Normative and mechanistic model of an adaptive circuit for efficient encoding and feature extraction

Nikolai M. Chapochnikov[1,2,✉], Cengiz Pehlevan[3], Dmitri B. Chklovskii[1,4]

[1]Flatiron Institute, Simons Foundation, New York, NY, USA

[2]current address: Department of Neurology, New York University School of Medicine, New York, NY, USA

[3]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

[4]Neuroscience Institute, New York University School of Medicine, New York, NY, USA

✉email: nchapochnikov@gmail.com

Abstract

One major question in neuroscience is how to relate connectomes to neural activity, circuit function, and learning. We offer an answer in the peripheral olfactory circuit of the *Drosophila* larva, composed of olfactory receptor neurons (ORNs) connected through feedback loops with interconnected inhibitory local neurons (LNs). We combine structural and activity data and, using a holistic normative framework based on similarity-matching, we propose a biologically plausible mechanistic model of the circuit. Our model predicts the ORN → LN synaptic weights found in the connectome and demonstrate that they reflect correlations in ORN activity patterns. Additionally, our model explains the relation between ORN → LN and LN – LN synaptic weight and the arising of different LN types. This global synaptic organization can autonomously arise through Hebbian plasticity, and thus allows the circuit to adapt to different environments in an unsupervised manner. Functionally, we propose LNs extract redundant input correlations and dampen them in ORNs, thus partially whitening and normalizing the stimulus representations in ORNs. Our work proposes a comprehensive framework to combine structure, activity, function, and learning, and uncovers a general and potent circuit motif that can learn and extract significant input features and render stimulus representations more efficient.

Significance

The brain represents information with patterns of neural activity. At the periphery, due to the properties of the external world and of encoding neurons, these patterns contain correlations, which are detrimental for stimulus discrimination. We study the peripheral olfactory neural circuit of the Drosophila larva, that preprocesses neural representations before relaying them to higher brain areas. A comprehensive understanding of this preprocessing is, however, lacking. Here, we propose a mechanistic and normative framework describing the function of the circuit and predict the circuit's synaptic organization based on the circuit's input neural activity. We show how the circuit can autonomously adapt to different environments, extracts stimulus features, and decorrelate and normalize input representations, which facilitates odor discrimination downstream.

Introduction

Thanks to technological advances in connectomics (Eichler et al., 2017; Scheffer et al., 2020) and neural population activity imaging (Aimon et al., 2019), more and more neural circuits will soon be characterized anatomically and physiologically at unprecedented scale and detail. However, it is not clear what insights can be obtained from combining such datasets and how to use them to advance our understanding of brain computation. To address this, we focus on the peripheral olfactory system of the *Drosophila* larva - a small and genetically tractable circuit for which a connectivity (Berck et al., 2016) and comprehensive activity imaging (Si et al., 2019) datasets are already available.

This circuit is an analogous, but simpler version of the well-studied olfactory circuit in adult flies and vertebrates (Wilson, 2013). It contains 21 olfactory receptor neurons (ORNs), each expressing a different receptor type with a different odor sensitivity profile (**Fig. 1A**). ORN axons are reciprocally connected to a web of multiple interconnected inhibitory local neurons (LNs) through feedforward excitation and feedback inhibition. The connectome dataset contains not just the presence or absence of a connection between two neurons but also the number of synaptic contacts in parallel (Berck et al., 2016), which is an estimate of the connection strength, since synaptic contacts do not vary significantly in size in the *Drosophila* (Scheffer et al., 2020). The activity dataset contains the responses of ORNs to 34 odors at 5 dilutions (**Fig. 2A**) and has been obtained by imaging Ca^{2+} concentration in their somas (Si et al., 2019).

Previous studies addressed the role of the inhibitory feedback provided by LNs in transforming the neural representation of odors from ORN somas to projection neurons (PNs), which are postsynaptic to ORNs. In adult *Drosophila*, this circuit was suggested to perform gain-control and divisive normalization (Olsen et al., 2010; Olsen & Wilson, 2008), which equalizes different odor concentrations and decorrelates input channels. In the zebrafish larva, an analogous circuit was suggested to whiten the input leading to pattern decorrelation which helps odors discrimination downstream (Friedrich, 2013; Wanner & Friedrich, 2020).

