli, J. R. Carlson, and T. Emonet, Journal of Neuroscience **33**, 6285 (2013).
[71] D. Rokni, V. Hemmelder, V. Kapoor, and V. N. Murthy, Nature neuroscience **17**, 1225 (2014).
[72] Y. Weiss, H. S. Chang, and W. T. Freeman, in *Snowbird Learning Workshop, Allerton, CA* (Citeseer, 2007).
[73] D. L. Donoho, A. Javanmard, and A. Montanari, IEEE transactions on information theory **59**, 7434 (2013).
[74] A. Ashok, L.-C. Huang, and M. A. Neifeld, JOSA A **30**, 831 (2013).
[75] S. Mansourian and M. C. Stensmyr, Current opinion in neurobiology **34**, 95 (2015).
[76] M. Chiani, Journal of Multivariate Analysis **129**, 69 (2014).
[77] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns, Human brain mapping **38**, 1541 (2017).
[78] S. Singh and B. Pøczos, arXiv preprint arXiv:1702.07803 (2017).
[79] Z. Szabó, The Journal of Machine Learning Research **15**, 283 (2014).
[80] S. X. Luo, R. Axel, and L. Abbott, Proceedings of the National Academy of Sciences **107**, 10713 (2010).
[81] K. Krishnamurthy, A. M. Hermundstad, T. Mora, A. M. Walczak, and V. Balasubramanian, arXiv preprint arXiv:1707.01962 (2017).

Supplemental Information: Statistical properties of the optimal sensitivity matrix for compressed sensing with nonlinear sensors

S1. ORN RESPONSE

A. Odor-receptor interaction and OR sensitivities

We first model odor-evoked ORN response using an equilibrium binding model, where odor receptors (ORs) have two states, ON and OFF. The activity of ORN is then determined by the fraction of ORs that are in the ON state. Odor receptor R_i is activated by the binding of an odorant j with concentration c_j, which can be described as a chemical reaction $R_i + c_j \underset{k_+}{\overset{k_-}{\rightleftharpoons}} R_i c_j$, where $R_i c_j$ is the receptor-odorant complex. At equilibrium, the activity of ORN is proportional to the fraction of odor-bound receptor, which reads

$$r_i = r_{\max} \frac{c_j/K_m^{ij}}{1 + c_j/K_m^{ij}} = r_{\max} \frac{c_j W_{ij}}{1 + c_j W_{ij}}, \tag{S1}$$

where $r_{\max}$ is the maximum neuronal activity, i.e., the spiking rate and we have set it to be 1 for simplicity. $K_m^{ij} \equiv k_-/k_+$ is the binding constant between odorant j and receptor i, $W_{ij} \equiv 1/K_m^{ij}$ is the sensitivity constant.

Response of ORNs to multiple odorants were characterized by a simple competitive binding model, thus odorants compete for the only binding site of the receptor. The response of R_i to the possible N odorants can be described by the following reactions:

$$\begin{aligned} c_1 + R_i &\underset{k_{1-}}{\overset{k_{1+}}{\rightleftharpoons}} R_i c_1, \\ &\cdots \\ c_j + R_i &\underset{k_{i-}}{\overset{k_{i+}}{\rightleftharpoons}} R_i c_j, \\ &\cdots \\ c_N + R_i &\underset{k_{N-}}{\overset{k_{N+}}{\rightleftharpoons}} R_i c_N. \end{aligned}$$

At equilibrium, the activity of ith ORN is determined by the fraction of R_i in the bound state

$$\theta \equiv \frac{\sum_j^N R_i c_j}{R_{tot}} = \frac{\sum_{j=1}^N c_j/K_m^{ij}}{1 + \sum_{j=1}^N c_j/K_m^{ij}}, \tag{S2}$$

where $K_m^{ij} \equiv k_{i-}/k_{i+}$ is the disassociation constant. Hence, the response of ith ORN is

$$r_i = r_{\max}\theta = r_{\max} \frac{\sum_j^N c_j W_{ij}}{1 + \sum_j^N c_j W_{ij}}, \tag{S3}$$

with $W_{ij} \equiv 1/K_m^{ij}$.

B. Inhibitory response

Model for both odor-evoked excitation and inhibition has been explained in detail in [64]. The ORN has four states that belong to two categories: ON and OFF. The difference of free energy between the ON and OFF states of an OR is modulated by the odorants binding to the receptor.

$$\begin{aligned} \Delta F &= F_{i,I} - F_{i,A} \\ &= E_{0,i} + \ln(1 + \sum_j c_j W_{ij}^A) - \ln(1 + \sum_j c_j W_{ij}^I), \end{aligned} \tag{S4}$$

where W_{ij}^A and W_{ij}^I are the sensitivity of the ON and OFF state of the ith OR to the jth odorant. For simplicity, we have assumed that an excitatory odorant only bind to the ON state ($W_{ij}^I = 0$), while an inhibitory odorant only bind to the OFF state ($W_{ij}^A = 0$). The activity of the ith ORN is given by

$$r_i = \frac{r_{\max}}{1 + e^{-\Delta F}} = r_{\max}[1 + \gamma_i \frac{1 + \sum_{j=1}^N c_j W_{ij}^I}{1 + \sum_{j=1}^N c_j W_{ij}^A}]^{-1}. \tag{S5}$$

where γ_i determines the basal activity r_i^0 of the ith ORN through $\gamma_i = r_{\max}/r_i^0 - 1$, and we have omitted the Boltzmann factor $k_B T$ in above equation for simplicity. In our model, basal activity of each receptors have been set to be the same, i.e., $r_0^i = r_0, \gamma_i = \gamma (i = 1, \cdots, M)$.

C. Measured sensitivities of olfactory receptors

Response of ORs to a panel of odorants have been measured for fly larvae [40], adult fly [39] and mouse [52]. For fly larvae, cognate odorants have been identified for each of the 18 ORs. Then a dose response curve was measured for each odorant-receptor pair. And the sensitivity of an OR to the odorant was estimated by fitting the dose response curve with a Hill function, the resulting sensitivity matrix was reported in [40]. For mouse receptors, odor mixtures drawn from a panel of 93 odorants with various concentrations were tested against 219 receptors, and 169 of them showed some level of response. A further dose-response test resulted in a response matrix for 52 ORs and 63 odorants [52]. We calculate the fraction of sensitive OR-odorant interactions directly from these sensitivity matrices.

For adult fly, the responses of 24 ORs to a panel of 110 odorants have been measured, together with the spontaneous firing rate of each OR in the absent of stimulus [39]. Since the measurement was carried out using the same concentration of odorants, for a rough estimation,

we fitted the response to Eq. S5 to estimate the relative sensitivity for both excitatory and inhibitory interactions. The resulting distributions of W^A and W^I are shown in figure 7F, and the fraction of inhibitory interactions is about 30%. If we take a more stringent criterion and only consider the strong interactions ($>$ 50Hz for excitatory interaction and $>$ %50 of suppression for inhibitory interaction), then the fraction of inhibition is about 11%. Thus the actual fraction of inhibitory interactions should be in the range of $0.1 \sim 0.3$, as shown in the shaded region in Fig. 7D.

S2. MAXIMUM ENTROPY CODING

As shown in Fig 1A in the main text, the "coding" problem is basically how the odor vector $\boldsymbol{c}$ is mapped to the ORN activity vector $\boldsymbol{r}$

$$r_i = F_i(\boldsymbol{c}) + \eta_i, \tag{S6}$$

where $F_i(\boldsymbol{c})$ is the nonlinear response function for the i-th neuron ($i = 1, 2, \ldots, M$) as introduced in previous section, and η_i is the noise, which may assume certain specific form given its biophysical/biochemical origin. For convenience, here we consider it as Gaussian noise with zero mean and a standard deviation σ_0.

