NEUROSCIENCE

A biological model of nonlinear dimensionality reduction

Kensuke Yoshida[1,2]* and Taro Toyoizumi[1,2]*

Obtaining appropriate low-dimensional representations from high-dimensional sensory inputs in an unsupervised manner is essential for straightforward downstream processing. Although nonlinear dimensionality reduction methods such as *t*-distributed stochastic neighbor embedding (*t*-SNE) have been developed, their implementation in simple biological circuits remains unclear. Here, we develop a biologically plausible dimensionality reduction algorithm compatible with *t*-SNE, which uses a simple three-layer feedforward network mimicking the *Drosophila* olfactory circuit. The proposed learning rule, described as three-factor Hebbian plasticity, is effective for datasets such as entangled rings and MNIST, comparable to *t*-SNE. We further show that the algorithm could be working in olfactory circuits in *Drosophila* by analyzing the multiple experimental data in previous studies. We lastly suggest that the algorithm is also beneficial for association learning between inputs and rewards, allowing the generalization of these associations to other inputs not yet associated with rewards.

INTRODUCTION

Our brain efficiently commands behavioral outputs based on complex sensory inputs. This process requires transformation of high-dimensional sensory inputs into concise neuronal representations useful for downstream processing. Recent advances in large-scale neuronal recording have highlighted the significance of neuronal population dynamics beyond single neuron activity (*1*–*3*). Numerous studies have reported that these population dynamics during tasks often operate within a low-dimensional manifold across various brain regions (*1*, *3*, *4*). Low-dimensional representations are expected to enhance robustness against variations and improve generalization (*5*).

In addition, it has been considered that representations of different categories might be gradually untangled through information processing in the brain (*6*, *7*). A simple metric for assessing these untangled representations is linear separability, which evaluates the feasibility of distinguishing different objects using a linear hyperplane. Linearly separable representations enable a straightforward downstream readout, while the entangled ones are more difficult to distinguish. Thus, originally entangled high-dimensional inputs are likely transformed into untangled low-dimensional representations (*4*, *8*).

Such representations have also been studied in the context of artificial neural networks. Deep neural networks are considered to solve complex tasks by acquiring useful untangled representations in a supervised manner (*4*, *9*). The dimensionality of learned representations in deep neural networks was reported to be expanded in the initial layers close to inputs and reduced in the latter layers close to outputs (*10*). These findings suggest that deep neural networks acquire untangled low-dimensional representations of input patterns to solve complex tasks. However, supervised learning usually requires a large number of labels, while the brain can learn more efficiently with fewer (or sometimes without) labels. The necessity of backpropagation for training deep neural networks also challenges their feasibility in biological systems, although some recent studies proposed alternative methods (*11*, *12*). The potential performance of simpler biological circuits than deep neural networks for obtaining good representations, especially in an unsupervised manner and without backpropagation, has not been sufficiently addressed.

Previous works have proposed that simple neural network models can implement linear dimensionality reduction methods such as principal components analysis (PCA) by a biologically plausible synaptic plasticity rule (*13*, *14*). However, a variety of more complex structures, including entanglements of data manifolds, cannot be extracted by such linear dimensionality reduction methods. To capture the manifold structure in a high-dimensional space and find latent low-dimensional structure, several nonlinear methods, such as *t*-distributed stochastic neighbor embedding (*t*-SNE) and uniform manifold approximation and projection, have been developed (*15*, *16*). Among these algorithms, the *t*-SNE is based on the relatively simple idea that the similarity matrices of the high-dimensional input and low-dimensional output representations are aimed to be closer. In addition, the *t*-SNE has been suggested to perform better than other methods, such as Sammon mapping, Isomap, and Locally Linear Embedding (*15*). Despite its simplicity, biologically plausible implementations are not known.

Here, we propose a biologically plausible algorithm, Hebbian *t*-SNE, which leverages a neural circuit structure present in the brain, uses a three-factor learning rule of synaptic plasticity (*17*, *18*), and functions effectively in situations where inputs are presented as streaming data. The Hebbian *t*-SNE performs as effectively as *t*-SNE in a simple network mimicking the olfactory circuit in *Drosophila*. The algorithm encompasses three key features: repeated presentation of inputs in random order, a global factor for comparing input and output changes over time, and a middle layer in the network composed of a large number of neurons with sparse activities. The repetitive presentation of inputs allows the network to infer the input and output similarities required for the *t*-SNE. The inferred similarities are then compared by the global factor in the model, which regulates synaptic plasticity. The middle-layer neurons enable the output representations to move independently of each other because input data are transformed into neuronal activities approximately orthogonal to each other in a high-dimensional space. This

[1]Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. [2]Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
*Corresponding author. Email: kensuke.yoshida@riken.jp (K.Y.); taro.toyoizumi@riken.jp (T.T.)

network structure is consistent with the olfactory circuit in *Drosophila*, where low-dimensional output representations are acquired in mushroom body output neurons (MBONs) (*19*) through a feedforward network with the input layer of projection neurons (PNs) and the middle layer of a large number of Kenyon cells (KCs) with sparse activities (*20*). To prove the ability of the Hebbian *t*-SNE, we first show that the Hebbian *t*-SNE works as efficiently as *t*-SNE for data in which linear transformations do not work well, such as entangled rings and Modified National Institute of Standards and Technology (MNIST) dataset. We then suggest that the Hebbian *t*-SNE might work in real biological circuits by applying it to experimental data of *Drosophila* olfactory circuits from previous studies (*21*, *22*). We lastly show that Hebbian *t*-SNE can conduct reward-association learning similarly to *Drosophila*, with generalization ability using geometric input structures.

RESULTS

Construction of the Hebbian *t*-SNE

The original *t*-SNE is based on the minimization of the Kullback-Leibler (KL) divergence between the input and output similarities, as we briefly review below. The *t*-SNE calculates the input similarity between patterns X^i and X^j as $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$, where N is the number of total input patterns and $p_{i|j}$ is the similarity of X^i with respect to X^j using the Gaussian measure, which is normalized so that the sum of $p_{i|j}$ over $i(\neq j)$ is one (Table 1). The output similarity q_{ij} between output patterns Y^i and Y^j is determined on the basis of *t*-distribution and normalized so that the sum of q_{ij} over i and j $(i \neq j)$ is one (Table 1). To minimize the KL divergence $C = \sum_j \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$ between the input and output similarities, the output representation Y^j are changed according to the negative gradient of the KL divergence $-\nabla_{Y^j} C = \sum_{i \neq j} [-4(p_{ij} - q_{ij})(1 + \| Y^j - Y^i \|^2)^{-1}(Y^j - Y^i)]$ in the *t*-SNE (*15*).

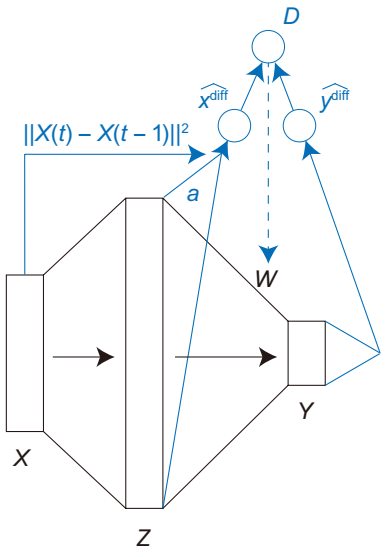
To achieve the *t*-SNE in a biologically plausible way, we considered a simple three-layer feedforward network (Fig. 1A), with the input layer X, the middle layer Z, and the output layer Y. This circuit structure is inspired by the *Drosophila* olfactory circuit, where ~50 different types of PNs project to ~2000 KCs, which then project to 34 MBONs (Fig. 1B) (*20*). The input and output layers are comprised of rate neurons equal in their numbers to the input and output dimensions, respectively. In this study, the output dimensions are fixed to two for visualization. The projection from the middle to output layers is linear, $Y = WZ$, with plastic synaptic weight matrix W. The middle layer contains rate neurons no less than the number of input patterns (or the number of patterns that need to be distinguished; see Discussion) (Fig. 1A). The patterns of middle-layer activity Z in response to different input patterns are assumed linearly independent. This enables the output representations to have the necessary flexibility for accurate *t*-SNE; without this linear independence, the output representations are restricted in a subspace with dimensions smaller than the number of input patterns, compromising the accuracy of the *t*-SNE. To achieve this constraint, we computed the middle-layer activity Z by projecting inputs X by fixed random weights to a high-dimensional space and, then, applied winner-take-all [but see an extension with k-winner-take-all in later sections, concordant with sparse KC activities in *Drosophila*; (*20*)]. This setting is consistent with the *Drosophila* olfactory circuit, where the synaptic strength from KCs to MBONs is plastic and regulated by dopaminergic neurons (DANs), while the projection from PNs to KCs is reported to be random and fixed (Fig. 1B) (*20*, *23*). Note that, although it was also recently reported that the connection from the PNs to the KCs is not random (*24*), our model remains robust to the specific form of the PN-to-KC transformation, unless the KC activities in response to different input patterns are nearly linearly dependent.

We considered a learning rule of synaptic weight W that mimics the *t*-SNE. We constructed the calculation of the input and output similarities in a biologically plausible way. A detailed mathematical derivation is presented in Materials and Methods. We assumed that the whole N input patterns $X^1, X^2, \ldots, X^N$ are provided in the random order a large number of times by repeated exposure to the same set of stimuli [or, in some systems, by the neuronal reactivation during sleep (*25*)]. The inputs are provided at each discrete time, denoted by vector $X(t) = [x_1(t), x_2(t), \ldots, x_r(t)]$ for time $t = 0,1,2, \ldots,$

Table 1. Comparison between *t*-SNE and Hebbian *t*-SNE. For the six rows from "Pair of inputs" to "Input-output comparison," each item in *t*-SNE corresponds to a counterpart in Hebbian *t*-SNE under the condition that $X(t-1) = X^i$ and $X(t) = X^j$. These correspondences allow the "Weight update" in *t*-SNE to be transformed into the form of Hebbian *t*-SNE. See Materials and Methods for a detailed description of the variables.

	t-SNE	Hebbian *t*-SNE
Pair of inputs	X^i, X^j	$X(t-1), X(t)$
Pair of outputs	Y^i, Y^j	$Y(t-1), Y(t)$
Normalized input similarity	$p_{i\|j} = \frac{\exp\left(-\|X^i - X^j\|^2/2\sigma_j^2\right)}{\sum_{i' \neq j} \exp\left(-\|X^{i'} - X^j\|^2/2\sigma_j^2\right)}$	$\widehat{x^{\text{diff}}}(t) = \sum_k \frac{x_k^{\text{diff}}(t)}{x_k^{\text{diff}}}$
Symmetric input similarity	$p_{ij} = \frac{p_{j\|i} + p_{i\|j}}{2N}$	Not needed
Output similarity	$q_{ij} = \frac{(1+\|Y^i - Y^j\|^2)^{-1}}{\sum_{i'} \sum_{i' \neq j'} (1+\|Y^{i'} - Y^{i'}\|^2)^{-1}}$	$\widehat{y^{\text{diff}}}(t) = \frac{y^{\text{diff}}(t)}{y^{\text{diff}}}$
Input-output comparison	$D_{i\|j} = -2\left(\frac{p_{i\|j}}{N} - q_{ij}\right)\left(1 + \|Y^j - Y^i\|^2\right)^{-1}$	$D(t) = -2\left[\frac{\widehat{x^{\text{diff}}}(t)}{N} - \widehat{y^{\text{diff}}}(t)\right] y^{\text{diff}}(t)$
Weight update	$\sum_j \sum_{i \neq j} \left[-4(p_{ij} - q_{ij})\left(1 + \|Y^j - Y^i\|^2\right)^{-1}(Y^j - Y^i) Z_j^T\right]$ $= \sum_j \sum_{i \neq j} D_{i\|j}(Y^j - Y^i)\left(Z_j^T - Z_i^T\right)$	$\sum_t D(t)[Y(t) - Y(t-1)]\left[Z(t)^T - Z(t-1)^T\right]$
Perplexity	$2^{-\sum_i p_{i\|j} \log_2 p_{i\|j}}$ (of pattern j)	$2^{-\sum_t a_k(t) \widehat{x^{\text{diff}}}(t) \log_2 \widehat{x^{\text{diff}}}(t)}$ (of axon k)

A



B

	Hebbian *t*-SNE	*Drosophila*
Input layer	*X*	Projection neurons (PNs)
Middle layer	*Z*	Kenyon cells (KCs)
Output layer	*Y*	Mushroom body output neurons (MBONs)
Global factor	*D*	Dopaminergic neurons (DANs)

Fig. 1. Model structure of the Hebbian *t*-SNE. (**A**) Three-layer feedforward model including the input layer *X*, the middle layer *Z*, and the output layer *Y*. The transformation from *X* to *Z* is fixed. The synaptic weight matrix *W* from the middle to output layers is plastic, regulated by the presynaptic and postsynaptic neurons and global factor *D*. The global factor *D* is determined by $\widehat{x^{\text{diff}}}$ and $\widehat{y^{\text{diff}}}$ that calculate input and output similarities, respectively. The input similarity $\widehat{x^{\text{diff}}}$ is based on the input difference $||X(t) - X(t-1)||^2$ and axonal signal *a* from the middle layer. (**B**) Hypothetical elements implementing Hebbian *t*-SNE in the *Drosophila* olfactory circuit.