However, the underlying mechanistic principles of computation are still not elucidated. For example, whereas different types of LNs have different connectivity patterns with ORNs in the *Drosophila* larva (Berck et al., 2016), the role of different LN types, their multiplicity, and their specific connectivity is not understood. Also, the peripheral olfactory circuit exhibits synaptic plasticity in response to olfactory environment changes (Arenas et al., 2012; Das et al., 2011; Devaud et al., 2001; Sachse et al., 2007; Sudhakaran et al., 2012), but the functional role of such plasticity is unclear.

To address these shortcomings, we use a combination of data analysis and modeling and develop a holistic theoretical framework that links circuit structure, function, activity data, and learning. Our contribution is fourfold. (1) We find that the ORNs → LN synaptic weights vectors reflect features of the independently acquired ORN activity patterns dataset (**Fig. 2, 3, 4**). (2) Building upon the similarity matching framework (Pehlevan et al., 2018), we develop a novel, biologically realistic, normative circuit model incorporating activity-dependent synaptic plasticity. (3) The model, driven by the ORN activity dataset, predicts the following observations in the structural dataset: the ORNs → LN synaptic weights (**Fig. 4**), the emergence of LNs groups (**Fig. 4**), and the relationship between feedforward ORN → LN and lateral LN - LN connection (**Fig. 5**). (4) Using our model, we characterize the circuit computation (**Fig. 6, 7**), and propose that LNs play a dual role in rending the neural representation of odors in ORNs more efficient and extracting useful features that are transmitted downstream. Furthermore, we show that the synaptic weights enabling this computation can be learned by the circuit in an unsupervised manner.

In this study, we further our understanding of LNs and their computations. We highlight the importance of minutely organized ORN - LN and LN - LN connection weights, which allows LNs to encode different significant features of input activity and dampen them in ORN axons. The transformation from the representation in ORN somas to that in ORN axons consists of a partial equalization of the PCA variances, which enables a more efficient stimulus encoding (Barlow, 1961). Indeed, this results in a decorrelation and equalization of ORNs and odor representations, which correspond to two fundamental computations in the brain: partial ZCA (zero-phase) whitening (Bell & Sejnowski, 1997; Kessy et al., 2018) and divisive normalization (Carandini & Heeger, 2012). In essence, we uncover an elegant neural circuit motif that can, via associative Hebbian plasticity, adapt to different stimuli environment and learn to extract features as well as to perform two critical computations. Thus, we present a framework that allows to quantitatively link synaptic weights in the structural data with the circuit's function and with the circuit adaptation to input correlations, thus making a crucial step towards more integrated understanding of neural circuits.