For a given odor environment, characterized by the mixture distribution $P_{\text{env}}(\boldsymbol{c})$, we can derive the distribution of ORN response

$$P_r(\boldsymbol{r}) = \int (2\pi\sigma_c^2)^{-\frac{M}{2}} \exp[-\sum_{i=1}^{M} \frac{(r_i - F_i(\boldsymbol{c}))^2}{2\sigma_0^2}] P_{\text{env}}(\boldsymbol{c}) d\boldsymbol{c}. \tag{S7}$$

The optimal entropy coding strategy is to determine the response functions, which are characterized by the sensitivity matrix W_{ij}, to maximize the mutual information between odor environment $P_{\text{env}}(\boldsymbol{c})$ and ORN response pattern $P_r(\boldsymbol{r})$, defined as

$$\begin{aligned} I(\boldsymbol{c}, \boldsymbol{r}) &= H(\boldsymbol{r}) - H(\boldsymbol{r}|\boldsymbol{c}) \\ &= -\int P_{\text{env}}(\boldsymbol{c}, \boldsymbol{r}) \log \frac{P_{\text{env}}(\boldsymbol{c}, \boldsymbol{r})}{P_r(\boldsymbol{r}) P_{\text{env}}(\boldsymbol{c})} d\boldsymbol{c} d\boldsymbol{r}. \end{aligned} \tag{S8}$$

In the limit of small noise, the conditional entropy $H(\boldsymbol{r}|\boldsymbol{c}) \to 0$, hence $I(\boldsymbol{c}, \boldsymbol{r}) \approx H(\boldsymbol{r})$, the differential entropy in the response space is

$$H(\boldsymbol{r}) = -\int P_r(\boldsymbol{r}) \ln[P_r(\boldsymbol{r})] d\boldsymbol{r}. \tag{S9}$$

We first considered the log-normal distribution of odor concentration, then $H(\boldsymbol{r})$ can be simplified as

$$\begin{aligned} H(\boldsymbol{r}) &\approx -\int P_{\text{env}}(\boldsymbol{c}) \ln\{\int (2\pi\sigma_0^2)^{-M/2} \exp[-\sum_{j=1}^{M} (F_j(\boldsymbol{c}) - F_j(\boldsymbol{c'}))^2/(2\sigma_0^2)] P_{\text{env}}(\boldsymbol{c'}) d\boldsymbol{c'}\} d\boldsymbol{c} \\ &\approx -\int P_{\text{env}}(\boldsymbol{c}) \ln P_{\text{env}}(\boldsymbol{c}) d\boldsymbol{c} + (M - n) \ln(\sqrt{2\pi}\sigma_0) + \frac{1}{2} \int P_{\text{env}}(\boldsymbol{c}) \ln[\det(\chi)] \text{dc} \end{aligned} \tag{S10}$$

where χ is a $N \times N$ Jacobian matrix with its matrix elements given as:

$$\chi_{kl} = \sum_{i=1}^{M} \frac{\partial r_i}{\partial c_k} \times \frac{\partial r_i}{\partial c_l}. \tag{S11}$$

The first two terms in the expression for $H(\boldsymbol{r})$, Eq.S10, are constants with clear physical meanings. The first term is just the entropy of the odor distribution, and the second term is due to dimension reduction. The third term is the most interesting term. It depends on the sensitivity matrix and it is this term that we need to optimize given the odor distribution $P_{\text{env}}(\boldsymbol{c})$ and the form of the response function F. We denote the third term as

$$\Delta H \equiv \frac{1}{2} \int P_{\text{env}}(\boldsymbol{c}) \ln[\det(\chi)] \text{dc}, \tag{S12}$$

which plays the central rule in our model and analysis. Here and after, we will call ΔH as our simplified target function which is to be maximized.

S3. CORRELATION AMONG RECEPTORS

In fig. 2C and fig. S1, we showed that the optimal sensitivity matrix W shows weak but statistically significant higher order structure by comparing the receptor-wise and odorant-wise rank correlations with randomly shuffled matrices. We now further verify this by showing that W is indeed different from random matrix, using other forms of statistical test.

We compared the eigenvalues of Spearman rank correlation matrix for the original W and its randomly shuffled versions. As shown in fig. S2 A and B, the original W has a much narrower distribution of eigenvalues (measured as the standard deviation σ_λ). The "sparse" structure of the optimal sensitivity matrix makes it difficult to perform direct statistical tests because the matrix variate distribution is complicated, to circumvent the difficulty, we first transformed the distribution of matrix elements to the standard Gaussian distribution. We denote the 1-D random variable of the elements in

the matrix as X, from which we constructed the empirical CDF $F(x) = P(X \leqslant x)$, we then used the function $y = N^{-1}(F(x))$ to project each element x in the matrix to y, where $N(x)$ is the CDF for the standard Gaussian distribution. Thus we mapped the optimal matrix W to a transformed matrix $\hat{W}$, their elements have one-to-one mapping according to the above transformation. After that, we can perform statistical test.

Our null hypothesis is that the sensitivity matrix is random, i.e., $\hat{W}_{ij} \sim \mathcal{N}(0,1)$, and the two-point covariance $\sigma^2_{ik,jl} = \delta_{ij}\delta_{kl}$ where (i,k) and (j,l) are matrix indices. Previous work shows that if $\hat{W} \in \mathbb{R}^{m\times n}$, whose entries are i.i.d. of standard Gaussian and λ_1 is the largest eigenvalue of the inner product matrix, $\hat{W}\hat{W}^T$, then for sufficient large n, m, the normalized largest eigenvalue follows the Tracy-Widom distribution of order 1 [76].

$$\lambda_{max} \equiv \frac{\lambda_1 - \mu_{mn}}{\sigma_{mn}} \to \mathcal{TW}_1, \tag{S13}$$

where

$$\mu_{mn} = (\sqrt{n-1/2} + \sqrt{m-1/2})^2,$$
$$\sigma_{mn} = \sqrt{\mu_{mn}}\left(\frac{1}{\sqrt{n-1/2}} + \frac{1}{\sqrt{m-1/2}}\right)^{\frac{1}{3}}.$$

To test if the transformed optimal matrix, $\hat{W}$ or its shuffled versions follow such distribution, we first generated 5000 new matrices by randomly shuffling the elements of $\hat{W}$. We then computed the largest eigenvalues of associated inner product matrices, the result is shown in Fig. S2 C. Compared with the distribution of the largest eigenvalues the random shuffled matrices and the asymptotic $\mathcal{TW}_1$, the normalized largest eigenvalue of $\hat{W}$ is an outlier, indicating that the null hypothesis can be rejected. Taken together, although all pairwise correlations are weak, the optimal sensitivity matrix significantly deviates from random matrix.

S4. MEAN FIELD APPROXIMATION

In this section, we describe our mean filed approximation that helps understanding the origin of the sparsity of optimal odor-receptor sensitivity matrix in detail. We will first consider the case with only one odorant that has broad concentration distribution and many receptors ($N = n = 1, M \gg 1$), and then the two odorants scenario ($N = n = 2, M \gg 1$), where we demonstrate the existence of "optimal" sparsity (ρ_w) of sensitivity matrix. We next extend the result to a more general scenario, i.e., $N, M \gg 1, n = 2$, where the upper bound of ρ_w can be approximated using the results from the case with $N = n = 2, M \gg 1$.

A. One odor and multiple receptors

Since we are considering the log-normally distributed odorant concentration, it is convenient to work in logscale. We now define $K_{ij} \equiv \ln(K^{ij}_m)$ and $\bar{c} = \ln(c)$. For the case $N = n = 1$ and $M \gg 1$, Eq. S12 becomes

$$\Delta H = \frac{1}{2}\int P_{\text{env}}(\bar{c}) \ln[\sum_{i=1}^{M} G(\bar{c} - K_i)]dc, \tag{S14}$$

where $G(x) = e^{2x}/(1+e^x)^4$ is the square of the derivative of the response function, it has a peak at $x = 0$ with a width ~ 1.

The optimal choices of K values can be obtained by solving the equations: $\partial\Delta H/\partial K_i = 0$ for all $i = 1, 2, \ldots, M$. In the limit of $M \to \infty$, the problem can be solved by assuming a distribution of K, $P_k(K)$, which leads to:

$$\Delta H = \frac{1}{2}\int P_{\text{env}}(\bar{c}) \ln[M \int G(\bar{c} - K)P_k(K)dK]d\bar{c}. \tag{S15}$$

Now the problem becomes optimizing ΔH with respect to the distribution $P_k(K)$ and with the constraint $\int P_k(K)dK = 1$. Using variational method, the optimum distribution is the one that satisfies

$$\int G(\bar{c} - K)P_k(K)dK = \lambda P_{\text{env}}(\bar{c}), \tag{S16}$$

where λ is a Lagrange multiplier that can be determined by the normalization constraint, and it is $\lambda = \int G(x)dx = \frac{1}{6}$. Eq. S16 can be solved by any deconvolution method:

$$P_k(K) = \lambda \int G^{-1}(\bar{c} - K)P_{\text{env}}(\bar{c})d\bar{c}, \tag{S17}$$

where G^{-1} is the inverse convolution function (operator) of G, i.e.,$\int G(\bar{c} - K)G^{-1}(K - \bar{c}')dK = \delta(\bar{c} - \bar{c}')$.