which elicits the activities of the middle and output activity vectors $Z(t) = [z_1(t), z_2(t), \ldots, z_s(t)]$ and $Y(t) = [y_1(t), y_2(t)]$, respectively. In the model, the normalized input similarity $p_{i|j}$ is estimated by neuronal activity $\widehat{x^{\text{diff}}}$, which receives inputs from the middle layer and is modulated by input-difference signal $\|X(t) - X(t-1)\|^2$ (Fig. 1A). This $\widehat{x^{\text{diff}}}$ computes how close $X(t-1)$ is to $X(t)$ in the neighborhood of $X(t)$ according to the input history. Specifically, axonal activity $A(t) = [a_1(t), a_2(t), \ldots, a_s(t)]$ from the middle-layer neurons is computed so that $a_k(t)$ is set to one only at time t when the $z_k(t)$ takes the largest value among the responses to the input patterns $X^1, X^2, \ldots, X^N$ and zero at other times. This axonal output can be biologically computed, for example, with an adaptive firing threshold in each neuron (see Materials and Methods). This distinction between the middle-layer activity and the axonal activity is used to evaluate $\widehat{x^{\text{diff}}}$ that involves the normalization within the neighborhood of $X(t)$ [but not of $X(t-1)$]. The transmitter release $X^{\text{diff}}(t) = \left[x_1^{\text{diff}}(t), x_2^{\text{diff}}(t), \ldots, x_s^{\text{diff}}(t)\right]$ of these axons is modeled as $x_k^{\text{diff}}(t) = a_k(t)\exp\left[-\|X(t) - X(t-1)\|^2 / \left(2\sigma_k^2\right)\right]$ with synapse specific scaling factor σ_k. The term $\|X(t) - X(t-1)\|^2$ might correspond to presynaptic modulation by neuromodulators (*26*) or astrocytes (*27*), while the term σ_k might represent the expression level of receptors for such presynaptic modulation at each synapse. Last, the postsynaptic neuronal activity downstream of these axons is determined to be $\widehat{x^{\text{diff}}}(t) = \sum_{k=1}^{s} x_k^{\text{diff}}(t) / \overline{x_k^{\text{diff}}}$ with synaptic weight $1 / \overline{x_k^{\text{diff}}}$. The synaptic weight $1 / \overline{x_k^{\text{diff}}}$ is regulated so that the conditional mean postsynaptic activity given the activation of this synapse [i.e., $a_k(t) > 0$] is one, similar to the normalization of the *t*-SNE (Table 1 and Materials and Methods). Note that this normalization is straightforwardly computable for each input pattern because the axonal activity $a_k(t)$ takes a positive value only in response to one input pattern. The output similarity is estimated using $y^{\text{diff}}(t) = [1 + \| Y(t) - Y(t-1)\|^2]^{-1}$. Then, the term $\widehat{y^{\text{diff}}}(t)$, which is obtained by dividing $y^{\text{diff}}(t)$ by its time average $\overline{y^{\text{diff}}}$, approximates the output similarity q_{ij} of the *t*-SNE. The resulting neuronal firing rate $D(t)$ that integrates the variables $\widehat{x^{\text{diff}}}(t)$, $\widehat{y^{\text{diff}}}(t)$, and $y^{\text{diff}}(t)$ is modeled as $D(t) = -2\left[\frac{\widehat{x^{\text{diff}}}(t)}{N} - \widehat{y^{\text{diff}}}(t)\right] \cdot y^{\text{diff}}(t)$, which compares the input and output similarities (Fig. 1A and Table 1). By using this neuronal activity $D(t)$, the negative gradient of KL divergence between input and output similarities with respect to synaptic weight w_{lm} is approximated by $\sum_t D(t) \cdot [y_l(t) - y_l(t-1)] \cdot [z_m(t) - z_m(t-1)]$. This update rule is biologically plausible. It is given by the product of the presynaptic component $z_m(t) - z_m(t-1)$, the postsynaptic component $y_l(t) - y_l(t-1)$, and the global factor $D(t)$. Using this gradient, the synaptic changes are computed by mini-batch and Adam (*28*). We call this learning rule Hebbian *t*-SNE. Note that mini-batch and Adam are considered biologically plausible because they use time-average values of the gradient and its square, which can be computable locally at individual synapses. The model determines the scaling factor σ_k for axon k similarly to the *t*-SNE. The *t*-SNE selects σ_j of pattern j such that the perplexity $2^{-\sum_i p_{i|j} \log_2 p_{i|j}}$ matches the target value determined by the user. The perplexity value, typically set between 5 and 50, determines the number of neighboring points substantially reflected in measuring the input similarity. Mimicking this method, we slowly change σ_k to bring $2^{-\sum_t a_k(t)\widehat{x^{\text{diff}}}(t)\log_2 \widehat{x^{\text{diff}}}(t)}$ close to the target perplexity. Here, the axonal activity $a_k(t)$ takes a positive value if and only if $X(t) = X^j$ (see Table 1 and Materials and Methods). We let the synaptic weights change after an initial run of 500 steps when the perplexity approached sufficiently close to the target value (Figs. 2E and 3E). We also considered the model in which the calculation of $\widehat{x^{\text{diff}}}(t)$ is simplified so that it receives inputs only from the input layer, which we call simplified Hebbian *t*-SNE (fig. S1).

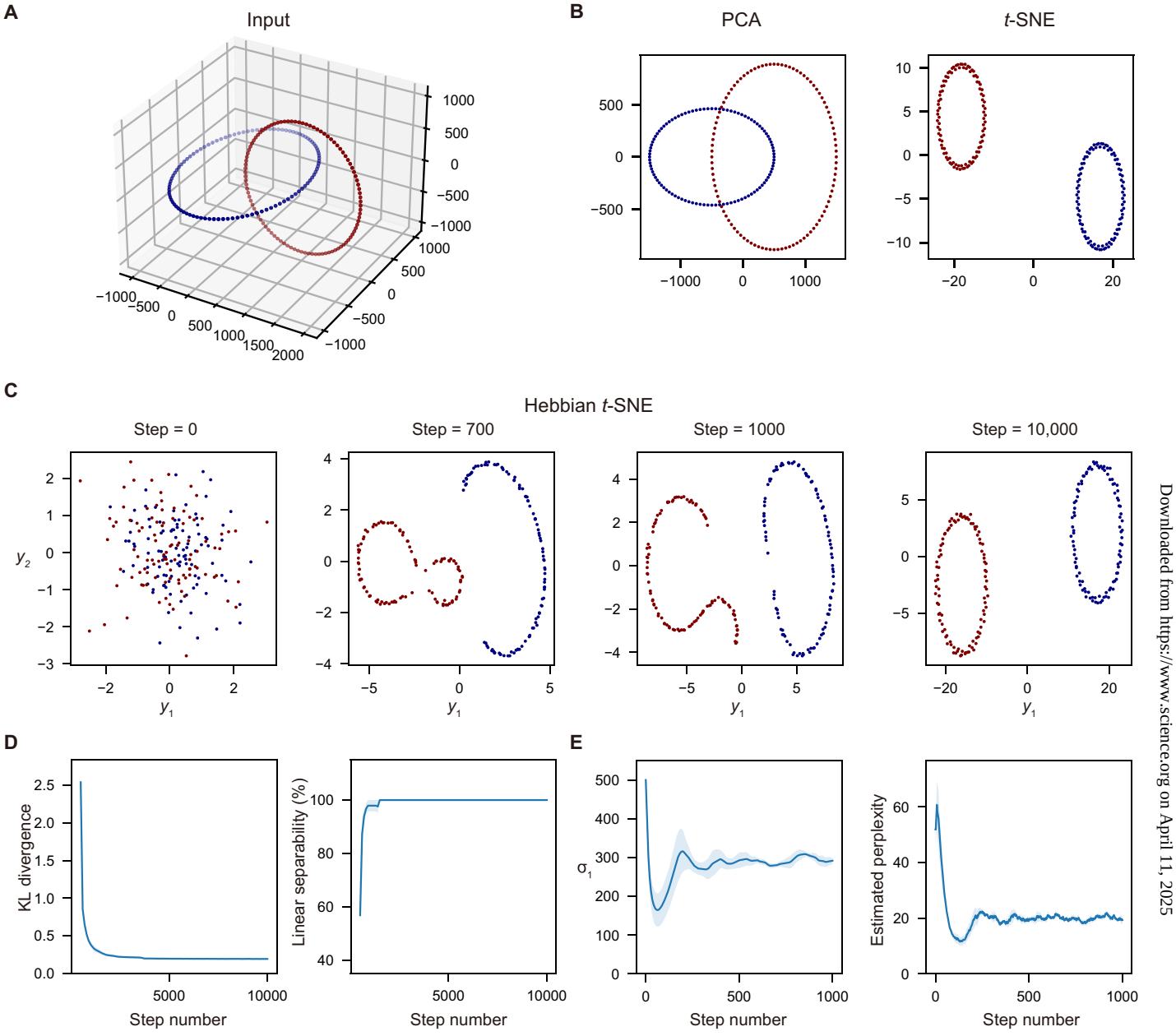


Fig. 2. Applying the Hebbian *t*-SNE to inputs distributed according to entangled rings. (**A**) Input data. Red and blue dots represent inputs corresponding to each ring. (**B**) The final two-dimensional output representations obtained by the PCA and *t*-SNE. While the PCA failed to separate the two rings, the *t*-SNE successfully separated the two rings. (**C**) The time course of the representation changes by the Hebbian *t*-SNE. The two rings were gradually separated along the iterative synaptic weight updates. (**D**) The changes of the KL divergence between the input and output similarity matrices and the linear separability of two rings in the output representation. The lines and shadows represent the means and SEMs in the five trials, respectively. (**E**) The changes of the σ_1 value and its related estimated perplexity in the model. The estimated perplexity gradually approached the target perplexity 20 along the changes of σ_1. The lines and shadows represent the means and SEMs in the five trials, respectively.

Therefore, we constructed the dimensionality reduction algorithm imitating *t*-SNE via a biologically plausible learning rule.

Dimensionality reduction of high-dimensional data by Hebbian *t*-SNE

To demonstrate the effectiveness of the Hebbian *t*-SNE across various datasets, we first considered the input of two entangled rings (Fig. 2A). While the linear map, such as the PCA, failed to separate entangled rings, the *t*-SNE successfully provided the separated representation of two rings (Fig. 2B). As with *t*-SNE, the Hebbian *t*-SNE succeeded in separating two rings (Fig. 2C). Because the coordinates of the output representations were initialized at random, the KL divergence between the original three-dimensional input and the two-dimensional output similarity matrices took a large value. The linear separability (see Materials and Methods for the definition) of two rings in the output representation was small at the initial condition (Fig. 2D). Following the iterative updates of the synaptic weights W, the KL divergence between the input and output similarity

matrices gradually decreased as expected (Fig. 2D) because the Hebbian *t*-SNE is constructed to reduce this value. Consequently, the two rings were properly separated and linear separability sufficiently increased (Fig. 2, C and D). In this process, the estimated perplexity rapidly approached the set value (20 in this case) in about the first 500 steps as expected (Fig. 2E). Given that the target perplexity determines the spatial scale for measuring input similarity structures, it was set to ensure that input similarity includes information about neighboring points within the same ring while excluding those in another ring. While the appropriate value for target perplexity generally depends on the number of data points and other input data characteristics, we set it between 20 and 40 throughout this manuscript. For the entangled rings, the simplified Hebbian *t*-SNE also succeeded in separating two rings as well as the Hebbian *t*-SNE (fig. S2). Therefore, the Hebbian *t*-SNE can be further simplified in this easy case.

To evaluate the efficacy of the Hebbian *t*-SNE for more complex higher-dimensional data, we considered the MNIST data (Fig. 3A) (*29*). As with the entangled rings, the *t*-SNE but not the PCA provided the low-dimensional representation separating the clusters of different numbers (Fig. 3B). This can be quantitatively assessed using the linear separability, defined as an accuracy rate of solving a multi-class classification problem with the linear support vector machine in a one-versus-one scheme, which was larger in the representation by *t*-SNE (71%) than in that by PCA (44%). Similar to the case of the entangled rings, following the iterative synaptic weight updates, the Hebbian

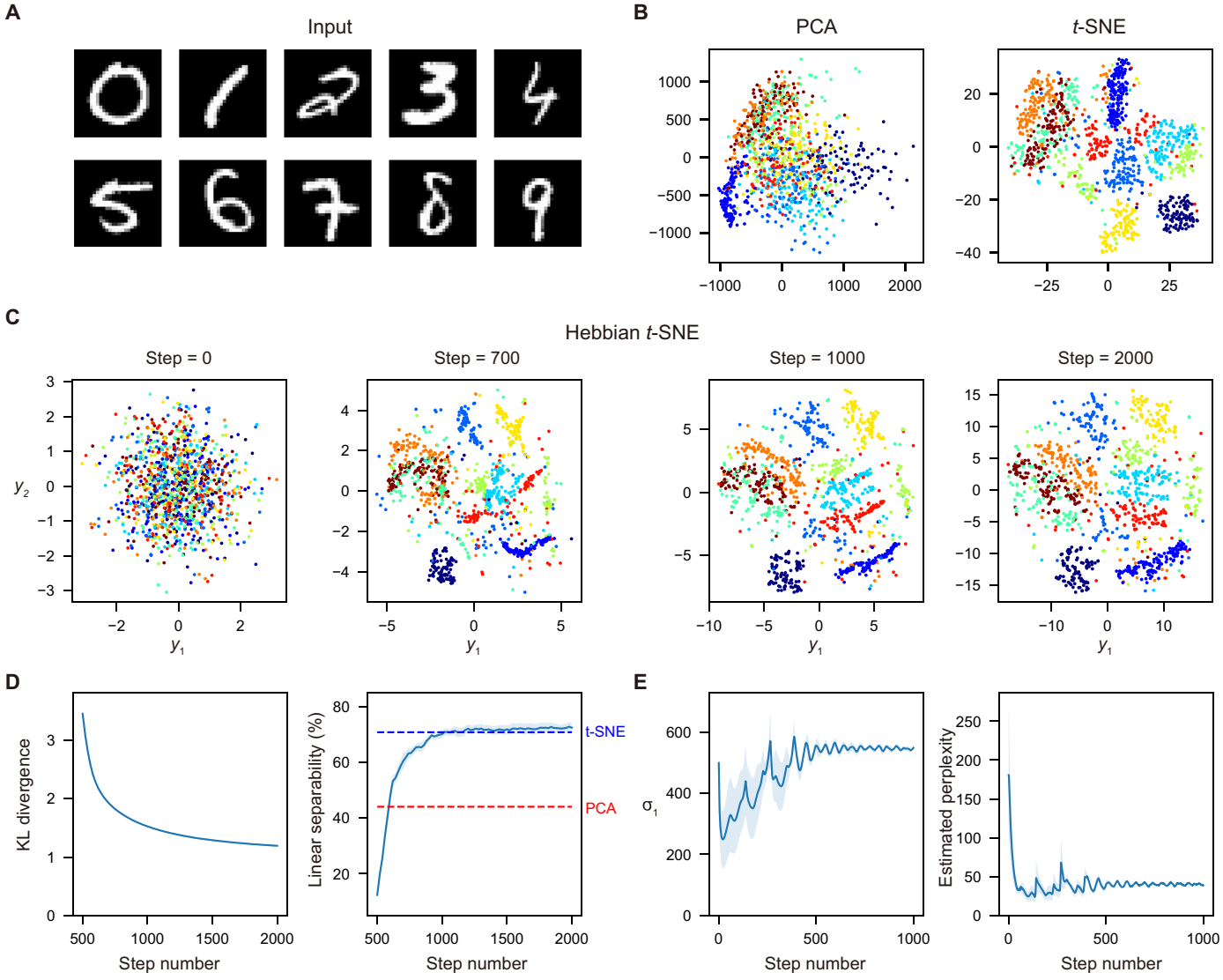


Fig. 3. Applying the Hebbian *t*-SNE to the MNIST data. (**A**) Sample MNIST data. (**B**) The two-dimensional representations by the PCA and *t*-SNE. Each color of the data represents a digit from zero to nine. While the PCA failed to find hidden representations, the *t*-SNE successfully separated inputs of different digits. (**C**) The time course of the changes of the representation by the Hebbian *t*-SNE. The Hebbian *t*-SNE gradually led to good representations separating different digits. (**D**) The changes in the KL divergence between the input and output similarity matrices and the linear separability of the 10 digits in the output representation. Learning starts at step 500. The lines and shadows represent the means and SEMs in the five trials, respectively. (**E**) The changes of the σ_1 value and its related estimated perplexity in the model. The estimated perplexity gradually approached the target perplexity 40 along the changes of σ_1. The lines and shadows represent the means and SEMs in the five trials, respectively.