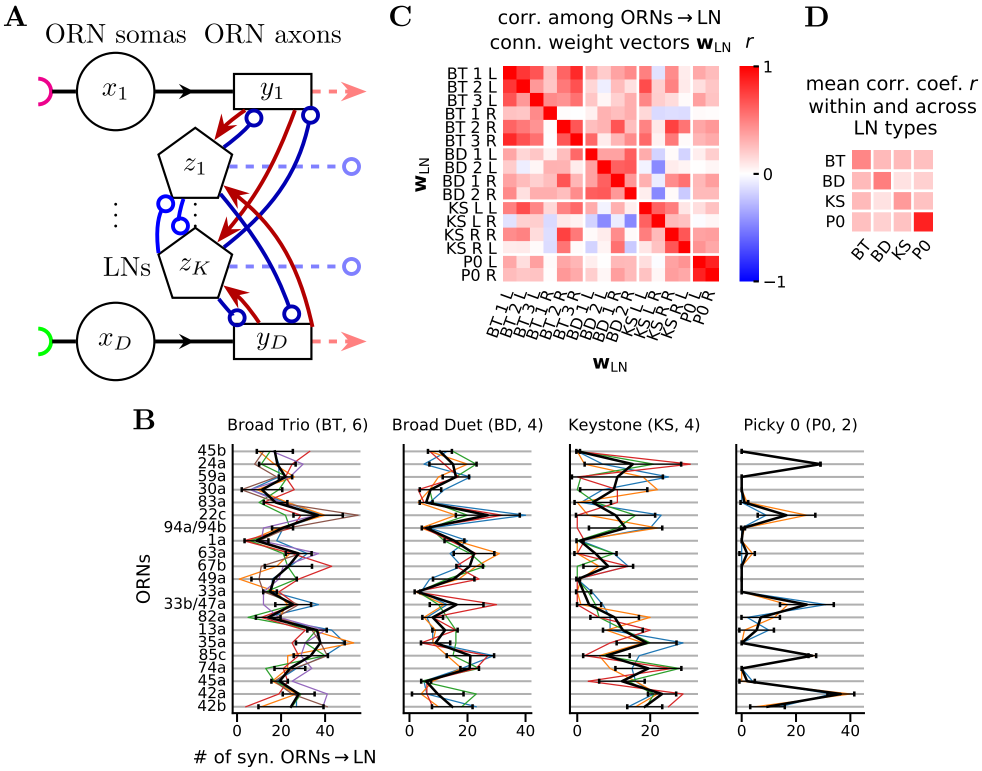

Fig. 1. Circuit connectivity and LN types

A Scheme of the ORN-LN circuit. Each of the D ORNs is depicted as a two-compartment unit with a soma (circle) and an axonal terminal (rectangle). The differently colored half circles on the left represent different chemical receptor types. K inhibitory local neurons (LNs, pentagons) reciprocally connect with ORN axons and between themselves. ORN axons and LNs transmit information further downstream (dashed lines). Red lines with arrowheads and blue lines with open circles represent excitatory and inhibitory connections, respectively. x_i, y_i, and z_i represent the activity of ORN somas, axons, and LNs, respectively.

B Feedforward ORNs → LN connection weight vectors, $\mathbf{w}_{\text{LN}}$ (colored lines), and average feedforward ORNs → LN type connection weight vectors, $\mathbf{w}_{\text{LNtype}}$ (black lines, mean ± s.d.) for each LN type (see also **Fig. S2A**).

C Correlation coefficients r between all $\mathbf{w}_{\text{LN}}$. L: left, R: right. KS L R is the Keystone with the soma positioned on the left side of the larva, connecting with the ORNs of the right side, and vice-versa for KS R L. Since Picky 0 receives synaptic input mainly on the dendrite, here we only use the connections synapsing onto the dendrite.

D Average rectified correlation coefficient $\langle r_+ \rangle$ ($r_+ := \max[0, r]$) between LN types calculated by averaging the rectified values from (**C**) in each rectangle with white border, excluding the diagonal entries of the full matrix.

Results

ORN-LN circuit

ORNs in the *Drosophila* larva carry odor information from antennas to the antennal lobe. There it is reformatted and handed over to PNs which transmit it to higher brain areas like the mushroom body and the lateral horn (Berck et al., 2016). LNs, which synapse bidirectionally with ORN axons and PN dendrites, strongly contribute to this reformatting through presynaptic and postsynaptic inhibition, as mainly shown in the adult fly (Asahina et al., 2009; Chou et al., 2010; Kim et al., 2015; Laurent, 2002; Nagel et al., 2014; Olsen et al., 2010; Olsen & Wilson, 2008).

Here, we focus on the circuit and computation presynaptic to PNs, i.e., occurring from ORN somas to ORN axons driven by LN inhibition. Specifically, we study the sub-circuit formed by all $D = 21$ ORNs and those 4 LN types (on each side of the brain) that provide direct inhibitory feedback onto the ORNs (Berck et al., 2016) (**Fig. 1A, S1**). The 4 LN types include 3 Broad Trio (BT) neurons, 2 Broad Duet (BD) neurons, 1 Keystone (KS, bilateral connections) neuron and 1 Picky 0 (P0) neuron (**Fig. S1, S2A**). This amounts to 8 ORNs - LN connections per side (3 BTs, 2 BDs, 2 KSs, and 1 P0s), and 16 on both sides.