To get the distribution of K from Eq. S16, we first approximate $G(x) = \frac{e^{2x}}{(1+e^x)^4}$ by the density function of Gaussian distribution, $\text{Gauss}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$(they have very similar shape). From the Fourier transformation of Eq. S16, we have

$$\mathcal{F}\{P_k\} = \frac{\mathcal{F}\{P_{\text{env}}\}}{\mathcal{F}\{G\}}, \tag{S18}$$

where

$$\mathcal{F}\{P_{\text{env}}\} = \frac{1}{2\pi}\exp[-iw\mu - \frac{1}{2}w^2\sigma_c^2] \tag{S19}$$

and

$$\mathcal{F}\{G\} = \frac{1}{2\pi}\exp[-\frac{1}{2}w^2\sigma_r^2]. \tag{S20}$$

So that we have

$$\mathcal{F}\{P_k\} = \frac{1}{2\pi}\exp[-\frac{1}{2}w^2(\sigma_c^2 - 1)], \tag{S21}$$

hence, by inverse Fourier transformation we have

$$P_k(K) \propto \exp[-\frac{K^2}{2(\sigma_c^2-1)}], \tag{S22}$$

which is a Gaussian distribution with a standard deviation $\sigma' = \sqrt{\sigma_c^2 - 1}$. By definition, $W_{ij} = \exp(-K_{ij})$, thus W_{ij} follows a log-normal distribution, which is Eq. 4 in main text.

As $M \to \infty$, plug Eq. S22 in Eq. S15, we have

$$I_1(M) = \Delta H + H(\boldsymbol{c}) = \frac{1}{2}\log(M) + \frac{1}{4}\log(2\pi e\sigma_c^2\lambda^2), \tag{S23}$$

where σ_c is the standard deviation of the input and the only term in $I_1(M)$ that depends on M is $\frac{1}{2}\log M$. Thus, when M is sufficiently large, $I_1(M)$ increases logarithmically with M.

B. Two odors and multiple receptors

In this scenario, we use both explicit computation and phenomenological model to approximate Eq. S15. In order to compute Eq. S15 explictly, we first write down the matrix elements of χ

$$\begin{aligned}
\chi_{11} &= \Delta_1^{-4}\sum_{i=1}^{M}\exp[2(\bar{c}_1 - K_{i1})],\\
\chi_{22} &= \Delta_2^{-4}\sum_{i=1}^{M}\exp[2(\bar{c}_2 - K_{i2})],\\
\chi_{12} &= \chi_{21} = (\Delta_1\Delta_2)^{-2}\sum_{i=1}^{M}\exp(\bar{c}_1 + \bar{c}_2 - K_{i1} - K_{i2})),
\end{aligned} \tag{S24}$$

where $\Delta_i \equiv [1 + \exp(\bar{c}_1 - K_{i1}) + \exp(\bar{c}_2 - K_{i2})]$ for $i = 1, 2$. From the χ matrix, we can calculate its determinant $\det(\chi) = \chi_{11}\chi_{22} - \chi_{12}^2$ and thus obtain the expression for ΔH: where $\vec{K_j} = (K_{1j}, K_{2j}, \ldots, K_{Mj})$ is a dimension-M sensitivity vector for odor j in the receptor space. The optimal design can be obtained by requiring $\partial\Delta H/\partial K_{ij} = 0$ for all $j = 1, 2$ and $i = 1, 2, \ldots, M$. The above formalism/procedure can also be used here to study the cases with $M = 2, M = 3, \ldots$. In our simulation with $N = 2$, we directly optimize the explicit form of entropy rather than estimating the entropy at each iteration.

In the limit of $M \to \infty$, the problem can be tackled using a distribution function $P_k(K_1, K_2)$. The matrix elements of χ now depend on this distribution. Using variational method similar to the case for one $N = 1$, the final equation for $P_k(K_1, K_2)$ is

$$\int \frac{P_{\text{env}}(\bar{\boldsymbol{c}})\int G(\bar{\boldsymbol{c}}, \vec{K}, \vec{K}')P_k(\vec{K}')d\vec{K}'}{\int G(\bar{\boldsymbol{c}}, \vec{K}, \vec{K}')P_k(\vec{K}')P_k(\vec{K})d\vec{K}d\vec{K}'} = \lambda, \tag{S25}$$

where

$$G(\bar{\boldsymbol{c}}, \vec{K}, \vec{K}') = \frac{e^{2(\bar{c}_1+\bar{c}_2)}[e^{-2(K_1+K_2')} - e^{-(K_1+K_2+K_1'+K_2')}]}{[(1+e^{\bar{c}_1-K_1}+e^{\bar{c}_2-K_2})(1+e^{\bar{c}_1-K_1'}+e^{\bar{c}_2-K_2'})]^4},$$

and $\vec{K} = (K_1, K_2)$ and $\vec{K}' = (K_1', K_2')$ are the sensitivity vectors, whose components are the sensitivity of a given receptor to one of the two odorants. By integrating over $\vec{K}$ on both sides of Eq. S25, we can determine $\lambda = \int P_{\text{env}}(\bar{\boldsymbol{c}})d\bar{\boldsymbol{c}} = 1$.

To simplify the notation, we now define $v_1 = \exp(-K_1), v_2 = \exp(-K_2)$. Now we have an explicit form of ΔH when $N = 2$, we shall first use appropriate approximation to demonstrate the existence of "sparseness" when $N = 2$, and use numerical calculations to determine how ρ and $p_s(v)$ depend on input statistics. Further, we extend the case to $n = 2, N > 2$, and demonstrate that the fraction of sensitive elements at $N = 2$ is an upper bound for scenarios where $n = 2, N > 2$.

For a large number of receptors and two odorants considered here, we can assume that the distributions for the two sensitivities v_1 and v_2 are independent with the joint distribution

$$\begin{aligned}
P(v_1, v_2) =&[\rho \times p_s(v_1) + (1-\rho)\times\delta(v_1)]\\
&\times[\rho \times p_s(v_2) + (1-\rho)\times\delta(v_2)],
\end{aligned} \tag{S26}$$

where $p_s(v)$ is a smooth distribution function that is the same for all odor-receptor pairs, and ρ is just the fraction of non-zero elements, i.e, the "sparsity" of the matrix.

We can now write down Eq. S24 explicitly in this particular case,

$$\begin{aligned}
\chi_{11} &= M\rho[\rho F_2(c_1, c_2) + (1-\rho)F_1(c_1)],\\
\chi_{22} &= M\rho[\rho F_2(c_2, c_1) + (1-\rho)F_1(c_2)],\\
\chi_{12} &= \chi_{21} = M\rho^2 F_{12}(c_1, c_2),
\end{aligned} \tag{S27}$$

where

$$\begin{aligned}
F_2(c_1, c_2) &= \int\int \frac{v_1^2}{(1+v_1c_1+v_2c_2)^4}p_s(v_1)p_s(v_2)dv_1dv_2,\\
F_1(c) &= \int \frac{v^2}{(1+vc)^4}p_s(v)dv,\\
F_{12}(c_1, c_2) &= F_{12}(c_2, c_1)\\
&= \int\int \frac{v_1v_2}{(1+v_1c_1+v_2c_2)^4}p(v_1)p(v_2)dv_1dv_2.
\end{aligned}$$

Plug Eq. S27 into Eq. S12, we have the expression for the differential entropy

$$\Delta H(\rho, p_s(v)) = \ln(\rho M) + \frac{1}{2}\int P(c_1, c_2)\ln\Sigma(c_1, c_2)dc_1dc_2, \tag{S28}$$

where $\Sigma(c_1, c_2)$ is defined as:

$$\begin{aligned}
\Sigma(c_1, c_2) =&[\rho F_2(c_1, c_2) + (1-\rho)F_1(c_1)]\\
&\times[\rho F_2(c_2, c_1) + (1-\rho)F_1(c_2)] - \rho^2 F_{12}^2(c_1, c_2).
\end{aligned} \tag{S29}$$

To compare with the one odor case, Eq. S28 can be further written as

$$I_2(\rho) = 2I_1(\rho M) + \frac{1}{2}\int p(c_1,c_2)\ln\{1+\rho[\frac{F_1(c_2,c_1)}{F_1(c_2)}+\frac{F_2(c_1,c_2)}{F_1(c_1)}-2] + \rho^2[1+\frac{F_2(c_1,c_2)F_2(c_2,c_1)}{F_1(c_1)F_1(c_2)}-\frac{F_2(c_1,c_2)}{F_1(c_1)}-\frac{F_2(c_2,c_1)}{F_1(c_2)}-\frac{F_{12}^2(c_1,c_2)}{F_1(c_1)F_1(c_2)}]\}dc_1dc_2. \tag{S30}$$

Given a value of ρ, and assuming $p_s(v)$ is the same as one odorant case, the second term of the above equation can be numerically integrated directly. As shown in Fig. 4D in the main text, I_2 achieves maximum value at an optimal $\rho^* < 1$.