t-SNE provided a low-dimensional representation that distinctly separated the clusters of different numbers (Fig. 3C). More quantitatively, the linear separability rapidly surpassed the PCA level and reached the *t*-SNE level with learning (Fig. 3D). The estimated perplexity rapidly approached the set value of 40 (Fig. 3E). We also evaluated the representations by the adjusted Rand index, which measures the similarity between the 10 clusters of different digits and the 10 clusters obtained by applying *K*-means clustering on each two-dimensional representation (see Materials and Methods). When evaluated using the adjusted Rand index, Hebbian *t*-SNE, like *t*-SNE, produced more well-separated representations of different digits than the PCA (fig. S3D). In addition, the Hebbian *t*-SNE outperformed other algorithms reported to be biologically plausible, such as self-organizing maps (SOMs) (*30*, *31*) and *K*-means clustering (*32*) applied to the original high-dimensional data, when evaluated by the metrics described above (fig. S3). In contrast, the simplified Hebbian *t*-SNE generated separated representations only in a limited number of data points, failing to develop good representations (fig. S4). The cluster representing each number was hard to find in the final representation (fig. S4A). Consequently, the linear separability was lower than the Hebbian *t*-SNE (fig. S4B). Hence, the Hebbian *t*-SNE, not the simplified *t*-SNE, obtained a good representation of the MNIST data. Note that, although the simplified Hebbian *t*-SNE was ineffective for complex data such as the MNIST, it might still be adapted in simple biological circuits.

Hebbian *t*-SNE might be working in the olfactory circuits in *Drosophila*

Here, we examined the potential role of the Hebbian *t*-SNE in biological circuits. The Hebbian *t*-SNE requires the middle layer that is composed of a large number of neurons with sparse neuronal activity. This structure is observed in the olfactory circuit in *Drosophila*, comprising PNs, KCs, and MBONs (*20*). Thus, we modeled the PNs, KCs, and MBONs as the input layer *X*, middle layer *Z*, and output layer *Y*, respectively (Fig. 1B), and investigated whether the Hebbian *t*-SNE might work during olfactory learning in *Drosophila*.

We first analyzed the responses of 24 olfactory receptors to 110 odors, which are classified into 10 functional groups on the basis of the chemical structures (*21*). Considering that chemically related odors are reported to elicit similar neuronal representations and perceptions (*33*–*35*), we evaluated whether the Hebbian *t*-SNE can find the representation reflecting this chemical classification. As olfactory receptor neurons expressing the same olfactory receptor transmit signals to identical PNs (*36*–*38*), we used these receptors' responses as surrogates of PN activities. First, the representations applying the PCA to the olfactory receptor responses crudely distinguished different chemical functional groups (Fig. 4A). In *Drosophila*, the transformation from the PN to KC activities is reported to be described as a combination of the random projection and lateral inhibition (*39*, *40*). Therefore, we estimated the KC activities

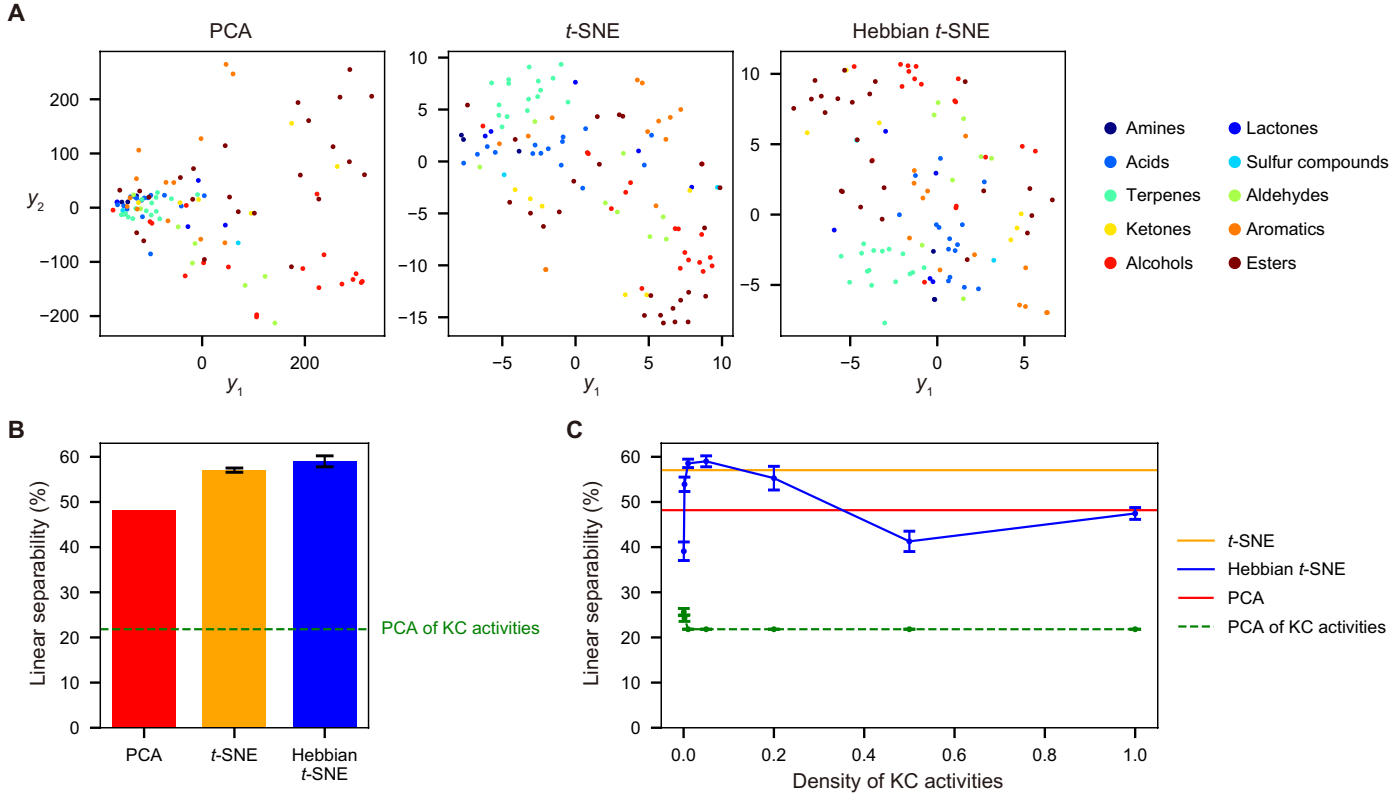


Fig. 4. The Hebbian *t*-SNE obtained a good chemical representation behind olfactory response in *Drosophila*. (**A**) The representations obtained by applying the PCA (left), *t*-SNE (middle), and Hebbian *t*-SNE (right) to olfactory receptor responses. The colors indicate the 10 functional groups (amines, lactones, acids, sulfur compounds, terpenes, aldehydes, ketones, aromatics, alcohols, and esters) based on the chemical structures. (**B**) The linear separability of the 10 functional groups in the two-dimensional representations obtained by the PCA, *t*-SNE, and Hebbian *t*-SNE. The error bar indicates SEMs in the 100 and 10 trials in the case of *t*-SNE and Hebbian *t*-SNE, respectively. The green dotted line indicates the linear separability obtained by applying the PCA to the modeled KC activities. (**C**) The linear separability of the 10 functional groups based on the two-dimensional output representations as the density of KC activities was varied. The error bars indicate SEMs in the 10 trials.

by applying random projection, thresholding, and normalization for the PN activities (see Materials and Methods), consistent with previous studies (*39*, *41*–*43*). We conducted Hebbian *t*-SNE using these estimated KC activities as the activities of the middle layer. As a result, the Hebbian *t*-SNE obtained representations distinguishing the different chemical functional groups better than the PCA of olfactory receptor responses (Fig. 4A). The linear separability of different chemical classes was higher in the Hebbian *t*-SNE (59%) than that in the PCA (48%), and it was comparable to that achieved by the *t*-SNE (57%) (Fig. 4B). This good representation by the Hebbian *t*-SNE was observed when the density of the KC activities was low (Fig. 4C). However, excessively sparse KC activities hindered proper discrimination because the KC activities for some inputs completely overlapped in this case. With KC representations that are too densely overlapping, independent learning of the output response to each input became impossible. Therefore, the intermediate density, including the experimentally observed level of 5% (*20*), was optimal for linear separability across different patterns. Note that the representations by applying the PCA to the modeled KC activities were unsuccessful because the input structure was distorted by the random transformation from PN to KC activities (Fig. 4, B and C). These results were consistent with the evaluation by the adjusted Rand index (fig. S5D; see Materials and Methods) and were robust to changes in the perplexity value (fig. S6). In addition, the representation by the Hebbian *t*-SNE outperformed the SOM and *K*-means clustering (fig. S5) as well as the MNIST case (fig. S3). The SOM was comparable in linear separability but inferior in adjusted Rand index to Hebbian *t*-SNE (fig. S5, A, C, and D). Applying *K*-means clustering to olfactory receptor responses did not reveal distinct clusters of chemical classes (fig. S5, B and D), likely due to the high dimensionality of the input data (*44*). Hence, the Hebbian *t*-SNE effectively captures the chemical structure behind high-dimensional olfactory sensory inputs, which implies its benefit for downstream olfactory processing.

Next, we investigated whether the representation by the Hebbian *t*-SNE can extract valence, which is expressed in the MBON activities of *Drosophila* (*19*). For this purpose, we analyzed the experimental data from (*22*), where the responses of 37 types of PNs to 84 odors were recorded together with the valence index, which evaluates how much flies prefer the odor by monitoring their behavioral responses. While this study did not record the activities of MBONs, the valence index is expected to be encoded in MBON activities because the previous study (*19*) found that odorants sharing similar valence index were clustered within the low-dimensional representation of MBON activities. Hence, we analyzed whether the Hebbian *t*-SNE could obtain a representation consistent with this valence index. We applied the Hebbian *t*-SNE to the responses of 37 types of PNs and compared the result with the PCA representations of the PN activities and the estimated KC activities, computed as in Fig. 4 (Fig. 5A). While the representations were only weakly

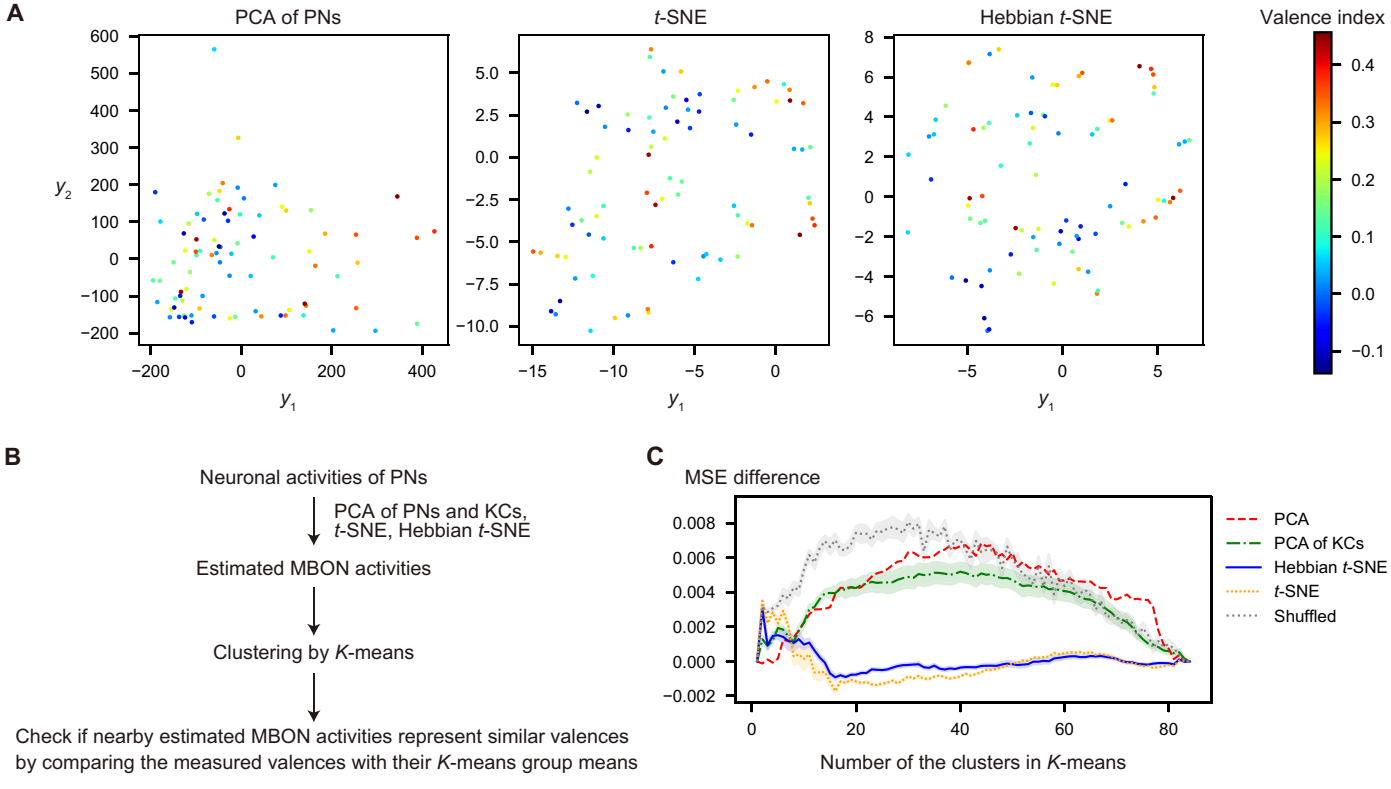


Fig. 5. The Hebbian *t*-SNE explained valence representation in *Drosophila* better than the PCA. (**A**) The two-dimensional representations by applying the PCA (left), *t*-SNE (middle), and Hebbian *t*-SNE (right) to the PN activities. The colors indicate the valence index. (**B**) The procedure of the analysis. (**C**) The mean squared errors (MSEs) between the valence index of each data point and its neighborhood cluster mean for varying numbers of *K*-means clusters. The MSE differences from the original PN activities are shown (see Materials and Methods). The representations were obtained by applying the PCA to the PN activities (red), PCA to the modeled KC activities (green), *t*-SNE to the PN activities (orange), and Hebbian *t*-SNE to the PN activities (blue). The gray line indicates the result of Hebbian *t*-SNE but after random shuffling of the valence index. The lines and shadows represent the means and SEMs in the 10 trials, respectively.