We use the number of synapses in parallel between two neurons as a proxy of the synaptic weight w because synapses in the *Drosophila* larva have been found to be of similar sizes (Scheffer et al., 2020; Takemura et al., 2013) and synaptic size correlates with strength (Holderith et al., 2012). In the linear approximation, the contribution of a connection to the postsynaptic neuron activity a_{post} is proportional to the product of w and the presynaptic neuron activity a_{pre}, i.e., $a_{post} \propto w \cdot a_{pre}$.

We focus our analysis on the feedforward ORNs → LN connection weight vectors, $\mathbf{w}_{\mathrm{LN}}$, whose $D = 21$ components are w's corresponding to the connections from different ORNs onto the same post-synaptic LN rather than the feedback LN → ORNs. Because all the components of such a weight vector share the same post-synaptic neuron their effect on the post-synaptic activity is directly comparable, i.e. the coefficient of proportionality in $a_{\mathrm{LN}} \propto \sum_i w_{\mathrm{LN},i} \cdot a_{pre,i}$ is the same. Conversely, the ws from one LN onto all 21 ORNs are not directly comparable among each other, because each connection affects a different postsynaptic ORN, which potentially has different electrical properties. Yet, the feedforward and feedback connection vectors are somewhat correlated (**Fig. S2**).

While Berck et al., 2016 divided the LNs into the above types based on their neuronal lineage, morphology, and qualitative connectivity, we also find that such types are innervated differently by ORNs (**Fig. 1B**). Indeed, the average correlations within LN type is higher than between LN types $\mathbf{w}_{\mathrm{LN}}$ (**Fig. 1C,D**). Thus, for a part of our study (**Fig. 2, 3, 4A,B**) we use the 4 average $\mathbf{w}_{\mathrm{LNtype}} = \frac{1}{n} \sum_{\mathrm{LN} \in \mathrm{LNtype}} \mathbf{w}_{\mathrm{LN}}$, where n is the number of connection vectors for that LN type.

Odor representations in ORNs are aligned with ORNs → Broad Trio connectivity weight vector

Several studies proposed that the LNs could facilitate decorrelation of the neural representation of odors (Friedrich, 2013; Friedrich & Laurent, 2001; Friedrich & Wiechert, 2014; Giridhar et al., 2011; Gschwend et al., 2015; Wanner & Friedrich, 2020). To perform such decorrelation, the circuit needs to be adapted to or "know about" the correlations in the activity patterns (Simoncelli & Olshausen, 2001). We investigated if this is the case in this olfactory circuit by testing whether the $\mathbf{w}_{\mathrm{LNtype}}$ contain signatures of ORN activity patterns.

An ensemble of ORN activity patterns $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ ($t = 1, ..., 170$) was obtained using Ca^{2+} fluorescence imaging of ORN somas in response to a set of 34 odorants at 5 dilutions (Si et al., 2019) (**Fig. 2A**). These odorants were chosen from the components of fruits and plant leaves from the larva's natural environment to stimulate ORNs as broadly and evenly as possible, with many odorants activating just a single ORN at the lowest concentration (i.e., the highest dilution).