C. The interference between mixtures of odors

Eq. S15 can be also approximated through a phenomenological equation by considering the interference between the two odorants. As in the main text, an intuitive mean-field equation can be written down

$$I_2(\rho M) = 2I_1(\rho M) - \alpha\rho^2 M\Delta I, \tag{S31}$$

where α is a phenomenological factor that quantifies how much the information conveyed by one receptor is "corrupted" when the receptor is active to two odorants in the same mixture, and $\rho \equiv \frac{m}{M}$ is what we defined as "sparsity" of the sensitivity matrix, where m is the average number of receptors activated by one odorant. Comparing Eq. S31 with Eq. S30, we know that the first term of Eq. S31 increases logarithmically with m , while the second term increases supralinearly with m, and there will be a cross-over point which leads to a peak of $I_2(\rho M)$.

We next extend our mean-field theory to more general cases where there are n odorants (can be larger than 2) in odor mixtures. While unable to derive analytically how ρ and $p_s(v)$ depend on input parameters, we will show that our previous result set an upper limit of ρ_w, thus introducing the interference would only cause the optimal ρ_w to be a smaller value. Start with $N = 3$ and $n = 2$,

$$I_3^2(\rho M) = 2I_1(\rho M) - \alpha\rho^2\Delta I - (\theta-\alpha)\rho^3\Delta I. \tag{S32}$$

The contribution of the higher order term depend on the difference between θ and α, and θ is a phenomenological factor that quantifies how much the information conveyed by one receptor is "corrupted" when the receptor is active to all 3 odors. It is obvious that $\theta - \alpha > 0$. This can be extended to more general case, i.e., $n > 2, N > 2$ where there are n odors in each sample and N odors altogether, we have

$$I_N^n(\rho M) = nI_1(\rho M) - \sum_{n_0=n}^{N}\alpha_{n_0}\binom{N}{n_0}\frac{\binom{N}{n}}{\binom{n_0}{n}}q^{n_0}(1-q)^{N-n_0}\Delta I \tag{S33}$$

with

$$\alpha_n < \alpha_{n+1} < \cdots < \alpha_N,$$

and α_n should increase with n sublinearly, because α_n describes the loss of information when a receptor is active to n odorants, and this information has a lower bound, as n increases, the loss increases but saturates gradually. It is easy to see from Eq. S32 that the interference between odor mixtures adds more negative terms, thus would lead to a smaller optimal sparsity level.

D. Analytical approximations extending to the case N>2, n=2

With $N > 2$, $M \to \infty$, and $n = 2$, the solution for the optimal sensitivity matrix can be also formulated as a distribution of the weights $P(K)$. Since the distribution of the odorants are i.i.d., it is reasonable to assume a similar symmetric structure for the optimal $P(K)$, suggesting that $P(K)$ should be the same for different odorants, and $P(K_1, K_2)$ should be the same for different odorant pairs.

Since only two odorants appear in any mixtures, the leading term of the constraints on the distribution of K therefore comes from the "pairwise interaction"(receptors that are sensitive to both odorants are likely to be "corrupted"). For simplicity, we ignore the "interference" between different samples. As we have shown in the previous section where we fixed the distribution of the sensitive part, ignoring the "interference" term would give us an upper bound of the optimal sparsity level.

Therefore we have reduced the problem to two odorants and infinite receptors, the interaction between any pairs of odorants becomes weaker. So Eq. S28 can approximate the target function. Given the constraint $\int p_s(v)dv = 1$, the problem becomes optimizing the following function with respect to $\rho, p_s(v)$:

$$\Delta G(\rho, p_s(v), \lambda) = \Delta H - \lambda(\int p_s(v) - 1), \tag{S34}$$

where λ is the Lagrange multiplier. Taking the derivative with respect to ρ, we have:

$$D_\rho \equiv \frac{1}{2}\int \frac{P(c_1,c_2)}{\Sigma(c_1,c_2)}\{(F_1(c_1)-F_2(c_1,c_2))[\rho F_2(c_1,c_2)+(1-\rho)F_1(c_2)] \\ +(F_1(c_2)-F_2(c_2,c_1))[\rho F_2(c_1,c_2)+(1-\rho)F_1(c_1)]+2\rho F_{12}^2(c_1,c_2)\}dc_1dc_2=\rho^{-1}. \tag{S35}$$

To obtain the optimal functional form of $p_s(v)$ directly through variational method is hard, therefore we further parameterized the distribution $p_s(v)$ with Gaussian distribution, i.e., $p_s(v)\sim\mathcal{N}(\mu_v,\sigma_v)$, which is supported by our numerical simulation. Therefore we can take derivative of Eq. S28 with respect to μ_v and σ_v

$$D_\zeta \equiv \frac{\partial \Delta H}{\partial \zeta} = \frac{1}{2}\int \frac{P(c_1,c_2)}{\Sigma(c_1,c_2)}\frac{\partial \Sigma(c_1,c_2)}{\partial \zeta}dc_1dc_2=0, \quad \zeta=\mu_v,\sigma_v, \tag{S36}$$

where

$$\Sigma_\zeta(c_1,c_2)=[\rho G_2^\zeta(c_1,c_2)+(1-\rho)G_1^\zeta(c_1)]\times[\rho F_2(c_1,c_2)+(1-\rho)F_1(c_2)] \\ +[\rho F_2(c_1,c_2)+(1-\rho)F_1(c_1)]\times[\rho G_2^\zeta(c_2,c_1)+(1-\rho)G_1^\zeta(c_2)]-2\rho^2G_{12}^\zeta(c_2,c_1)F_{12}(c_2,c_1),\ \zeta=\mu_v,\sigma_v,$$

and

$$G_1^\zeta \equiv \frac{\partial F_1(c)}{\partial \zeta}=\int \frac{v^2}{(1+vc)^4}\frac{\partial p_s(v)}{\partial \zeta}dv, \quad G_2^\zeta \equiv \frac{\partial F_2(c_1,c_2)}{\partial \zeta}=\int \frac{v_1^2}{(1+v_1c_1+v_2c_2)^4}\frac{\partial [p_s(v_1)p_s(v_2)]}{\partial \zeta}dv, \\ G_{12}^\zeta \equiv \frac{\partial F_{12}(c_1,c_2)}{\partial \zeta}=\int \frac{v_1v_2}{(1+v_1c_1+v_2c_2)^4}\frac{\partial [p_s(v_1)p_s(v_2)]}{\partial \zeta}dv_1dv_2,\ \zeta=\mu_v,\sigma_v.$$

We then solve the above three coupled equations by turning the problem into an optimization problem

$$\{\mu_v,\sigma_v,\rho\}=\arg\min_{\{\mu_v,\sigma_v,\rho\}}[(D_\rho-\rho^{-1})^2+D_\mu^2+D_\sigma^2]. \tag{S37}$$

The solution is shown in Fig S5, which agree with the direct simulation qualitatively (see First column of Fig. 3). The quantitative difference is due to the omission of the interference between odor samples.

NUMERICAL OPTIMIZATION PROCEDURES

In this section, we explain in detail the procedure of numerical simulation to search for optimal sensitivity matrix. We first explain how we estimate the target function, i.e., differential entropy, at each step of the iteration, and how the genetic algorithm works to find a local minimum of the target function in each trial. Then, we explain the difference in optimization for the excitation-only case and the optimization procedure for receptors with both excitatory and inhibitory response.