correlated with the valence index when evaluated in the whole range (fig. S7A), nearby points tended to have a more similar valence index in the representation obtained by the Hebbian *t*-SNE than the PCA representations of the PN and modeled KC activities. For an objective assessment, we performed the *K*-means clustering on each two-dimensional representation. We compared the representations by the mean squared error (MSE) between experimentally measured valence index values and their group mean within each cluster (Fig. 5B; see Materials and Methods). The representation by the Hebbian *t*-SNE was locally more homogeneous in the valence index than the PCA representations of the PN and modeled KC activities, and it was comparable to the *t*-SNE representation (Fig. 5C). In addition, the representation by the Hebbian *t*-SNE exhibited a lower MSE than those by applying the SOM and *K*-means clustering to the PN activities (fig. S7, B and C). Given that odors with similar valence tend to cluster in the MBON representation (*19*), this result suggests that the Hebbian *t*-SNE might be the most plausible computation conducted in this circuit among these methods. The outcomes were robust to different choices of the perplexity value (fig. S8). These results indicate the possibility that nonlinear dimensionality reduction, such as the Hebbian *t*-SNE, can be implemented in *Drosophila* olfactory circuits.

Hebbian *t*-SNE induces association learning with generalization

The MBON representations are not formed entirely in an unsupervised manner. In *Drosophila*, the association between odors and rewards is detected by DANs, whose activities promote synaptic plasticity from KCs to MBONs (*19*, *45*, *46*). In addition, a recent study reported that DANs are also activated by the innate value of each odor (*47*). Thus, we explored the collaborative role of the Hebbian *t*-SNE in conjunction with partially available rewards and the innate value of odors. For simplicity, we subsequently use the term "rewards" to denote the effects of both the externally provided rewards and the innate value of odors. Mimicking the synaptic plasticity mediated by DANs, we considered a simple rule in which the reward associated with odors induces modification of synaptic weight from the middle layer to a subset of the output neurons (y_1 in the case below) in addition to synaptic changes by the Hebbian *t*-SNE. This corresponds to adding a reward-modulation term in the objective function of the *t*-SNE (see Materials and Methods). Thus, the total synaptic changes are computed using $\Delta w_{lm}^{\mathrm{Hebb}} + \sum_t R(X(t)) \cdot \delta_{l1} \cdot z_m(t)$, where $\Delta w_{lm}^{\mathrm{Hebb}}$ is a synaptic change by the Hebbian *t*-SNE, $R(X(t))$ represents the reward associated with the input $X(t)$, δ_{l1} is Kronecker delta function that restricts changes only in synapses to y_1, and the $z_m(t)$ factor induces synaptic changes only when presynaptic neuron m is active (see Materials and Methods for details). Note that biologically, positive associations are often implemented by the depression of synapses onto the MBONs associated with avoidance behavior in *Drosophila* (*20*, *45*, *48*). These neurons' activity might be aligned with the negative-y_1 direction in our model. Given that the MBON activities predict attraction and repulsion behaviors to odors (*49*), we investigated whether this reward-modulated Hebbian *t*-SNE could obtain the representation separating positively and negatively rewarded inputs.

We considered input data composed of four entangled rings (Fig. 6A). We applied the reward-modulated Hebbian *t*-SNE in the case that positive and negative rewards were given to a subset of inputs within each ring (Fig. 6A). Two rings were associated with positive rewards, and the others were associated with negative rewards. As a result, the reward-modulated Hebbian *t*-SNE separated the positive-rewarded and negative-rewarded rings even when only most of the inputs were paired with reward signals (Fig. 6B). When the reward proportion was set to 0.1, the reward-modulated Hebbian *t*-SNE obtained a representation similar to the case with reward proportion 1.0 (Fig. 6B). This is qualitatively evaluated using a reward separability, which assesses the separability of the rings associated with positive and negative rewards based on their values of y_1 (Fig. 6C; see Materials and Methods for a definition). The reward separability improved substantially even with a small reward proportion ~0.1 (Fig. 6D). Thus, the reward signal was generalized to the inputs that were not rewarded but belonged to the same ring as rewarded inputs. This is consistent with the olfactory learning in *Drosophila* in which avoidance is generalized among similar odors (*50*).

To investigate the generalization ability further, we compared the reward-modulated Hebbian *t*-SNE with linear and kernel perceptrons. Systematically changing the reward proportion suggested that the Hebbian *t*-SNE had a higher generalization ability than these perceptrons (Fig. 6E). The simple linear perceptron, given the three-dimensional inputs (see Materials and Methods), could not separate the rings associated with positive and negative rewards because they were entangled. The kernel perceptron used linear perceptron after transforming the original inputs to a high-dimensional representation using the Gaussian kernel (see Materials and Methods). This kernel perceptron had better performance than the linear perceptron. Yet, the Hebbian *t*-SNE had superior generalization ability to this method (Fig. 6E). While the kernel perceptron only incorporates the distance from each point to the rewarded points, similar to the nearest neighbor algorithm, the Hebbian *t*-SNE is expected to have a stronger generalization ability using geometric structures in the original inputs. We also considered the data along an S-shaped manifold, in which one end is positively rewarded, and the other is negatively rewarded (fig. S9). In this case, the reward-modulated Hebbian *t*-SNE also had superior generalization ability to the other two methods. Furthermore, when the S-shaped manifold had discontinuity toward the negative- or positive-rewarded data (Fig. 6F), the data around the center of the S-shaped curve acquired either positive or negative valence, respectively (Fig. 6G), reflecting the geometric continuity along the S-shaped manifold. This result suggests that the generalization by the reward-modulated Hebbian *t*-SNE depends on the distribution of whole input patterns, including the non-rewarded data. This can be experimentally tested by providing different distributions of odor mixtures of positive-rewarded and negative-rewarded odors in *Drosophila* (see Discussion). In summary, the reward-modulated Hebbian *t*-SNE facilitates association learning with generalization ability.

DISCUSSION

We showed that the Hebbian *t*-SNE, constructed by the simple feedforward neuronal network with the Hebbian synaptic plasticity learning rule, performed comparably to *t*-SNE on datasets including the entangled rings and MNIST data. We then suggested the possibility that the Hebbian *t*-SNE might be implemented for learning low-dimensional representations of the high-dimensional odor space

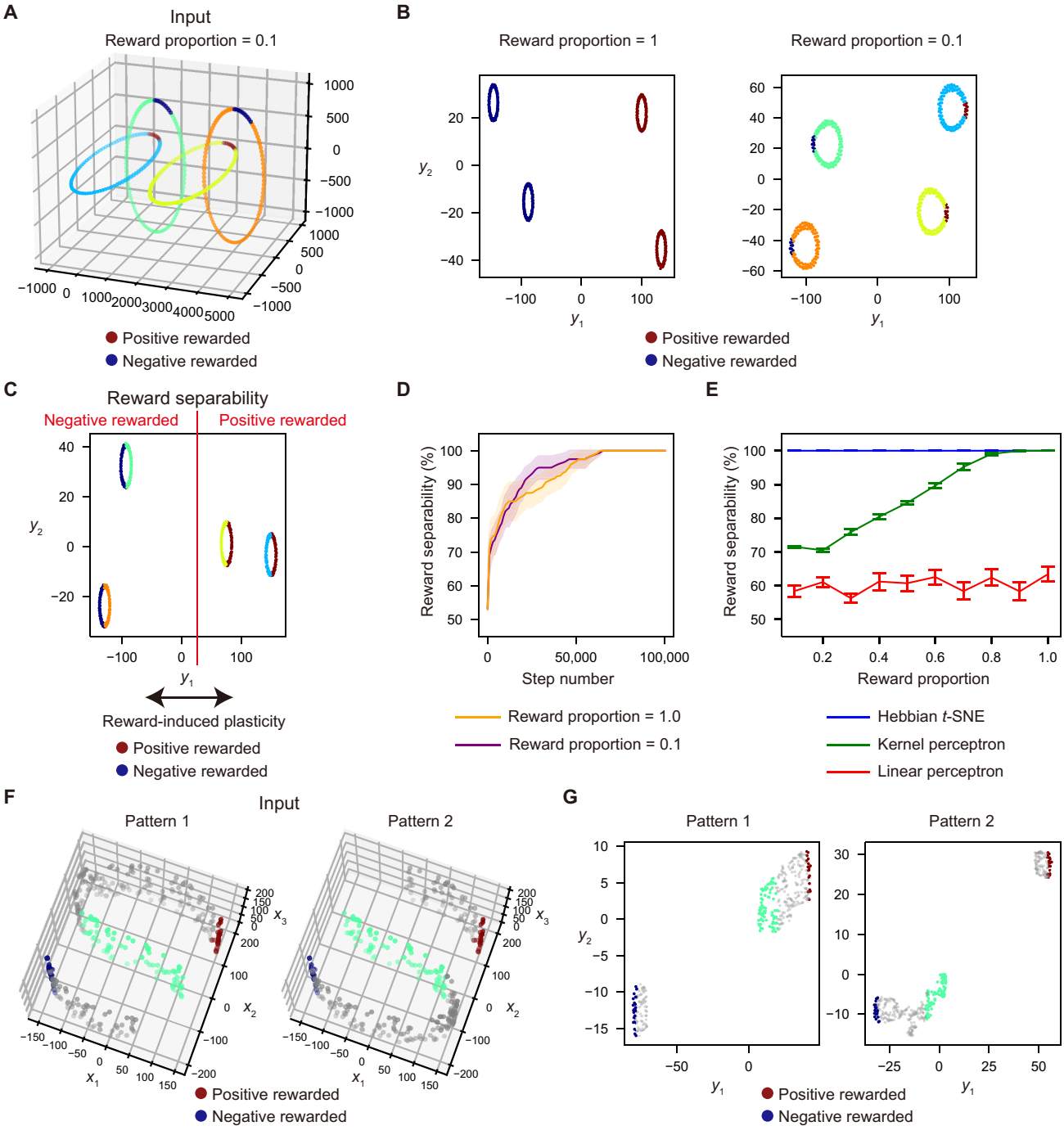


Fig. 6. Reward-modulated Hebbian *t*-SNE. (**A**) The original data of the four entangled rings in the three-dimensional space. The red and blue points indicate positive- and negative-rewarded data, respectively. The reward proportion is the proportion of the rewarded data in the total data. (**B**) The representations obtained with distinct reward proportions. The red and blue data points were biased to the positive and negative direction of the y_1 coordinate by the reward, respectively. (**C**) Reward separability is defined as the classification accuracy of whether each data point belongs to a positively or negatively rewarded ring based on output y_1 when its threshold is optimally chosen. The reward proportion was 0.5 in this example. (**D**) The time course of reward separability when the reward proportion was 1.0 (orange) and 0.1 (purple). The lines and shadows represent the means and SEMs in the 10 trials, respectively. (**E**) The reward separability of the Hebbian *t*-SNE (blue), linear perceptron for three-dimensional inputs (red), and linear perceptron after transforming inputs into high-dimensional space (green) (see Materials and Methods) for different reward proportions. The error bars represent SEMs in the 10 trials. (**F**) Three-dimensional input patterns along disconnected S-shaped sheets, a part of which is cut. The red and blue points indicate positive- and negative-rewarded data, respectively. The number of the rewarded points is equal to that in fig. S9A at reward proportion 0.1. The light green points are some neutral data points located around the center of the S-shape, colored for visualization purposes. The manifolds are cut so that the light green data are connected only to the positive-rewarded data in pattern 1 and the negative-rewarded data in pattern 2. (**G**) The representations obtained by the reward-modulated Hebbian *t*-SNE. The light green points were located near the positive- and negative-rewarded data in patterns 1 and 2, respectively.

in *Drosophila*. We further showed that the reward-modulated Hebbian *t*-SNE enabled association learning with generalization, as with the olfactory systems in *Drosophila*. These findings suggest that even simple biological circuits can conduct nonlinear dimensionality reduction and obtain untangled representations in an unsupervised way.

The proposed model is considered biologically plausible in the sense that it uses a three-factor learning rule of synaptic plasticity (*17*, *18*) in a neural circuit structure observed in the brain and operates effectively when inputs are presented as streaming data. In particular, the model is based on previous experimental findings in neuroscience, especially of the network architecture and synaptic plasticity rule in the olfactory circuit in *Drosophila*. The model proposes multiple essential characteristics, the role of which could be tested in future experiments. First, high-dimensional activities must be created by a large number of neurons in the middle layer. In *Drosophila*, KCs are experimentally reported to exhibit higher-dimensional representations of odors compared to PNs (*39*). Similar to the *Drosophila* olfactory circuit, a comparable neuronal circuit has been identified in the cerebellum, including a large number of granule cells in the middle layer (*20*, *43*). In our model, inputs are transformed into high-dimensional space in the middle layer, allowing output representations to move independently. Therefore, the number of middle-layer neurons influences the precision for distinguishing similar data. Because very similar inputs (e.g., similar odors) need not necessarily be distinguished in many cases, the number of neurons in the aforementioned biological circuits [~2000 KCs in *Drosophila* and ~50 to 70 billion cerebellar granule cells in humans; (*51*)] might be sufficient to achieve representations with enough precision. Our model predicts that intermediate sparseness is optimal for achieving high-dimensional activities in a fixed number of middle-layer neurons (Fig. 4C). The sparseness of KC activities is known to be modulated by inhibitory neurons called anterior paired lateral (APL) neurons (*40*). Thus, experimentally modulating APL neurons could test the model prediction.