Activity patterns $\mathbf{x}^{(t)}$ elicited by different odorants are correlated with the synaptic weight vector $\mathbf{w}_{\mathrm{BT}}$ to a different degree (**Fig. 2B-D**), yet are such correlations statistically significant? To determine this, we first calculate the Pearson correlation coefficients r between the four $\mathbf{w}_{\mathrm{LNtype}}$ and the ensemble of $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ (**Fig. 2E**). Each $\mathbf{w}_{\mathrm{LNtype}}$ exhibits a different "connectivity tuning curve" shape (**Fig. 2F**), $\mathbf{w}_{\mathrm{BT}}$ being the most broadly aligned to the $\mathbf{x}^{(t)}$ of this stimuli set, $\mathbf{w}_{\mathrm{P0}}$ the most sharply aligned to a few $\mathbf{x}^{(t)}$, and the $\mathbf{w}_{\mathrm{BD}}$ and $\mathbf{w}_{\mathrm{KS}}$ the most weakly aligned. To test if the $\mathbf{w}_{\mathrm{LNtype}}$ are significantly aligned with the ensemble $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$, we compare the relative cumulative frequency (RCF) of r in the data with the RCFs of r obtained after randomly shuffling the entries of each $\mathbf{w}_{\mathrm{LNtype}}$ (**Fig. 2G,H**). We use the maximum deviation from the mean RCF from shuffled connection vector to measure significance and find that only $\mathbf{w}_{\mathrm{BT}}$ is significantly aligned to $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ (**Fig. 2H,I**).

Furthermore, we find that $\mathbf{w}_{\mathrm{BT}}$ is significantly aligned with the first PCA direction of $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ (**Fig. S6A,B**), but none of remaining $\mathbf{w}_{\mathrm{LNtype}}$ significantly aligned with any of the top 5 PCA directions (**Fig. 3**). We choose to compare with the top 5 (instead of 4, as the number of $\mathbf{w}_{\mathrm{LNtype}}$) PCA directions of $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ to cover more activity direction, thus accounting for the fact that this activity dataset does not have the same statistics of odors as the true larva environment, and likely has a different order of PCA directions. We performed PCA without centering $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$, to avoid any preprocessing on the activity data and mimic what the circuit is experiencing. The first PCA direction is thus relatively close to the mean activity direction.