E. Entropy estimation

The response of ORNs $\boldsymbol{r}$ is generated through Eq. S3 using 50000 odor stimuli sampled from $P_{\text{env}}(\boldsymbol{c})$. The entropy of $P_r(\boldsymbol{r})$ is estimated using a copula-based method, the Gaussian Copula Mutual Information(GCMI) estimator[77, 78], which uses the concept of statistical copula to give a Gaussian parametric estimation of mutual information for variables with any marginal distributions. $I(\boldsymbol{r})$ then can be decomposed into the sum of the entropy of each marginal distribution $I(r_i)$ and the joint mutual information of the random variable $\text{MI}(\boldsymbol{r})$, thus

$$I(\boldsymbol{r})=\sum_i^M I(r_i)-\text{MI}(\boldsymbol{r}), \tag{S38}$$

where $I(r_i)$ is estimated by KDP algorithm provided in the *Information theoretical estimators toolbox* [79].

F. Genetic algorithm for optimization

We use the covariance matrix adaptation-evolution strategy (CMA-ES) algorithm to do the numerical

optimization[49, 50]. First, a population of initial solutions are randomly sampled, their performance is estimated according to Eq. S38, and part of the best "performers" are selected and new population of solutions are generated according to their statistics. The iteration keeps going until the solution converges. Since this method does not guarantee global optimal solution, for each combination of different parameters N,M,n and σ_c, we perform simulation from random starting points, and the results are robust with different initial points (similar values of final target function). In all our simulations, the algorithm runs for 5000 iterations to ensure convergence.

G. Introducing inhibitory response

In the case where both odor-evoked excitation and inhibition exist, the sensitivity matrix W can be divided into two parts, W^A and W^I. Since the target function $H(\boldsymbol{r})$ does not continuously change as the values of W_{ij} go from positive to negative values, in stead of directly searching in the W-space, we search solutions in the "normalized response space". Thus, the elements of W_{ij} is mapped to a quantity f_{ij} which we refer to as a "standard response", based on whether it is excitatory or inhibitory,

$$f_{ij}^{ex} = r_{\max}[1 + \gamma \frac{1}{1 + W_{ij}^A}],$$
$$f_{ij}^{in} = r_{\max}[1 + \gamma(1 + W_{ij}^I)],$$

where γ is the constant determining the basal activity of ORN. In this way, at each iteration, we update f_{ij} continuously and W_{ij} can be derived from inversely transforming f_{ij} at the final stage.

S5. DOWNSTREAM DECODING TASKS

The potential advantage of having maximum entropy coding at ORNs are demonstrated using two tasks: reconstruction and classification. For a set of parameters (N, M, n, σ_c), we have the optimal sensitivity matrix W^*, whose nonzero elements follow log-normal distribution. The coding scheme is changed by only changing the sparsity of W without changing the distribution of sensitive elements, i.e., they are from the same distribution as in W^*. By doing so, we can control the coding capacity by varying ρ_w.

A. Reconstruction task

We use a multilayer artificial neural network (ANN) as the decoder to reconstruct odor vectors. During the training phase, 10^6 sparse odor stimuli are generated from $P_{\text{env}}(\boldsymbol{c})$, and the activity patterns of ORNs they elicit are calculated by Eq. 1, which are the inputs of the first layer of the decoding network. The decoder infers both the composition and intensity of odors, so the cost function consists of two parts, the "identity error" and the "intensity error" (Eq. 8 in main text). After training, the performance of ANN is tested using a new set of odors that are drawn from the same distribution as used in the training phase. We train the network with two different settings.

In the first scenario, the first layer weight matrix (corresponding to our receptor-odorant sensitivity matrix) is fixed during training phase. The fixed weight matrix is controlled only by its sparsity, i.e, ρ_w, its nonzero entries are drawn from a log-normal distribution with the same mean and variance as our optimized sensitivity matrix. We then compare how the loss function of the trained network changes w.r.t. ρ_w.

In the second scenario, we allow the sensitivity matrix to be "plastic", i.e., it updates during training process. Intuitively, allowing the weights in the first layer to be updated is equivalent to adding one more layer to the neural network, thus it should improve the performance. Surprisingly, with "plastic" encoder the decoder even performs slightly worse. This result indicates that with sparse, wide range signals and highly nonlinear "sensors", adaptive receptor sensitivities do not necessarily outperform fixed receptor weights (hard-wired), providing a plausible explanation of the stereotyped signaling processing at olfactory periphery.

B. Classification task

Odors are presented as clusters whose centroids are sampled from $P_{\text{env}}(\boldsymbol{c})$, members in centroids represent variations. In each simulation, each of the 100 centroids (clusters) is randomly assigned with a valence: attraction or aversion and each cluster contains 50 variations with the same label as the corresponding centroid, resulting in an ensemble of 5000 stimuli, which are randomly divided into training set (80%) and testing set (20%). The representations at ORNs are used as the input of the single hidden layer decoder, 500 neurons in hidden layer (KC) and single output neuron (MBON) are used. To model the sparse KC response observed in experiments[53, 55], numbers of connection from ORNs/PN to KC are randomly drawn from a binomial distribution, resulting on average 7 inputs from random ORNs/PNs to a KC (if $M < 7$, the maximum number of connection is M). The strength of connections are in the range $[0, 1]$, which are randomly drawn from the truncated Gaussian distribution $\mathcal{N}(0.5, 0.25)$. To further encapsulate the global inhibition from APL to KCs[54], following[80, 81], the input of ith KC evoked by an ORNs/PNs response vector $\vec{r}$ is $h_i = \langle \vec{J}_i, (\vec{r} - \langle \vec{\mu}, \vec{r} \rangle \vec{\mu}) \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner-product, $\vec{J}_i$ is the weight vector from ORNs/PN to KC, and μ is the average ONRs/PNs response vector over all odors. The activity of each KCs is a rectified linear function of its input $\Theta(h_i - \theta)$, where $\Theta(x) = x$ if $x > 0$, and

$\Theta(x) = 0$ otherwise. The threshold is chosen such that on average 10% KCs response to each odor, matching the *in vivo* experimental data[55, 56].

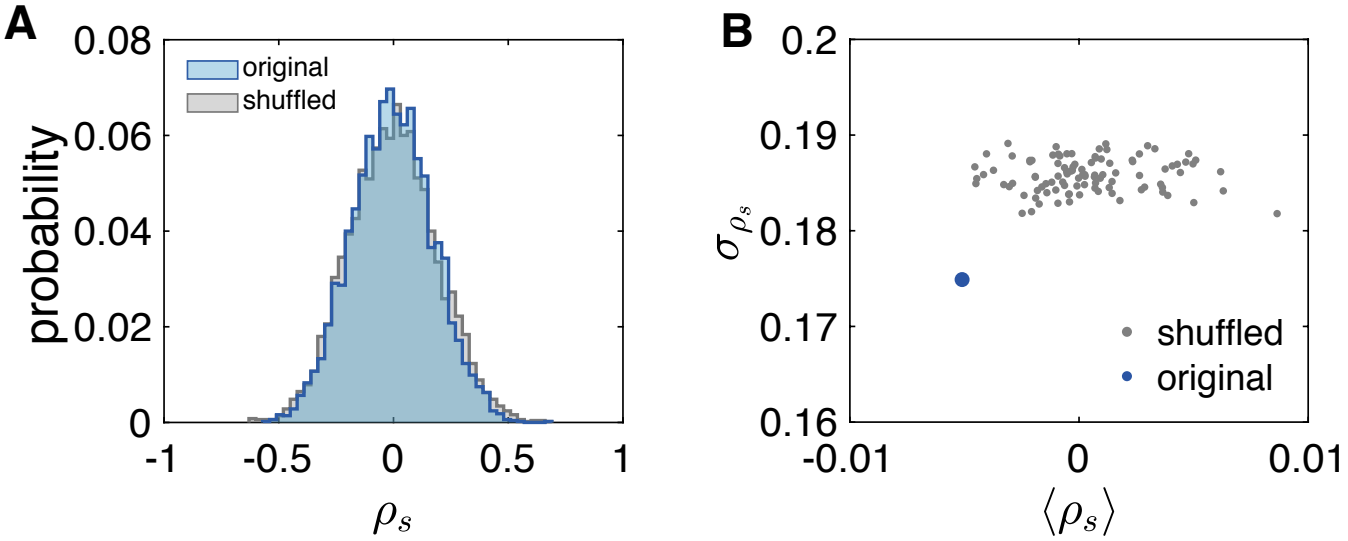

FIG. S1. (A) Histogram of column-wise (odor-wise) Spearman rank correlation coefficient ρ_s of the optimal matrix (blue, the matrix in Fig. 2 in main text) and its randomly shuffled matrix (blue). (B) Comparison of the average and standard deviation of ρ_s for the original optimal sensitivity matrix (blue dot) and 100 randomly shuffled matrices (gray dots).