Second, the computation of the global factor D plays a critical role in this model. The proposed synaptic plasticity rule is written in the form of the three-factor learning rule (*17*, *18*), composed of the presynaptic activity, postsynaptic activity, and the global factor. Specifically, the presynaptic and postsynaptic terms are described as the difference in their activities at neighboring time points. This formulation aligns with previously proposed differential Hebbian learning, potentially corresponding to experimentally observed spike-timing–dependent plasticity in spiking neurons (*52*, *53*). In addition to the presynaptic and postsynaptic activities, the global factor D regulates network outcomes by incorporating common information among synapses. In the example of the olfactory circuit in *Drosophila*, one suggested global factor is the activity of DANs, which is known to modulate the synaptic plasticity from KCs to MBONs (*45*, *47*, *54*). Anatomically, the neurons corresponding to $\widehat{x^{\mathrm{diff}}}$ and $\widehat{y^{\mathrm{diff}}}$ that project to DANs might be specific types of KCs and MBONs (*55*, *56*). Although the activities corresponding to $\widehat{x^{\mathrm{diff}}}$ and $\widehat{y^{\mathrm{diff}}}$ have not yet been reported to our knowledge, similar components that calculate the difference between the current neuronal activity [$X(t)$ and $Y(t)$] and previous neuronal activity [$X(t-1)$ and $Y(t-1)$] might be computed using a neuronal adaptation mechanism. One study suggested that the flies' responses to odors showed a relatively slow adaptation over a timescale of several minutes (*22*). Therefore, it might be possible that the adaptation at a timescale compatible with the Hebbian *t*-SNE also exists and calculates the time derivative of neuronal activities. There is now no experimental evidence regarding the detailed formulation of $\widehat{x^{\mathrm{diff}}}$ and $\widehat{y^{\mathrm{diff}}}$, which should be further investigated experimentally to validate the biological plausibility of the model. The suggested role of the DANs in this model (i.e., comparing the input and output similarities) is a prediction to be further investigated in the future.

The third point is the biological mechanism by which inputs are repeatedly presented in a random order in our model. One possibility is that such a process can be realized by experiencing real inputs many times. Another possibility is to exploit offline learning with memory reactivation (*57*–*59*). During rest and non-rapid eye movement sleep, neural activities during awake experiences are reported to be reactivated (*25*, *60*, *61*). Repeated neuronal reactivation might be used to reorganize memory, as our model suggests. Representation learning during rest and sleep periods is an interesting future topic.

While previous studies have investigated learning algorithms in the *Drosophila* olfactory circuit where the association of inputs and rewards is given (i.e., supervised learning) (*62*–*65*), this study proposes an unsupervised learning algorithm that operates without labels. Some studies have suggested that odor exposure without reward or punishment alters MBON activity and fly behavior through synaptic plasticity induced by DANs (*47*, *66*). In the future, changes in representation resulting from such passive odor exposures could be experimentally compared with the predictions of our model. In addition, it would be interesting to explore how the unsupervised learning addressed in this study complements the supervised learning discussed in previous literature. In the final part (Fig. 6), we investigated the combination with a simple example of semi-supervised learning, demonstrating that the reward-modulated Hebbian *t*-SNE conducts association learning with generalization ability. The generalization can also be realized by the overlap of the KC activities. A previous study reported that flies' responses are generalized to similar odors, which share the KCs' activities with the original odor (*45*). In this case, the generalization happens simply because the two odors share the synapses from KCs to MBONs. This overlap of KCs between similar odors is an important aspect related to locality-sensitive hashing used in similarity searches (*41*, *67*). A recent experimental study reported that KC activities are not merely the result of random projection of PN activities but instead reflect the similarity of their distribution across natural sources (*24*). In addition, a theoretical study proposed a biologically plausible clustering algorithm mimicking K-means clustering by using the neuronal circuit from PNs to KCs (*32*). These studies highlight the significant role of KC representations. We demonstrated that the Hebbian *t*-SNE also works in the case that the representations in the middle layer Z overlap with each other (Figs. 4 and 5 and figs. S5 to S8). Further, our study suggests the complementary possibility that generalization can occur even without shared KC activities, using the input geometric structures. In Fig. 6, we suggested that the reward-modulated Hebbian *t*-SNE has superior generalization ability to the kernel perceptron. The generalization capability of the kernel perceptron exclusively relied on the overlapped KC activity in the model. Therefore, these results suggested an additional benefit of the Hebbian *t*-SNE using input geometric structures complementary to the KC activity overlaps. Crucially, the contribution to the generalization by the Hebbian *t*-SNE independent of the KC activity overlap can be tested by the experiment using odor mixtures in *Drosophila*

(Fig. 6, F and G). Our model predicts that, when one odor A is associated with reward and another odor B is associated with punishment, the valence of the 50% mixture of the two odors should depend on the continuity of odor mixtures used for learning. The acquired valence of the 50% mixture in this model differs depending on whether a mixture gap segregates positively or negatively rewarded odors (Fig. 6G), even when the KC activity overlap is identical. This distinction can also set our model apart from the algorithms in previous studies (*62*–*65*), where learning is driven purely by labeled (rewarded) inputs, without considering the effect of unlabeled inputs. Thus, these two different generalization strategies can be experimentally tested in the future.

A series of previous studies have addressed biologically plausible dimensionality reduction by using an idea called similarity matching (*68*). The similarity-based cost function, which includes the dot product (Gramian) of inputs and outputs as similarity measures, is minimized by a biologically plausible neuronal network, which is realized by effective variable substitution of the specific cost function (*14*, *32*, *68*–*70*). However, the cost function of the *t*-SNE is the KL divergence between input similarities measured by a Gaussian distribution and output similarities measured by a *t*-distribution (*15*), which cannot be simply realized by the similarity matching framework. The practical usefulness of the *t*-SNE suggests the importance of this formulation of the cost function. In particular, using a heavy-tailed distribution (e.g., *t*-distribution) for measuring output similarity is crucial for avoiding the crowding problem (*15*). To compute the cost function of the *t*-SNE, our model uses the differences in neural activity between neighboring time points. Further, we demonstrate the advantages of the Hebbian *t*-SNE over other biologically plausible algorithms such as PCA (*13*, *14*), SOM (*30*, *31*), and *K*-means clustering (*32*).

One study reported that conducting the dimensional reduction with the *t*-SNE before providing the data in the convolutional neural network (CNN) is useful for improving performance (*71*). Although this study assessed the representation by the linear separability in line with the previous studies, e.g., (*72*), the obtained representation by the Hebbian *t*-SNE could also be useful for more complex downstream models such as CNNs. Recent studies have proposed algorithms for general representation learning in deep neural networks related to competitive Hebbian learning (*73*) or self-supervised learning (*8*, *74*). In particular, a recent study proposed biologically plausible algorithms that bring the outputs corresponding to temporally proximate inputs closer together in time-series data (*8*). In contrast, our model aims to obtain low-dimensional representations that reflect input structures by measuring similarity as the distance in the input space. Consequently, the resulting output similarity relies on the temporal proximity of inputs in (*8*) and the spatial proximity of inputs in ours. While these models (*8*, *73*, *74*) focus on enhancing representation learning in deep neural networks, our model specifically targets dimensionality reduction in a neural network similar to the *Drosophila* olfactory system, which can be beneficial for efficient resource management. In addition, while the previous work (*73*) uses a CNN, where weight sharing may not be biologically plausible, our work proposes a simple network model that mimics actual *Drosophila* circuits. Further, our model derives the learning rule from a global cost function, whereas these models (*8*, *73*, *74*) do not deduce a cost function that summarizes the operation of the entire deep neural network. In summary, our model suggests that simple biological circuits with Hebbian synaptic plasticity can conduct nonlinear dimensionality reduction.

MATERIALS AND METHODS

Hebbian *t*-SNE

In this section, we concisely introduce the Hebbian *t*-SNE algorithm. See the next section for how this algorithm is derived based on the original *t*-SNE. Let us consider a three-layer neural network model. The three layers were denoted by $X = (x_1, x_2, \ldots, x_r)$, $Z = (z_1, z_2, \ldots, z_s)$, and $Y = (y_1, y_2)$. The whole N patterns $X^1, X^2, \ldots, X^N$ were presented in a random order numerous times. The values of X, Z, and Y at time t were denoted by $X(t) = [x_1(t), x_2(t), \ldots, x_r(t)]$, $Z(t) = [z_1(t), z_2(t), \ldots, z_s(t)]$, and $Y(t) = [y_1(t), y_2(t)]$ ($t = 0,1,2, \ldots$), respectively. The patterns of neighboring epochs were assumed to be different, thus $X(t) \neq X(t-1)$. The middle-layer activities Z in response to different input patterns were required to be linearly independent for an accurate implementation of the *t*-SNE. Except for Figs. 4 and 5 and figs. S5 to S8, the transformation from X to Z was set as a winner-take-all. The projection from Z to Y was linear, thus $Y(t) = W(t)Z(t)$, where $W(t) = [w_{lm}(t)]$ is a synaptic weight matrix. We considered a learning rule of synaptic weight W, inspired by the *t*-SNE. We adopted a batch update with T steps to speed up the simulation. The synaptic weights were updated with Adam (*28*), described as

$$w_{lm}(nT+1) = w_{lm}(nT) + \eta \frac{\hat{u}_{lm}(nT)}{\sqrt{\hat{v}_{lm}(nT)} + \epsilon}$$

with $\eta = 0.1$ and $\epsilon = 10^{-8}$. The terms $\hat{u}_{lm}(nT)$ and $\hat{v}_{lm}(nT)$ were described as

$$\hat{u}_{lm}(nT) = \frac{u_{lm}(nT)}{1-\rho_1^{n-n_s}}$$

$$\hat{v}_{lm}(nT) = \frac{v_{lm}(nT)}{1-\rho_2^{n-n_s}}$$

$$u_{lm}(nT) = \rho_1 u_{lm}[(n-1)T] + (1-\rho_1)\Delta w_{lm}(nT)$$

$$v_{lm}(nT) = \rho_2 v_{lm}[(n-1)T] + (1-\rho_2)\left[\Delta w_{lm}(nT)\right]^2$$

$$\Delta w_{lm}(nT) = \frac{N(N-1)}{T} \sum_{t=(n-1)T+1}^{nT} \left\{ D(t) \cdot \left[z_m(t) - z_m(t-1)\right] \cdot \left[y_l(t) - y_l(t-1)\right] \right\} \quad (1)$$

with $\rho_1 = 0.9$ and $\rho_2 = 0.999$. Note that the factor $\Delta w_{lm}(nT)$ represents the gradient of the loss function in Eq. 6. Synaptic weights were changed after the initial warmup steps ($n > n_s$ with $n_s = 500$), when the perplexity approached sufficiently close to the target value (Figs. 2E and 3E). The initial values $u_{lm}(n_sT)$ and $v_{lm}(n_sT)$ were set to zero.

Next, we describe how the global factor $D(t)$ was computed. The neuronal activity corresponding to $D(t)$ received inputs $\widehat{x^{\mathrm{diff}}}(t)$ and $\widehat{y^{\mathrm{diff}}}(t)$ (see Fig. 1 and Table 1), which estimate input and output similarities, respectively. The neuronal activity $\widehat{x^{\mathrm{diff}}}(t)$ was calculated using the middle-layer activity Z and the input change $\|X(t) - X(t-1)\|^2$. First, the axonal activity $a_k(t)$ based on the middle-layer activity $z_k(t)$ was set to one at time t when the $z_k(t)$ exhibited the largest positive value among the responses to the input patterns $X^1, X^2, \ldots, X^N$, and otherwise $a_k(t) = 0$. This mechanism can be biologically implemented,

for example, by gradually lowering each neuron's firing threshold from an initially very high value until the average firing rate becomes nonzero. This process ensures that each axonal activity $a_k(t)$ exhibits a positive value generically for only one input pattern. [In nongeneric cases, where $z_k(t)$ took the largest value for multiple input patterns, $a_k(t)$ was set to one for a randomly selected input pattern among them and zero otherwise.] Subsequently, each synapse had a synapse specific variable σ_k and a postsynaptic weight $1/\overline{x_k^{\mathrm{diff}}}$. The transmitter release of the synapse k was determined by $x_k^{\mathrm{diff}}(t) = a_k(t)\exp\left[-\frac{\|X(t)-X(t-1)\|^2}{2\sigma_k^2}\right]$. Last, the postsynaptic neuronal activity $\widehat{x^{\mathrm{diff}}}(t)$ was described using the transmitter release $x_k^{\mathrm{diff}}(t)$ and the postsynaptic weight $1/\overline{x_k^{\mathrm{diff}}}$ as

$$\widehat{x^{\mathrm{diff}}}(t) = \sum_{k=1}^{s} \frac{x_k^{\mathrm{diff}}(t)}{\overline{x_k^{\mathrm{diff}}}}$$

The postsynaptic weight $1/\overline{x_k^{\mathrm{diff}}}$ were batch updated so that the $\widehat{x^{\mathrm{diff}}}(t)$ was appropriately normalized according to

$$\overline{x_k^{\mathrm{diff}}}(nT+1) = \overline{x_k^{\mathrm{diff}}}(nT) + \frac{\overline{x_k^{\mathrm{diff}}}(nT)}{\tau_x} \cdot \frac{\sum_{t=(n-1)T+1}^{nT} a_k(t)\left[-1+(N-1)\cdot\widehat{x^{\mathrm{diff}}}(t)\right]}{\sum_{t=(n-1)T+1}^{nT} a_k(t)} \tag{2}$$

for $n \geq 1$ with $\tau_x = 100$. This normalization makes sure that the contribution from each pattern is equal regardless of the number K of simultaneously active axons (see the next section for details). The variable σ_k was regulated to achieve the desired perplexity of the t-SNE by introducing $H_k(t)$, which estimates information entropy computed as

$$\log\sigma_k(nT+1) = \log\sigma_k(nT) - \alpha\left[2^{H_k(nT+1)} - \mathrm{Perp}\right] \tag{3}$$

$$H_k(nT+1) = H_k(nT) + \frac{1}{\tau_p}\left\{-H_k(nT) - (N-1)\frac{\sum_{t=(n-1)T+1}^{nT} a_k(t)\cdot\widehat{x^{\mathrm{diff}}}(t)\cdot\log_2\left[\widehat{x^{\mathrm{diff}}}(t)+\epsilon\right]}{\sum_{t=(n-1)T+1}^{nT} a_k(t)}\right\} \tag{4}$$

for $n \geq 1$, with $\tau_p = 100$, $\alpha = 10^{-3}$, and $\epsilon = 10^{-8}$, which prevents the logarithm from being imaginary. The term Perp corresponds to the target perplexity in the t-SNE, which was set to 20 in Figs. 2, 4, 5, and 6 (A to E) and figs. S5 and S7, 40 in Figs. 3 and 6 (F and G) and figs. S3 and S9, and 30 in figs. S6 and S8.