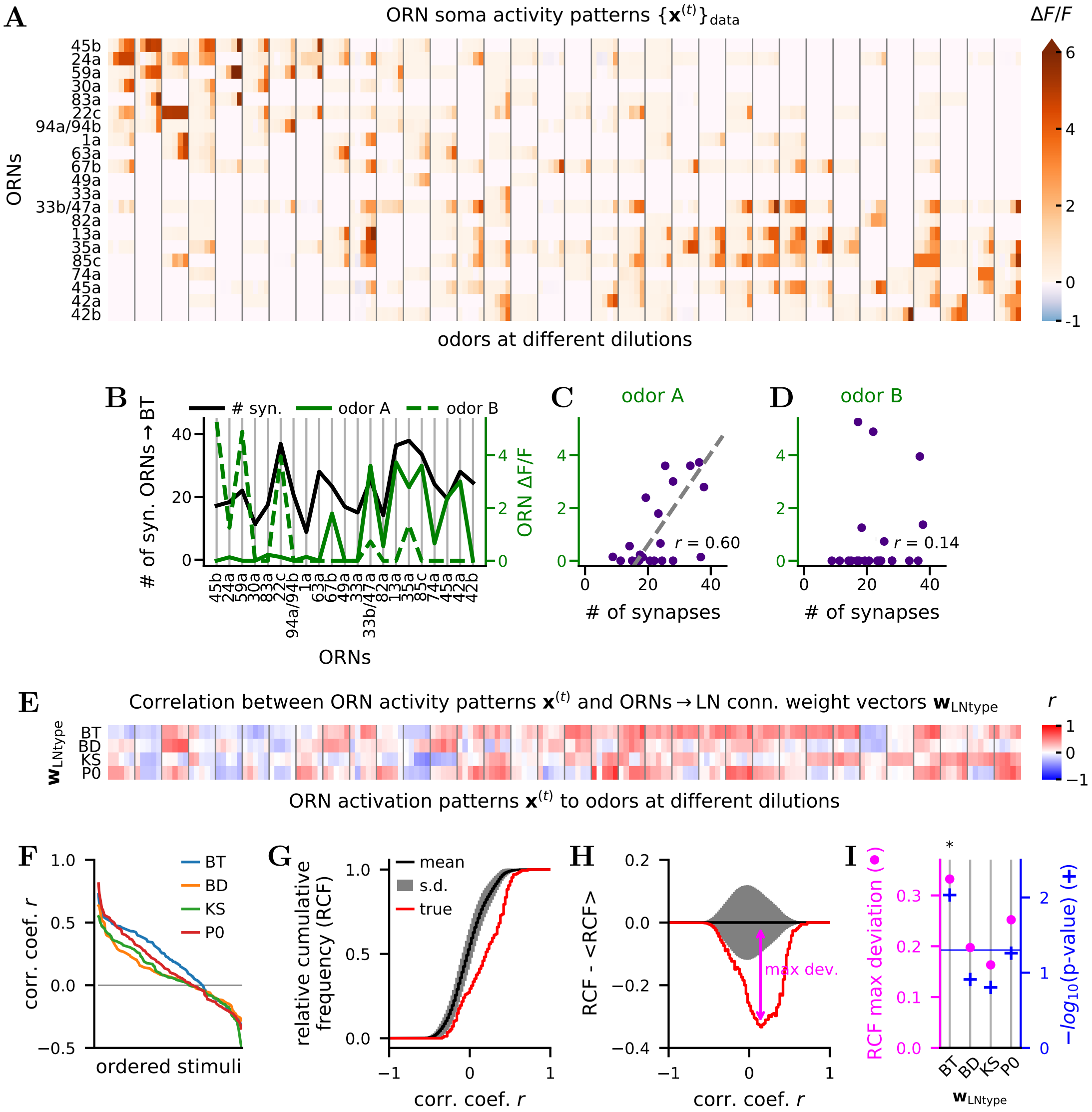

Fig. 2. Alignment of ORNs → LN connectivity weight vectors with odor representations in ORN activity
A Activity patterns $\{\mathbf{x}^{(t)}\}_{\text{data}}$ at ORN soma in response to 34 odors at 5 dilutions from Si et al., 2019. Different odors are separated by vertical gray lines. For each odor, there are 5 columns corresponding to 5 dilutions: $10^{-8}, ..., 10^{-4}$. See **Fig. S3** for odor labels and scaled $\mathbf{x}^{(t)}$.
B $\mathbf{w}_{\text{BT}}$ superimposed with ORNs activity patterns $\mathbf{x}^{(\text{A})}$ and $\mathbf{x}^{(\text{B})}$ in response to the ligands 2-heptanone (odor A) and 2-acetylpyridine (odor B) at dilution 10^{-4}.
C-D Scatter plot representation of (**B**). $\mathbf{w}_{\text{BT}}$ is more strongly tuned to $\mathbf{x}^{(\text{A})}$ $(r = 0.6)$ than to $\mathbf{x}^{(\text{B})}$ $(r = 0.14)$.
E Correlation coefficients between $\mathbf{w}_{\text{LNtype}}$ with the $\mathbf{x}^{(t)}$ from (**A**) (see also **Fig. S4A**).
F LN "connectivity tuning curves": correlation coefficients sorted in decreasing order from (**E**) for each $\mathbf{w}_{\text{LNtype}}$.
G Red line: relative cumulative frequency (RCF) of the correlation coefficients r of the first row of (**E**). Black line and gray band: mean ± s.d. from the RCFs generated by 10,000 instances of shuffling the entries of $\mathbf{w}_{\text{BT}}$. Bin size: 0.02.
H Same as (**G**) with the mean RCF subtracted. We define the maximum deviation as the maximum negative difference between the true and the mean RCF of correlation coefficients.
I RCF maximum deviation and $\log_{10}$ of false discovery rate (FDR, Benjamini and Hochberg, 1995) adjusted p-values for each $\mathbf{w}_{\text{LNtype}}$ (see also **Fig. S4B**). *: significance with FDR at 5%.