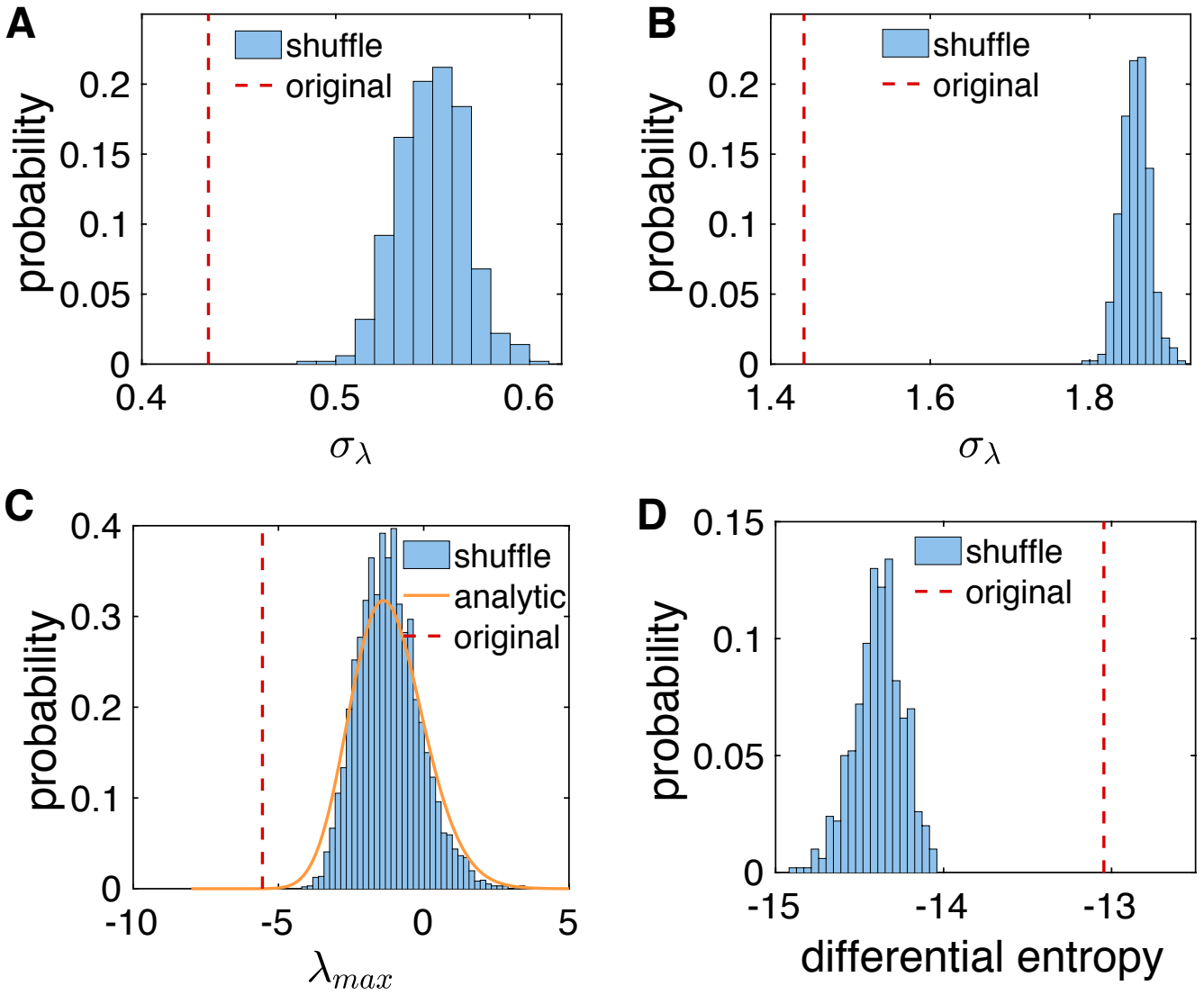

FIG. S2. Optimal sensitivity matrix is more orthogonal (see Fig. 2A) than random matrix. (A) Comparison of the eigenvalues of row-wise Spearman rank correlation matrix. Histogram is the standard deviation of eigenvalues from 500 randomly shuffled matrices, red dashed line corresponds to the optimal sensitivity matrix. Notice that, as a property of correlation matrix, the average of eigenvalues is 1 for both shuffled and original correlation matrix. (B) Similar as in A, column-wise, only considering the eigenvalues that are significant larger than 0. Notice that for the parameter used here: $N = 100, M = 30$, according to random matrix theory, there should be a fraction of $1 - (N - M)/N = 0.3$ nonzero eigenvalues of the column-wise correlation matrix. (C) Eigenvalues of row-wise inner product matrix. In order to compare with random matrix, the entries of optimal sensitivity matrix W_{ij} were first transformed into variable $\hat{W}_{ij}$ that follow standard Gaussian distribution, i.e., $\hat{W}_{ij} \sim \mathcal{N}(0, 1)$. Histogram of the largest eigenvalue of $\hat{W}\hat{W}^T$, 1000 $\hat{W}$ were generated through randomly shuffling the original transformed matrix. Solid line is the analytical result for large random matrix Eq. S13. (D) Comparison of the differential entropy of ORN response $I(\boldsymbol{r})$ with optimal sensitivity matrix (dashed line) and with randomly shuffled matrix (histogram). Parameters: $N = 100, M = 30, n = 2, \sigma_c = 2$.

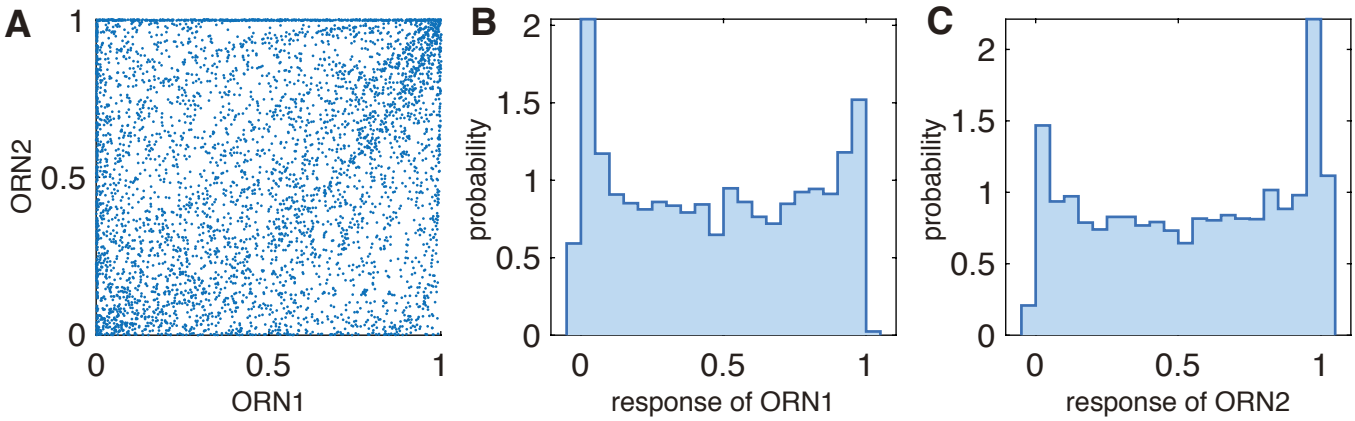

FIG. S3. Optimal receptor-odor sensitivity matrix ensures near uniformly distributed and decorrelated ORN response patterns. (A) Scatter plot of responses of two ORNs to a sample of 10^4 odor stimuli drawn from $P_{\text{env}}(\boldsymbol{c})$. (B) and (C) histogram of corresponding ORN response showing roughly uniform distribution in the whole response range. Parameters are: $N = 50, M = 13, n = 3, \sigma_c = 2$.