The neuronal activity $\widehat{y^{\mathrm{diff}}}(t)$ was calculated using $y^{\mathrm{diff}}(t)$ as

$$\widehat{y^{\mathrm{diff}}}(t) = \frac{y^{\mathrm{diff}}(t)}{\overline{y^{\mathrm{diff}}}}$$

$$y^{\mathrm{diff}}(t) = \frac{1}{1+\|Y(t)-Y(t-1)\|^2}$$

where the time average $\overline{y^{\mathrm{diff}}}$ was batch updated as

$$\overline{y^{\mathrm{diff}}}(nT+1) = \overline{y^{\mathrm{diff}}}(nT) + \frac{1}{\tau_y}\left[-\overline{y^{\mathrm{diff}}}(nT) + \frac{N(N-1)}{T}\sum_{t=(n-1)T+1}^{nT} y^{\mathrm{diff}}(t)\right]$$

for $n \geq 1$, with time constant $\tau_y = 100$. Last, the neuronal activity $D(t)$ was calculated as

$$D(t) = -2\cdot\left[\frac{1}{N}\widehat{x^{\mathrm{diff}}}(t) - \widehat{y^{\mathrm{diff}}}(t)\right]\cdot y^{\mathrm{diff}}(t)$$

The variables $w_{lm}(t)$, $\overline{x_k^{\mathrm{diff}}}(t)$, $\overline{y^{\mathrm{diff}}}(t)$, $\sigma_k(t)$, and $H_k(t)$ were equal to the values of time $t-1$ except for the case described above.

Comparison between the *t*-SNE and Hebbian *t*-SNE

In this section, we derive the Hebbian t-SNE algorithm on the basis of the t-SNE (*15*). First, we briefly review the original t-SNE. In the t-SNE, the similarity p_{ij} between inputs X^i and X^j is defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

where N denotes the total number of inputs and $p_{i|j}$ is described as

$$p_{i|j} = \frac{\exp\left(-\|X^i - X^j\|^2/2\sigma_j^2\right)}{\sum_{j'\neq j}\exp\left(-\|X^{j'} - X^j\|^2/2\sigma_j^2\right)}$$

for $i \neq j$ and $p_{i|i} = 0$ for all i. The variables σ_j are determined by searching the value that achieves a suitable perplexity specified by the user for each data point j. The perplexity of data point j is defined as

$$P\left(p_{\cdot|j}\right) = 2^{-\sum_i p_{i|j}\log_2 p_{i|j}} \tag{5}$$

The similarity q_{ij} of outputs Y^i and Y^j is described as

$$q_{ij} = \frac{\left(1+\|Y^j - Y^i\|^2\right)^{-1}}{\sum_{j'}\sum_{i'\neq j'}\left(1+\|Y^{j'} - Y^{i'}\|^2\right)^{-1}}$$

for $i \neq j$ and $q_{i|i} = 0$ for all i. The t-SNE aims to minimize the KL divergence between the input similarities p_{ij} and the output similarities q_{ij}. Thus, the objective function C is described as

$$C = \sum_j\sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{6}$$

The gradient of this function C is given by

$$\nabla_{Y^j}C = \sum_{i\neq j} 4\left(p_{ij} - q_{ij}\right)\left(1+\|Y^j - Y^i\|^2\right)^{-1}\left(Y^j - Y^i\right)$$

Therefore, the mapping to output Y is changed according to this gradient. Note that Z_j in Table 1 is a column vector whose jth component is one and the other components are zero. If the mapping to output is represented as $Y^j = WZ^j$, then the gradient with respect to the synaptic weight matrix $W = (w_{lm})$ can be described as

$$\frac{\partial C}{\partial w_{lm}} = \sum_{j=1}^{N} \frac{\partial y_l^j}{\partial w_{lm}} \frac{\partial C}{\partial y_l^j} = \sum_{j=1}^{N} z_m^j \cdot \sum_{i \neq j} 4\left(p_{ij} - q_{ij}\right)\left(1 + \| Y^j - Y^i \|^2\right)^{-1}\left(y_l^j - y_l^i\right)$$

where y_l^j and z_m^j are the lth and mth components of Y^j and Z^j, respectively.

In the following paragraphs, we prove that the expected value of the synaptic changes in the Hebbian t-SNE $\Delta w_{lm}(nT)$ in Eq. 1 is proportional to the negative gradient $-\frac{\partial C}{\partial w_{lm}}$ under the condition that, for any input, at least one of the axonal activities $a_k(t)$ is positive; the batch size T is sufficiently large; and the time-average variables $\overline{x_k^{\text{diff}}}$, $\overline{y^{\text{diff}}}$, and H_k are calculated over sufficiently long time. From here, $A \longrightarrow B$ indicates that the value A approaches to the value B under this condition.

We first consider the case of $X(t) = X^j$ and $X(t-1) = X^i$. Let $k_1, k_2, \ldots, k_K$ be a set of all indices k such that $a_k(t) = 1$ in this case. By assumption, at least one such k exists. Because each axon k responds only to a single input pattern, the axonal activities $a_{k_1}, a_{k_2}, \ldots, a_{k_K}$ take positive value only when the input pattern is X^j. Moreover, because $a_{k_1}(t) = a_{k_2}(t) = \ldots = a_{k_K}(t)$ at all time t, the values of $\overline{x_k^{\text{diff}}}$ and σ_k are equal for all $k = k_1, k_2, \ldots, k_K$. Therefore, we obtain

$$\widehat{x^{\text{diff}}}(t) = \frac{x_{k_1}^{\text{diff}}(t)}{\overline{x_{k_1}^{\text{diff}}}} + \frac{x_{k_2}^{\text{diff}}(t)}{\overline{x_{k_2}^{\text{diff}}}} + \cdots + \frac{x_{k_K}^{\text{diff}}(t)}{\overline{x_{k_K}^{\text{diff}}}} = K \cdot \frac{x_{k_1}^{\text{diff}}(t)}{\overline{x_{k_1}^{\text{diff}}}}$$

By using this formula, the update of $\overline{x_{k_1}^{\text{diff}}}$ in Eq. 2 is described as

$$\begin{aligned}
\overline{x_{k_1}^{\text{diff}}}(nT+1) &= \overline{x_{k_1}^{\text{diff}}}(nT) + \frac{1}{\tau_x} \cdot \frac{\overline{x_{k_1}^{\text{diff}}}(nT)}{\sum_{t=(n-1)T+1}^{nT} a_{k_1}(t)} \cdot \sum_{t=(n-1)T+1}^{nT} a_{k_1}(t) \cdot \left[-1 + (N-1) \cdot \widehat{x^{\text{diff}}}(t)\right] \\
&= \overline{x_{k_1}^{\text{diff}}}(nT) + \frac{1}{\tau_x} \cdot \frac{\overline{x_{k_1}^{\text{diff}}}(nT)}{\sum_{t=(n-1)T+1}^{nT} a_{k_1}(t)} \cdot \sum_{t=(n-1)T+1}^{nT} a_{k_1}(t) \cdot \left[-1 + (N-1) \cdot K \cdot \frac{x_{k_1}^{\text{diff}}(t)}{\overline{x_{k_1}^{\text{diff}}}(t)}\right] \\
&= \overline{x_{k_1}^{\text{diff}}}(nT) + \frac{1}{\tau_x} \cdot \left[-\overline{x_{k_1}^{\text{diff}}}(nT) + K(N-1) \frac{\sum_{t=(n-1)T+1}^{nT} a_{k_1}(t) x_{k_1}^{\text{diff}}(t)}{\sum_{t=(n-1)T+1}^{nT} a_{k_1}(t)}\right]
\end{aligned}$$

The last equality is followed by $\overline{x_{k_1}^{\text{diff}}}(t) = \overline{x_{k_1}^{\text{diff}}}(nT)$ for t in $(n-1)T + 1 \leq t \leq nT$ because of the batch update. Therefore, noting $x_{k_1}^{\text{diff}}(t) = \exp\left(-\| X^i - X^j \|^2 / 2\sigma_{k_1}^2\right)$ when $X(t) = X^j$ and $X(t-1) = X^i$, we obtain after many iterations

$$\overline{x_{k_1}^{\text{diff}}} \to K \sum_{j' \neq j} \exp\left(-\| X^{j'} - X^j \|^2 / 2\sigma_{k_1}^2\right)$$

because of the low of large numbers and because every pattern pair appears with equal probability over the duration T. By using these formulas

$$\widehat{x^{\text{diff}}}(t) = \sum_k \frac{x_k^{\text{diff}}(t)}{\overline{x_k^{\text{diff}}}} = K \frac{x_{k_1}^{\text{diff}}(t)}{\overline{x_{k_1}^{\text{diff}}}} \to p_{i|j} \tag{7}$$

Note that σ_{k_1} achieves a desired perplexity in the same way as with the t-SNE for the following reason. From Eq. 7, we find

$$-(N-1)\frac{\sum_{t=(n-1)T+1}^{nT} a_{k_1}(t) \cdot \widehat{x^{\text{diff}}}(t) \cdot \log_2\left[\widehat{x^{\text{diff}}}(t) + \epsilon\right]}{\sum_{t=(n-1)T+1}^{nT} a_{k_1}(t)} \to -\sum_{i \neq j} p_{i|j} \log_2\left(p_{i|j} + \epsilon\right)$$

Following Eq. 4, the $H_{k_1}(t)$ approaches the left-hand side of this equation. Hence

$$2^{H_{k_1}(t)} \to 2^{-\sum_{i \neq j} p_{i|j} \log_2\left(p_{i|j} + \epsilon\right)} \to P\left(p_{\cdot|j}\right) \quad (\epsilon \to 0)$$

where $P(p_{\cdot|j})$ represents the perplexity of data point j in Eq. 5. Because the σ_{k_1} is regulated so that $2^{H_{k_1}(t)}$ approaches the target perplexity in Eq. 3, σ_{k_1} achieves the desired perplexity in the same way as the t-SNE.

Also

$$y^{\text{diff}}(t) = \left(1 + \| Y^j - Y^i \|^2\right)^{-1}$$
$$\overline{y^{\text{diff}}}(t) \to \sum_{j'} \sum_{i' \neq j'} \left(1 + \| Y^{j'} - Y^{i'} \|^2\right)^{-1}$$

which leads to $\widehat{y^{\text{diff}}}(t) = \frac{y^{\text{diff}}}{\overline{y^{\text{diff}}}} \to q_{ij}$. Therefore, using the definition of $D(t)$ and $D_{i|j}$ in Table 1

$$\begin{aligned}
D(t) &= -2 \cdot \left[\frac{1}{N}\widehat{x^{\text{diff}}}(t) - \widehat{y^{\text{diff}}}(t)\right] \cdot y^{\text{diff}}(t) \\
&\to -2 \cdot \left(\frac{1}{N} p_{i|j} - q_{ij}\right) \cdot \left(1 + \| Y^j - Y^i \|^2\right)^{-1} \\
&= D_{i|j}
\end{aligned}$$

Then, averaging over t with random pattern presentations, we obtain

$$\begin{aligned}
\Delta w_{lm}(nT) &= \frac{N(N-1)}{T} \sum_{t=(n-1)T+1}^{nT} \left\{D(t) \cdot \left[z_m(t) - z_m(t-1)\right] \cdot \left[y_l(t) - y_l(t-1)\right]\right\} \\
&\to \sum_j \sum_{i \neq j} D_{i|j} \cdot \left(z_m^j - z_m^i\right) \cdot \left(y_l^j - y_l^i\right) \\
&= -\sum_j \sum_{i \neq j} \left[\left(\frac{1}{N} p_{i|j} - q_{ij}\right) \cdot \left(1 + \| Y^j - Y^i \|^2\right)^{-1} \cdot \left(z_m^j - z_m^i\right) \cdot \left(y_l^j - y_l^i\right)\right. \\
&\quad \left. + \left(\frac{1}{N} p_{j|i} - q_{ji}\right) \cdot \left(1 + \| Y^i - Y^j \|^2\right)^{-1} \cdot \left(z_m^i - z_m^j\right) \cdot \left(y_l^i - y_l^j\right)\right] \\
&= -2 \sum_j \sum_{i \neq j} \left[\left(p_{ij} - q_{ij}\right) \cdot \left(1 + \| Y^j - Y^i \|^2\right)^{-1} \cdot z_m^j \cdot \left(y_l^j - y_l^i\right)\right. \\
&\quad \left. + \left(p_{ji} - q_{ji}\right) \cdot \left(1 + \| Y^i - Y^j \|^2\right)^{-1} \cdot z_m^i \cdot \left(y_l^i - y_l^j\right)\right] \\
&= -4 \sum_j \sum_{i \neq j} \left[\left(p_{ij} - q_{ij}\right) \cdot \left(1 + \| Y^j - Y^i \|^2\right)^{-1} \cdot z_m^j \cdot \left(y_l^j - y_l^i\right)\right] \\
&= -\frac{\partial C}{\partial w_{lm}}
\end{aligned}$$

In summary, the proposed algorithm realizes t-SNE in the biological neural circuit.

Simplified Hebbian *t*-SNE

The only difference from the original model was in the formulation of $\widehat{x^{\mathrm{diff}}}(t)$. In the simplified model, the term $\widehat{x^{\mathrm{diff}}}(t)$ was described as

$$x^{\mathrm{diff}}(t) = \exp\left[-\frac{\| X(t) - X(t-1) \|^2}{2\sigma_{\mathrm{all}}(t)^2}\right]$$

$$\widehat{x^{\mathrm{diff}}}(t) = \frac{x^{\mathrm{diff}}(t)}{\overline{x^{\mathrm{diff}}}}$$

where the variable $\overline{x^{\mathrm{diff}}}$ was updated by

$$\overline{x^{\mathrm{diff}}}(nT+1) = \overline{x^{\mathrm{diff}}}(nT) + \frac{1}{\tau_x}\left[-\overline{x^{\mathrm{diff}}}(nT) + \frac{N-1}{T}\sum_{t=(n-1)T+1}^{nT} x^{\mathrm{diff}}(t)\right]$$

for $n \geq 1$, with time constant $\tau_x = 100$. The variance σ_{all} common to all input patterns was updated by

$$\log\sigma_{\mathrm{all}}(nT+1) = \log\sigma_{\mathrm{all}}(nT) - \alpha\left[2^{H(nT+1)} - \mathrm{Perp}\right]$$

$$H(nT+1) = H(nT) + \frac{1}{\tau_p}\left\{-H(nT) - \frac{N-1}{T}\sum_{t=(n-1)T+1}^{nT} \widehat{x^{\mathrm{diff}}}(t)\cdot\log_2\left[\widehat{x^{\mathrm{diff}}}(t) + \epsilon\right]\right\}$$

for $n \geq 1$, with $\tau_p = 100$, $\alpha = 10^{-3}$, and $\epsilon = 10^{-8}$, which prevents the logarithm from being imaginary, and Perp corresponds to the target perplexity in the *t*-SNE, which was set to 20 in fig. S2 and 40 in fig. S4.