Next, to test whether the connection vectors $\mathbf{w}_{\text{LNtype}}$ might be linear combinations of the PCA directions of $\{\mathbf{x}^{(t)}\}_{\text{data}}$, we examine the alignment of the subspace spanned by the 4 $\mathbf{w}_{\text{LNtype}}$ and the one spanned by the top 5 PCA directions of $\{\mathbf{x}^{(t)}\}_{\text{data}}$ (**Fig. S5**). We define a measure $0 \leq \Gamma \leq 4$, approximately representing the number of aligned directions between these 2 subspaces (**Methods**) and find $\Gamma \approx 2$. This value significantly deviates from the expected Γ from subspaces generated by 4 and 5 Gaussian random normal vectors in 21 dimensions ($p < 10^{-4}$) and subspaces generated from the 4 connectivity vectors with shuffled entries and the 5 original activity vectors from PCA ($p < 0.01$). Approximately 1 more dimension is significantly aligned between the 2 subspaces than expected by random, supporting the results from **Fig. 3C**.

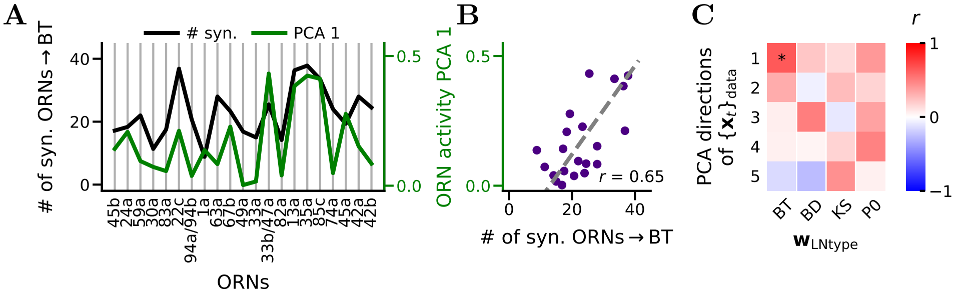

Fig. 3. Alignment of $\mathbf{w}_{\text{BT}}$ with the top PCA direction of ORN activity patterns $\{\mathbf{x}^{(t)}\}_{\text{data}}$

A $\mathbf{w}_{\text{BT}}$ superimposed in the 1st PCA direction of $\{\mathbf{x}^{(t)}\}_{\text{data}}$.

B Scatter plot representation of (**A**).

C Correlation coefficient r between the top 5 principal directions of $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and the four $\mathbf{w}_{\text{LNtype}}$ (see also **Fig. S6C,D,G**). Two-sided p-values were calculated by shuffling the entries of each $\mathbf{w}_{\text{LNtype}}$. 50,000 permutations used. *: significance with FDR at 5%.

In summary, we find that $\mathbf{w}_{\text{BT}}$ is adapted to ORNs activity patterns $\{\mathbf{x}^{(t)}\}_{\text{data}}$ as demonstrated by (1) the significant alignment of $\mathbf{w}_{\text{BT}}$ with individual activity patterns $\mathbf{x}^{(t)}$, (2) the significant alignment of $\mathbf{w}_{\text{BT}}$ with the top PCA direction of $\{\mathbf{x}^{(t)}\}_{\text{data}}$, and (3) by a significantly large Γ. This supports the idea that the circuit is at least partially adapted to ORN activity patterns. This analysis fails, however, to reveal the relation between ORN activity and LNs other than BT.

A normative and mechanistic model of the ORN-LN circuit

A detailed bottom-up modeling of the circuit requires the knowledge of the multiple unavailable physiological parameters such as ion channel distributions and neural morphologies. We therefore take here a route that circumvents these unknowns and harvests the benefits of normative approaches: similar to physics, we guess the circuit cost function, derive the governing equations, and see if their predictions agree with experiments.

Similarity-matching objective functions have been shown to be capable of extracting PCA subspaces and can be optimized by biologically plausible neural circuits with Hebbian synaptic learning rules (Pehlevan et al., 2018). Motivated by the result that the ORN-LN circuit might be adapted to at least one PCA direction of the input, we postulated a similarity-matching inspired objective function (equation (18)), such that its online optimization equations maps onto the neural dynamics of the ORN-LN circuit (equations (19), (20)) and Hebbian plasticity update rules for ORN-LN and LN-LN synapses (equation (21)). Biologically, the circuit synaptic weights could be "learned" either over evolutionary time scales, and/or during the animal lifetime.