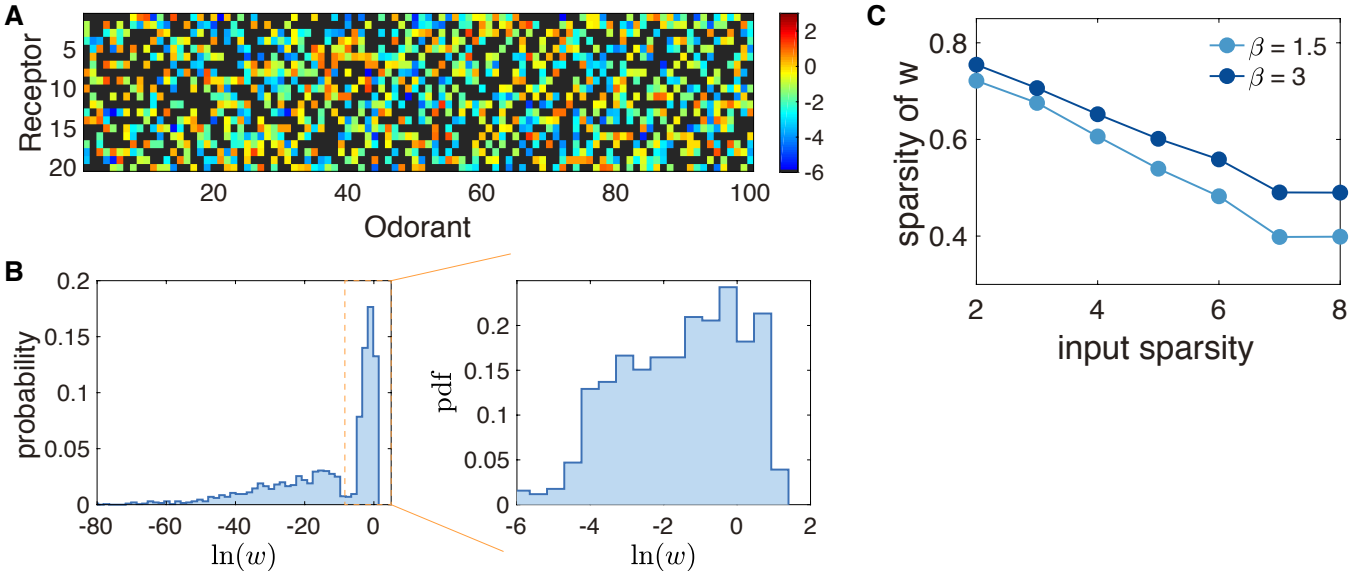

FIG. S4. Optimal sensitivity matrix for power law distributed odor concentration, i.e., $p(c) \sim c^{-\beta}$. (A) Heatmap of a typical optimal matrix W, color indicates "sensitive" elements and black indicates "insensitive" elements. (B) Histogram of all the W_{ij} values. It shows a bimodal distribution: an insensitive part with essentially zero W_{ij} (notice the log scale) and a sensitive part with nonzero W_{ij}. (C)The sparsity of optimal W with respect to different odor complexity (number of odorants in mixture). Parameters: $N = 100$,$M = 20$,$n = 5$,$\sigma_c = 2$.

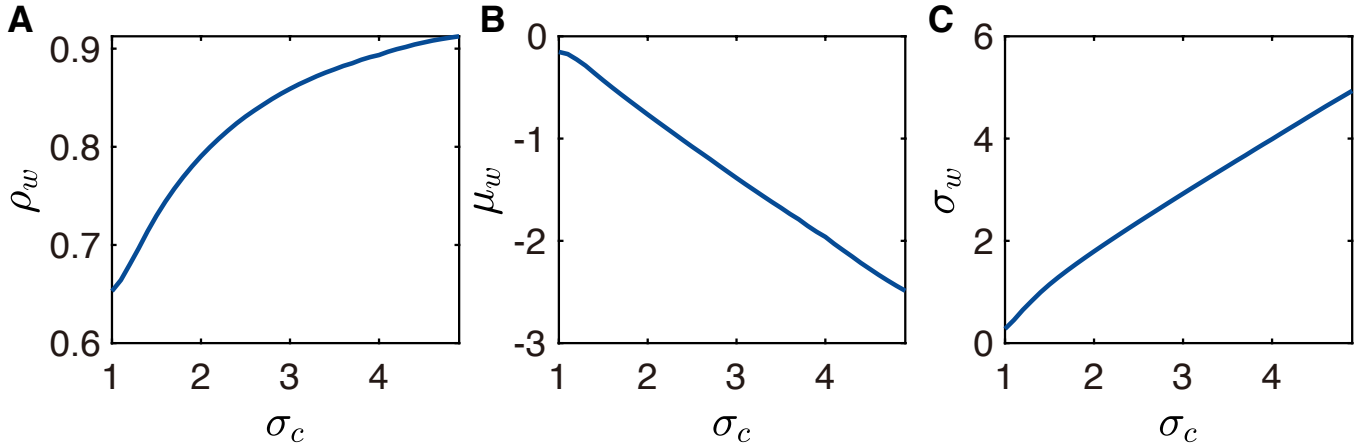

FIG. S5. Statistics of the optimal sensitivity matrix from the mean-field theory with two odorants in mixtures. The sparsity (A), average (B) and standard deviation (C) of optimal W (log scale) versus the standard deviation of log-concentration. Note the qualitatively similar trends as in Fig. 3A.

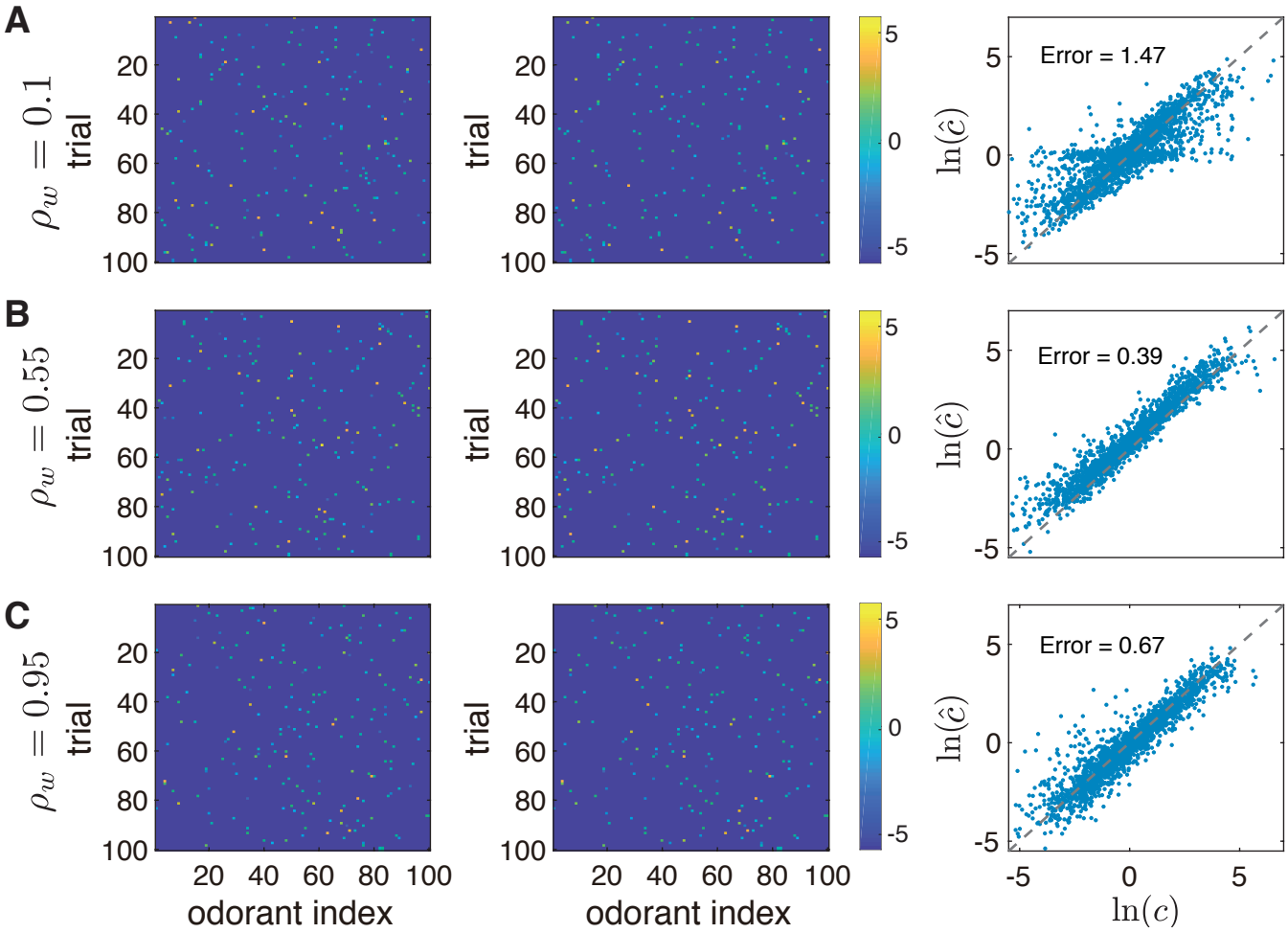

FIG. S6. Comparison of the reconstruction error when the mapping from odor space to ORN space only differ in the sparsity of sensitivity matrix W (with fixed average and standard deviation of nonzero entries, $\mu_w = -1$, $\sigma_w = 2$). Odor stimuli were randomly generated from $P_{\text{env}}(\boldsymbol{c})$. Left: 100 example sparse vectors $\boldsymbol{c}$ whose nonzero entries are coded with color (logscale). Middle: the reconstructed sparse vector $\hat{c}$. Right: scatter plot of the original and reconstructed nonzero elements. Best performance is achieved when $\rho_w = 0.55$. Parameters: $N = 100, M = 20, n = 2, \sigma_c = 2, \sigma_0 = 0.1$.