Reward-modulated Hebbian *t*-SNE

We introduced a reward-modulation term into the objective function as follows: $L = C - \lambda C_{\mathrm{R}}$ with $\lambda = 1$ and $C_{\mathrm{R}} = \sum_{k=1}^{N} R(X^k) y_1^k$, where $R(X^k)$ represents the reward associated with the input X^k. The gradient was described as

$$\frac{\partial C_{\mathrm{R}}}{\partial w_{lm}} = \sum_{k=1}^{N} z_m^k \cdot \delta_{l1} \cdot R(X^k)$$

where δ is Kronecker delta. By defining $\Delta w_{lm}^{\mathrm{R}}(nT) = \frac{N}{T}\sum_{t=(n-1)T+1}^{nT} z_m(t)\cdot\delta_{l1}\cdot R(X(t))$, then

$$\Delta w_{lm}^{\mathrm{R}}(nT) \to \sum_{k=1}^{N} z_m^k \cdot \delta_{l1} \cdot R(X^k) = \frac{\partial C_{\mathrm{R}}}{\partial w_{lm}}$$

Therefore, we added the $\Delta w_{lm}^{\mathrm{R}}$ to the Hebbian *t*-SNE as a reward modulation as follows

$$\Delta w_{lm}(nT) = \frac{N(N-1)}{T}\sum_{t=(n-1)T+1}^{nT} \left\{D(t)\cdot\left[z_m(t) - z_m(t-1)\right]\cdot\left[y_l(t) - y_l(t-1)\right]\right\} + \lambda\Delta w_{lm}^{\mathrm{R}}(nT)$$

In Fig. 6 and fig. S9, the reward $R(X(t))$ was set to $7.5\cdot10^{-6}/p$ and $-7.5\cdot10^{-6}/p$ for positive- and negative-rewarded input patterns, respectively, where p denotes the proportion of rewarded patterns. The reward $R(X(t))$ was set to zero for other non-rewarded input patterns.

Linear methods with transformed high-dimensional data

In Fig. 6 and fig. S9, the original three-dimensional input patterns were transformed into high-dimensional activities $Z = (z_1, z_2, \ldots, z_N)$, where N is the total number of input patterns. Each original input X^j was transformed into $Z^j = \left(z_1^j, z_2^j, \ldots, z_N^j\right)$ with $z_i^j = \exp\left[-\| X^j - X^i \|^2 / \left(2\sigma^2\right)\right]$. The optimal σ value was searched among 5,10,…,1000 and determined to maximize the average reward separability across the reward proportions ranging from 0.1 to 1.0. Consequently, σ was set to 25 in Fig. 6 and 915 in fig. S9.

In Fig. 6 and fig. S9, the linear perceptron, the linear perceptron adapted for transformed high-dimensional data, and the Hebbian *t*-SNE were compared. The generalization performance of the linear perceptron was calculated as follows. The linear perceptron, implemented with the scikit-learn package (*75*) in Python, was trained only using the positive- and negative rewarded points. All input patterns were then projected into a one-dimensional space using the obtained weight of the linear perceptron. The reward separability of the projected one-dimensional value was calculated.

Estimation of the KC activities

In Figs. 4 and 5 and figs. S5 to S8, the KC activities $Z = (z_1, z_2, \ldots, z_{2000})$ were estimated by using a theoretical model composed of random projection, thresholding, and normalization, consistent with the previous studies (*39*, *41*). We constructed the random weight matrix W_{rand} from the PNs X to KCs Z, in which each KC received inputs from seven randomly chosen PNs and those synaptic weights were set to one. Because the other synaptic weights were set to zero, the sum of each row of W_{rand} was seven. For a given PN activity X^j, the corresponding Z^j was defined as follows. First, $\widehat{Z}^j = \left(\widehat{z}_1^j, \widehat{z}_2^j, \ldots, \widehat{z}_{2000}^j\right)$ was computed as

$$\widehat{Z}^j = \phi\left[W_{\mathrm{rand}}\left(X^j\right)\right]$$

where the function ϕ is a threshold-linear function that outputs the original value when the input is positive and is among the largest p proportion of the total 2000 neurons and zero otherwise. The density p of the KC activities was set to 0.05 according to previous experimental observation (*20*) (but explored for various densities in Fig. 4C and fig. S6C). Then, the estimated KC activity Z^j was obtained by normalizing $\widehat{Z}^j$, described as $Z^j = \widehat{Z}^j / \left(\sum_{i=1}^{2000}\widehat{z}_i^j\right)$. The thresholding and normalization can be realized through global inhibition by APL neurons in the *Drosophila* olfactory circuit (*40*).

Other simulation details

The initial value of each component in the weight matrix W was randomly sampled from the standard Gaussian distribution. The initial value of σ_k was set to 500. The initial values of $\overline{y^{\mathrm{diff}}}(t)$, $\overline{x_k^{\mathrm{diff}}}(t)$, and $H_k(t)$ were described as

$$\overline{y^{\mathrm{diff}}}(1) = \frac{N(N-1)}{T}\sum_{t=1}^{T} y^{\mathrm{diff}}(t)$$

$$\overline{x_k^{\mathrm{diff}}}(1) = \epsilon + (N-1)\frac{\sum_{t=1}^{T}\left[a_k(t)\sum_{k=1}^{s} x_k^{\mathrm{diff}}(t)\right]}{\sum_{t=1}^{T} a_k(t)}$$

$$H_k(1) = -(N-1)\frac{\sum_{t=1}^{T} a_k(t)\cdot\widehat{x^{\mathrm{diff}}}(t)\cdot\log_2\left[\widehat{x^{\mathrm{diff}}}(t) + \epsilon\right]}{\sum_{t=1}^{T} a_k(t)}$$

with $\epsilon = 10^{-8}$. In the simplified Hebbian *t*-SNE, the initial values were described as

$$\overline{x^{\mathrm{diff}}}(1) = \epsilon + \frac{N-1}{T}\sum_{t=1}^{T} x^{\mathrm{diff}}(t)$$

$$H(1) = -\frac{N-1}{T}\sum_{t=1}^{T} \widehat{x^{\mathrm{diff}}}(t)\cdot \log_2\left[\widehat{x^{\mathrm{diff}}}(t) + \epsilon\right]$$

with $\epsilon = 10^{-8}$.

The batch size T was set to $[N(N-1)/10]$, where $[x]$ is a floor function. The number of total steps was 2000 in Figs. 3 to 5 and figs. S3 to S8, 10,000 in Fig. 2 and fig. S2, 100,000 in Fig. 6 (A to E), 20,000 in Fig. 6 (F and G), and 50,000 in fig. S9. The KL divergence plotted in Figs. 2 and 3 and figs. S2 and S4 was calculated using the fixed optimal variances σ_j that achieve the target perplexity.

Original data

The two entangled rings were described as $(r\sin 2i\pi/n, r\cos 2i\pi/n, 0)$ and $(r + r\sin 2i\pi/n, 0, r\cos 2i\pi/n)$, where $i = 0,1,\ldots,n-1$, $r = 1000$, and $n = 100$. The four entangled rings were described as $(r\sin 2i\pi/n, r\cos 2i\pi/n, 0)$, $\left(\frac{4r}{3} + r\sin 2i\pi/n, 0, r\cos 2i\pi/n\right)$, $\left(\frac{8r}{3} + r\sin 2i\pi/n, r\cos 2i\pi/n, 0\right)$, and $(4r + r\sin 2i\pi/n, 0, r\cos 2i\pi/n)$, where $i = 0,1,\ldots,n-1$, $r = 1000$, and $n = 100$. When the reward proportion was p, input patterns of $i < [np]$ were rewarded in Fig. 6 (A to E).

The S-shaped curve data in fig. S9 included 400 points generated by $[150\sin\theta, 100\ \mathrm{sgn}(\theta)(\cos\theta - 1), 200d]$, where $\mathrm{sgn}(\theta)$ represents a sign function, θ and d were sampled from the uniform distribution between $-3\pi/2$ and $3\pi/2$ and the uniform distribution between 0 and 1, respectively. In Fig. 6 (F and G), input patterns of the 20 to 40% largest or smallest θ values were excluded from the input data. When the reward population was p, input patterns of the $[200p]$ largest and smallest θ values were positively and negatively rewarded, respectively, in Fig. 6 (F and G) and fig. S9. The MNIST data were constructed by randomly choosing 1200 images from the MNIST dataset.

The experimental data of the receptor responses in Fig. 4 and figs. S5 and S6 were obtained from table S1 in (*21*). The experimental data of the PN activities and valence index in Fig. 5 and figs. S7 and S8 are described in (*22*). The raw data of the valence index were provided by Badel *et al.* (*22*).

Comparison with other methods

The PCA and *t*-SNE were conducted with the scikit-learn package (*75*) in Python. The perplexity of the *t*-SNE was set to 20 in Figs. 2, 4, and 5 and figs. S5 and S7, 30 in figs. S6 and S8, and 40 in Fig. 3 and fig. S3.

The SOM was implemented by using the minisom package (*76*). The map size was determined so that the total point is close to $5\sqrt{N}$, where N is the total number of inputs, based on a previously reported empirical method (*77*). Specifically, the map size was set to $M \times M$, where M is the integer closest to $\sqrt{5\sqrt{N}}$. The other hyperparameters were selected via grid search to minimize the Kaski-Lagus error (*78*, *79*). In fig. S7, a small noise was added to each point to allow further subdivision into smaller clusters using *K*-means clustering. In the visualization of the SOM, a small noise was added to prevent the overlap of points.

In addition, we compared the results of Hebbian *t*-SNE with those obtained by applying *K*-means clustering directly to the high-dimensional input. The *K*-means clustering was implemented by using the scikit-learn package (*75*). The number of clusters in figs. S3 and S5 was set to 10, corresponding to the number of true labels. The number of clusters in fig. S7 was determined to maximize the silhouette score, which resulted in 21 clusters. To enable the analysis in fig. S7, after clustering with *K*-means clustering, each cluster was randomly arranged in two-dimensional representations. In addition, a small noise was added within each cluster to allow further subdivision into smaller clusters using *K*-means clustering in fig. S7.

The definition of linear and reward separability

The linear separability was defined as the classification accuracy achieved using a linear support vector machine, implemented by using the scikit-learn package (*75*). The reward separability was defined as the maximum accuracy of a classifier that judges the data whose y_1 coordinate exceeds the threshold as belonging to the positive-rewarded group and all other data as belonging to the negative-rewarded group. The threshold was set to maximize the classification accuracy. The positive- and negative-rewarded groups were defined as follows. In Fig. 6 (A to E), the positive- and negative-rewarded groups encompassed their respective rewarded points as well as unrewarded points within the same rings as the respective rewarded points. In fig. S9, data points were divided into positive- and negative-rewarded groups along the S-shaped manifold as indicated by different colors in fig. S9.

Analysis related to clustering metrics

In figs. S3, S5, and S6, we obtained 10 clusters $U = \{U_1, U_2, \ldots, U_{10}\}$ by applying *K*-means clustering to each low-dimensional representation. We used 10 ground-truth clusters $V = \{V_1, V_2, \ldots, V_{10}\}$ on the basis of MNIST labels or chemical structures. We computed the adjusted Rand index, a measure of the similarity between two clusterings adjusted for chance, between the clusters U obtained by *K*-means clustering and the ground-truth clusters V using the scikit-learn package (*75*).

The analysis related to the odor and valence index

In Fig. 5, we obtained the K clusters $C_1, C_2, \ldots, C_K$ by applying the *K*-means clustering for each representation. We calculated the mean of the squared error M between the true valence index and its group mean, described as

$$M = \frac{1}{N}\sum_{k=1}^{K}\sum_{X^i \in C_k}\left[\mathrm{VI}(X^i) - \frac{1}{|C_k|}\sum_{X^j \in C_k}\mathrm{VI}(X^j)\right]^2$$

where N is the total number of odors and $\mathrm{VI}(X^i)$ represents the experimentally measured valence index of the odor X^i. We also calculated the value of M for the high-dimensional PN activities, denoted as M_0, as a baseline. We plotted the value of $M - M_0$ as the MSE difference for various numbers of the cluster number K in Fig. 5C and figs. S7C and S8C. In the shuffled case in Fig. 5C and figs. S7C and S8C, the shuffled valence index was used instead of the true valence index.

In figs. S7 and S8, the normalized Euclidean distances so that the maximum distance was one were plotted. A one-sided statistical test was used for the Pearson correlation coefficient.

Supplementary Materials

This PDF file includes:

Figs. S1 to S9

REFERENCES AND NOTES

1. J. A. Gallego, M. G. Perich, L. E. Miller, S. A. Solla, Neural manifolds for the control of movement. *Neuron* **94**, 978–984 (2017).
2. S. Vyas, M. D. Golub, D. Sussillo, K. V. Shenoy, Computation through neural population dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
3. A. E. Urai, B. Doiron, A. M. Leifer, A. K. Churchland, Large-scale neural recordings call for new insights to link brain and behavior. *Nat. Neurosci.* **25**, 11–19 (2022).
4. S. Chung, L. F. Abbott, Neural population geometry: An approach for understanding biological and artificial neural networks. *Curr. Opin. Neurobiol.* **70**, 137–144 (2021).
5. S. Fusi, E. K. Miller, M. Rigotti, Why neurons mix: High dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
6. J. J. DiCarlo, D. D. Cox, Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
7. J. J. DiCarlo, D. Zoccolan, N. C. Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
8. M. S. Halvagal, F. Zenke, The combination of Hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *Nat. Neurosci.* **26**, 1906–1915 (2023).
9. U. Cohen, S. Chung, D. D. Lee, H. Sompolinsky, Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 746 (2020).
10. S. Recanatesi, M. Farrell, M. Advani, T. Moore, G. Lajoie, E. Shea-Brown, Dimensionality compression and expansion in deep neural networks. arXiv:1906.00443 [cs.LG] (2019).
11. T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, G. Hinton, Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346 (2020).
12. B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, K. P. Kording, A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
13. T. Isomura, T. Toyoizumi, Error-gated Hebbian rule: A local learning rule for principal and independent component analysis. *Sci. Rep.* **8**, 1835 (2018).
14. C. Pehlevan, T. Hu, D. B. Chklovskii, A Hebbian/anti-Hebbian neural network for linear subspace learning: A derivation from multidimensional scaling of streaming data. *Neural Comput.* **27**, 1461–1495 (2015).
15. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
16. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML] (2018).
17. L. Kuśmierz, T. Isomura, T. Toyoizumi, Learning with three factors: Modulating Hebbian plasticity with errors. *Curr. Opin. Neurobiol.* **46**, 170–177 (2017).
18. N. Frémaux, W. Gerstner, Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. Neural. Circuits* **9**, 85 (2016).
19. T. Hige, Y. Aso, G. M. Rubin, G. C. Turner, Plasticity-driven individualization of olfactory coding in mushroom body output neurons. *Nature* **526**, 258–262 (2015).
20. M. N. Modi, Y. Shuai, G. C. Turner, The *Drosophila* mushroom body: From architecture to algorithm in a learning circuit. *Annu. Rev. Neurosci.* **43**, 465–484 (2020).
21. E. A. Hallem, J. R. Carlson, Coding of odors by a receptor repertoire. *Cell* **125**, 143–160 (2006).
22. L. Badel, K. Ohta, Y. Tsuchimoto, H. Kazama, Decoding of context-dependent olfactory behavior in *Drosophila*. *Neuron* **91**, 155–167 (2016).
23. S. J. C. Caron, V. Ruta, L. F. Abbott, R. Axel, Random convergence of olfactory inputs in the *Drosophila* mushroom body. *Nature* **497**, 113–117 (2013).
24. J.-Y. Yang, T. F. O'Connell, W.-M. M. Hsu, M. S. Bauer, K. V. Dylla, T. O. Sharpee, E. J. Hong, Restructuring of olfactory representations in the fly brain around odor relationships in natural sources. bioRxiv 528627 [Preprint] (2023). https://doi.org/10.1101/2023.02.15.528627.
25. D. Ji, M. A. Wilson, Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* **10**, 100–107 (2007).
26. F. Nadim, D. Bucher, Neuromodulation of neurons and synapses. *Curr. Opin. Neurobiol.* **29**, 48–56 (2014).
27. A. Panatier, J. Vallée, M. Haber, K. K. Murai, J.-C. Lacaille, R. Robitaille, Astrocytes are endogenous regulators of basal transmission at central synapses. *Cell* **146**, 785–798 (2011).
28. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG] (2014).
29. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
30. T. Kohonen, Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
31. J. A. Hertz, A. Krogh, R. G. Palmer, *Introduction To The Theory Of Neural Computation* (Avalon Publishing, 1991), vol. 1, pp. 44–70.
32. C. Pehlevan, A. Genkin, D. B. Chklovskii, "A clustering neural network model of insect olfaction" in *2017 51st Asilomar Conference on Signals, Systems, and Computers* (IEEE, 2017), pp. 593–600.
33. S. S. Schiffman, Physicochemical correlates of olfactory quality: A series of physicochemical variables are weighted mathematically to predict olfactory quality. *Science* **185**, 112–117 (1974).
34. A. Dravnieks, Odor quality: Semantically generated multidimensional profiles are stable. *Science* **218**, 799–801 (1982).
35. S. L. Pashkovski, G. Iurilli, D. Brann, D. Chicharro, K. Drummey, K. M. Franks, S. Panzeri, S. R. Datta, Structure and flexibility in cortical representations of odour space. *Nature* **583**, 253–258 (2020).
36. E. A. Hallem, J. R. Carlson, The odor coding system of *Drosophila*. *Trends Genet.* **20**, 453–459 (2004).
37. H. Kazama, R. I. Wilson, Origins of correlated activity in an olfactory circuit. *Nat. Neurosci.* **12**, 1136–1144 (2009).
38. H. Amin, A. C. Lin, Neuronal mechanisms underlying innate and learned olfactory processing in *Drosophila*. *Curr. Opin. Insect Sci.* **36**, 9–17 (2019).
39. K. Endo, Y. Tsuchimoto, H. Kazama, Synthesis of conserved odor object representations in a random, divergent-convergent network. *Neuron* **108**, 367–381.e5 (2020).
40. A. C. Lin, A. M. Bygrave, A. de Calignon, T. Lee, G. Miesenböck, Sparse, decorrelated odor coding in the mushroom body enhances learned odor discrimination. *Nat. Neurosci.* **17**, 559–568 (2014).
41. S. Dasgupta, C. F. Stevens, S. Navlakha, A neural algorithm for a fundamental computing problem. *Science* **358**, 793–796 (2017).
42. B. Babadi, H. Sompolinsky, Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).
43. A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, L. F. Abbott, Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164.e7 (2017).
44. T. Ronan, Z. Qi, K. M. Naegle, Avoiding common pitfalls when clustering biological data. *Sci. Signal.* **9**, re6 (2016).
45. T. Hige, Y. Aso, M. N. Modi, G. M. Rubin, G. C. Turner, Heterosynaptic plasticity underlies aversive olfactory learning in *Drosophila*. *Neuron* **88**, 985–998 (2015).
46. M. Adel, N. Chen, Y. Zhang, M. L. Reed, C. Quasney, L. C. Griffith, Pairing-dependent plasticity in a dissected fly brain is input-specific and requires synaptic CaMKII enrichment and nighttime sleep. *J. Neurosci.* **42**, 4297–4310 (2022).
47. A. Kato, K. Ohta, K. Okanoya, H. Kazama, Dopaminergic neurons dynamically update sensory values during olfactory maneuver. *Cell Rep.* **42**, 113122 (2023).
48. D. Owald, J. Felsenberg, C. B. Talbot, G. Das, E. Perisse, W. Huetteroth, S. Waddell, Activity of defined mushroom body output neurons underlies learned olfactory behavior in *Drosophila*. *Neuron* **86**, 417–427 (2015).
49. Y. Aso, D. Sitaraman, T. Ichinose, K. R. Kaun, K. Vogt, G. Belliart-Guérin, P.-Y. Plaçais, A. A. Robie, N. Yamagata, C. Schnaitmann, W. J. Rowell, R. M. Johnston, T.-T. B. Ngo, N. Chen, W. Korff, M. N. Nitabach, U. Heberlein, T. Preat, K. M. Branson, H. Tanimoto, G. M. Rubin, Mushroom body output neurons encode valence and guide memory-based action selection in *Drosophila*. *eLife* **3**, e04580 (2014).
50. P. Masek, M. Heisenberg, Distinct memories of odor intensity and quality in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 15985–15990 (2008).
51. C. S. von Bartheld, J. Bahney, S. Herculano-Houzel, The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *J. Comp. Neurol.* **524**, 3865–3895 (2016).
52. B. Kosko, Differential Hebbian learning. *AIP Conf. Proc.* **151**, 277–282 (1986).
53. S. Zappacosta, F. Mannella, M. Mirolli, G. Baldassarre, General differential Hebbian learning: Capturing temporal relations between events in neural networks and the brain. *PLOS Comput. Biol.* **14**, e1006227 (2018).
54. Y. Aso, D. Hattori, Y. Yu, R. M. Johnston, N. A. Iyer, T.-T. B. Ngo, H. Dionne, L. F. Abbott, R. Axel, H. Tanimoto, G. M. Rubin, Rubin The neuronal architecture of the mushroom body provides a logic for associative learning. *eLife* **3**, e04577 (2014).
55. I. Cervantes-Sandoval, A. Phan, M. Chakraborty, R. L. Davis, Reciprocal synapses between mushroom body and dopamine neurons form a positive feedback loop required for learning. *eLife* **6**, e23789 (2017).
56. F. Li, J. W. Lindsey, E. C. Marin, N. Otto, M. Dreher, G. Dempsey, I. Stark, A. S. Bates, M. W. Pleijzier, P. Schlegel, A. Nern, S. Takemura, N. Eckstein, T. Yang, A. Francis, A. Braun, R. Parekh, M. Costa, L. K. Scheffer, Y. Aso, G. S. X. E. Jefferis, L. F. Abbott, A. Litwin-Kumar, S. Waddell, G. M. Rubin, The connectome of the adult Drosophila mushroom body provides insights into function. *eLife* **9**, e62576 (2020).

57. E. L. Roscow, R. Chua, R. P. Costa, M. W. Jones, N. Lepora, Learning offline: Memory replay in biological and artificial reinforcement learning. *Trends Neurosci.* **44**, 808–821 (2021).
58. K. Yoshida, T. Toyoizumi, Information maximization explains state-dependent synaptic plasticity and memory reorganization during non-rapid eye movement sleep. *PNAS Nexus* **2**, pgac286 (2023).
59. K. Yoshida, T. Toyoizumi, Computational role of sleep in memory reorganization. *Curr. Opin. Neurobiol.* **83**, 102799 (2023).
60. H. S. Kudrimoti, C. A. Barnes, B. L. McNaughton, Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and EEG dynamics. *J. Neurosci.* **19**, 4090–4101 (1999).
61. U. Dag, Z. Lei, J. Q. Le, A. Wong, D. Bushey, K. Keleman, Neuronal reactivation during post-learning sleep consolidates long-term memory in *Drosophila*. *eLife* **8**, e42786 (2019).
62. R. Huerta, T. Nowotny, Fast and robust learning by reinforcement signals: Explorations in the insect brain. *Neural Comput.* **21**, 2123–2151 (2009).
63. R. Huerta, S. Vembu, J. M. Amigó, T. Nowotny, C. Elkan, Inhibition in multiclass classification. *Neural Comput.* **24**, 2473–2507 (2012).
64. J. E. M. Bennett, A. Philippides, T. Nowotny, Learning with reinforcement prediction errors in a model of the *Drosophila* mushroom body. *Nat. Commun.* **12**, 2569 (2021).
65. D. Lipshutz, A. Kashalikar, S. Farashahi, D. B. Chklovskii, A linear discriminant analysis model of imbalanced associative learning in the mushroom body compartment. *PLOS Comput. Biol.* **19**, e1010864 (2023).
66. D. Hattori, Y. Aso, K. J. Swartz, G. M. Rubin, L. F. Abbott, R. Axel, Representations of novelty and familiarity in a mushroom body compartment. *Cell* **169**, 956–969.e17 (2017).
67. C. Ryali, J. Hopfield, L. Grinberg, D. Krotov, "Bio-inspired hashing for unsupervised similarity search" in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, H. D. Iii, A. Singh, Eds. (PMLR, 2020), pp. 8295–8306.
68. C. Pehlevan, D. B. Chklovskii, Neuroscience-inspired online unsupervised learning algorithms: Artificial neural networks. *IEEE Signal Process. Mag.* **36**, 88–96 (2019).
69. A. M. Sengupta, M. Tepper, C. Pehlevan, A. Genkin, D. Chklovskii, Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks. bioRxiv 338947 [Preprint] (2018). https://doi.org/10.1101/338947.
70. Y. Bahroun, A. Acharya, D. Chklovskii, A. M. Sengupta, "Similarity-preserving neural networks from GPLVM and information theory" in *Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems* (NeurIPS, 2022), pp. 1–8.
71. L. Gao, D. Gu, L. Zhuang, J. Ren, D. Yang, B. Zhang, Combining t-distributed stochastic neighbor embedding with convolutional neural networks for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **17**, 1368–1372 (2020).
72. N. Deperrois, M. A. Petrovici, W. Senn, J. Jordan, Learning cortical representations through perturbed and adversarial dreaming. *eLife* **11**, e76384 (2022).
73. A. Journé, H. G. Rodriguez, Q. Guo, T. Moraitis, Hebbian deep learning without feedback. arXiv:2209.11883 [cs.NE] (2022).
74. B. Illing, J. Ventura, G. Bellec, W. Gerstner, Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *Adv. Neural Inf. Process. Syst.* **34**, 30365–30379 (2021).
75. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
76. G. Vettigli, MiniSom: Minimalistic and NumPy-based implementation of the self organizing map (2018); https://github.com/JustGlowing/minisom/.
77. J. Vesanto, "Neural network tool for data mining: SOM toolbox" in *Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000)* (Citeseer, 2000), pp. 184–196.
78. S. Kaski, K. Lagus, "Comparing self-organizing maps" in *Artificial Neural Networks – ICANN 96*, Lecture Notes in Computer Science (Springer, 1996), pp. 809–814.
79. F. Forest, M. Lebbah, H. Azzag, J. Lacaille, A survey and implementation of performance metrics for self-organized maps. arXiv:2011.05847 [cs.NE] (2020).
80. K. Yoshida, kkyoshida/bio-ndr: bio-ndr-v1.0.0, Zenodo (2024); https://doi.org/10.5281/zenodo.13970192.

Acknowledgments: We thank H. Kazama for providing the valence index data in (*22*), valuable comments on the manuscript, and helpful discussions. We are grateful to the RIKEN TRIP initiative (RIKEN Quantum) for insightful discussions. **Funding:** This study was supported by RIKEN Center for Brain Science (T.T. and K.Y.), JST CREST program JPMJCR23N2 (T.T.), KAKENHI Grant-in-Aid JP21J10564 (K.Y.) and JP23K19415 (K.Y.) from JSPS, and Masason Foundation (K.Y.). **Author contributions:** Writing—original draft: K.Y. and T.T. Conceptualization: K.Y. and T.T. Investigation: K.Y. Writing—review and editing: K.Y. Methodology: K.Y. and T.T. Resources: K.Y. and T.T. Funding acquisition: K.Y. and T.T. Data curation: K.Y. Validation: K.Y. Supervision: K.Y. and T.T. Formal analysis: K.Y. Software: K.Y. Project administration: K.Y. and T.T. Visualization: K.Y. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All experimental data used in this study are from the previous studies. The experimental data of the receptor responses in Fig. 4 and figs. S5 and S6 are available at (*21*). The experimental data of PN activities in Fig. 5 and figs. S7 and S8 are available at (*22*). The experimental data of the valence index from (*22*), used in Fig. 5 and figs. S7 and S8, and the source code are available at https://github.com/kkyoshida/bio-ndr and Zenodo (*80*).

Submitted 18 April 2024
Accepted 6 January 2025
Published 5 February 2025
10.1126/sciadv.adp9048