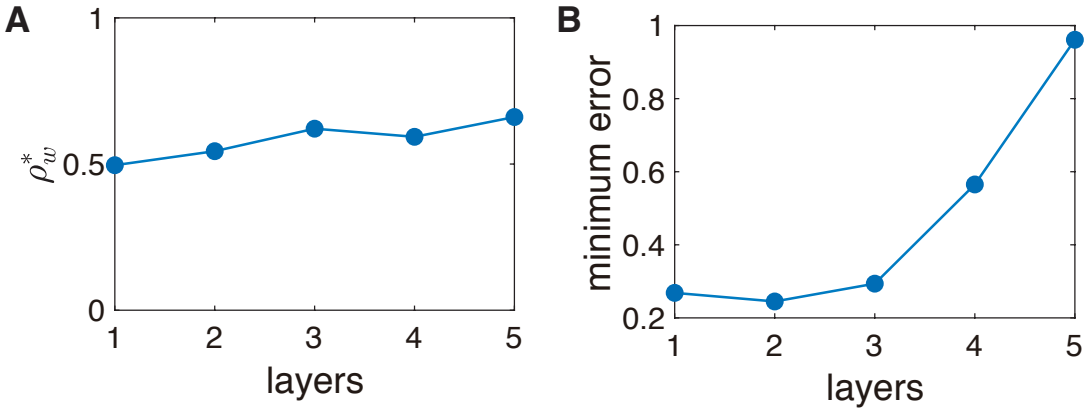

FIG. S7. Performance of reconstruction task with different number of hidden layers. (A) Sparsity of sensitivity matrix that achieves best performance v.s the number of hidden layers in the reconstruction network. (B) Minimum reconstruction error with respect to the number of hidden layers. Parameters: $N = 100$,$M = 20$,$n = 2$,$\sigma_c = 2$, $\sigma_0 = 0.05$ and each hidden layer contains 100 units.

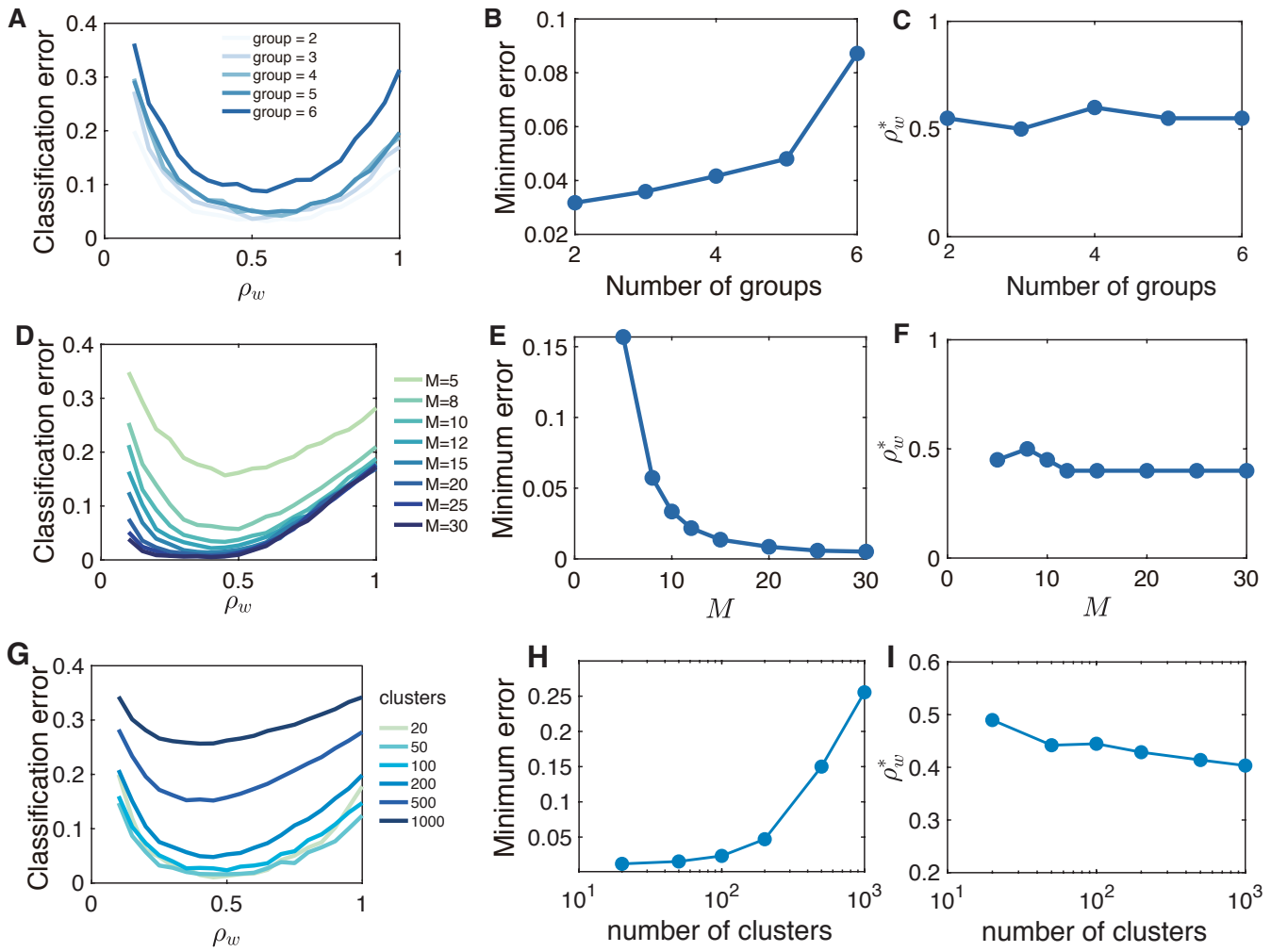

FIG. S8. Classification performance versus the sparsity of the receptor-odor sensitivity matrix with different parameters of the decoder. (A-C) Classification performance for different number of groups (labels) of odor stimuli. (A) Classification error (fraction of mis-classification) v.s sparsity of W. Colors indicate different number of groups that odor stimuli have been assigned. (B) Minimum classification error with respect to number odor groups. (C) For different number of odor groups, optimal performance is achieved around $\rho_w = 0.5$. Parameters used: $N = 100, M = 12, n = 2, \sigma_c = 2, \Delta S = 0.1, \sigma_0 = 0.1$, and 100 clusters. (D-F) Classification performance for different number of ORNs, M. (D) Classification error v.s number of ORNs. Colors indicate different M. (E) Minimum classification error with respect to number odor ORNs. (F) For different numbers of ORN, optimal performance is achieve around $\rho_w^* = 0.5$. Parameters used: $N = 100, M = 12, n = 3, \sigma_c = 2, \Delta S = 0.1, \sigma_0 = 0.1$ and two groups of labels. (G-I) Classification performance for different number of clusters of odor stimuli. (G) Classification error v.s sparsity of W. Colors indicate different number of clusters. (H) Minimum classification error with respect to number of clusters. (I) Optimal sparsity of W v.s the number of clusters. Parameters used: $N = 100, M = 20, n = 3, \sigma_c = 2, \Delta S = 0.1, \sigma_0 = 0.1$, 100 clusters, and two groups of labels. In all these figures, line are the average classification error, error bars have been omitted to avoid clutter.