Given a set of T inputs $\left[\mathbf{x}^{(1)}, ..., \mathbf{x}^{(T)}\right] = \left\{\mathbf{x}^{(t)}\right\}_{t=1...T}$ representing the activity patterns of ORN somas, the model provides us with the learned connection weights between D ORNs and K LNs: $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K]$ as well as between LNs: $\mathbf{M} = \{m_{i,j}\}_{i,j=1...K}$. $m_{i,i}$ relates to the leak term of LN i. $[\mathbf{w}_1, ..., \mathbf{w}_K]$ and $\mathbf{M}$ set the input-output relationship of the circuit and determine the activity patterns of ORN axons: $\left\{\mathbf{y}^{(t)}\right\}_{t=1...T}$ and LNs: $\left\{\mathbf{z}^{(t)}\right\}_{t=1...T}$. In addition to K, the number of LNs, the model contains only one effective parameter ρ characterizing the strength of the feedback inhibition.

We consider two models. First is a Linear Circuit LC-K, (equations (19), arising from the unconstrained objective function (18)), for which we derived an analytical solution for $[\mathbf{w}_1, ..., \mathbf{w}_K]$, $\mathbf{M}$, $\left\{\mathbf{y}^{(t)}\right\}$, and $\left\{\mathbf{z}^{(t)}\right\}$ (**Supplementary Information**). Although linearity might be an oversimplification of the biological reality, it allows us to build up intuition. Second is a Non-Negative Circuit, NNC-K, (equations (20), arising from objective function (18), containing non-negativity constraints on the ORN axon and LN activity), which might be more biologically plausible. The results below for the NNC arise from numerical simulations.

Predictions of the ORN - LN connection weight vectors

We start by analyzing the prediction of our model in terms of circuit connectivity. In the LC-K, the $\{\mathbf{w}_k\}_{k=1...K}$ span the subspace of the top K PCA directions of the input $\left\{\mathbf{x}^{(t)}\right\}$ (**Supplementary Information**):

$$\mathbf{w}_k = \sum_{i=1}^{K} a_{k,i} \mathbf{u}_i \tag{1}$$

where $\{\mathbf{u}_i\}_{i=1...K}$ are the top K PCA directions of the dataset $\left\{\mathbf{x}^{(t)}\right\}$, $\{a_{i,j}\}_{i,j=1...K}$ are coefficients such that all $\mathbf{w}_k$ are linearly independent. Thus, the $\mathbf{w}_k$ in the LC do not necessarily correspond to specific PCA directions and are not orthogonal, and there is a degree of freedom in the $\{a_{i,j}\}$, making the solution of the optimization not unique. Such synaptic organization assure that LNs in the LC extract the top K PCA subspace of the input (below). This structural prediction is tested and only partially verified in the data above (**Fig. 3**): the first PCA direction of $\left\{\mathbf{x}^{(t)}\right\}_{\text{data}}$ significantly aligns with $\mathbf{w}_{\text{BT}}$, but there is no full alignment between the connectivity $\{\mathbf{w}_{\text{LNtype}}\}$ and activity ORN principal subspaces.

Next, we study the predictions of the NNC-4 ($K = 4$ as the number of LN types). We numerically optimize the objective function (18) with $\left\{\mathbf{x}^{(t)}\right\}_{t=1...T} = \left\{\mathbf{x}^{(t)}\right\}_{\text{data}}$ (**Fig. 2A**), $K = 4$, $\rho = 1$ and obtain $\left\{\mathbf{y}^{(t)}\right\}$, $\left\{\mathbf{z}^{(t)}\right\}$, and $[\mathbf{w}_1, ..., \mathbf{w}_4]$ (**Fig. S6C**). Intuitively, the $\{\mathbf{w}_k\}$ relate to cluster centers in soft K-means or to features in non-negative matrix factorization and the $z_k^{(t)}$ are the

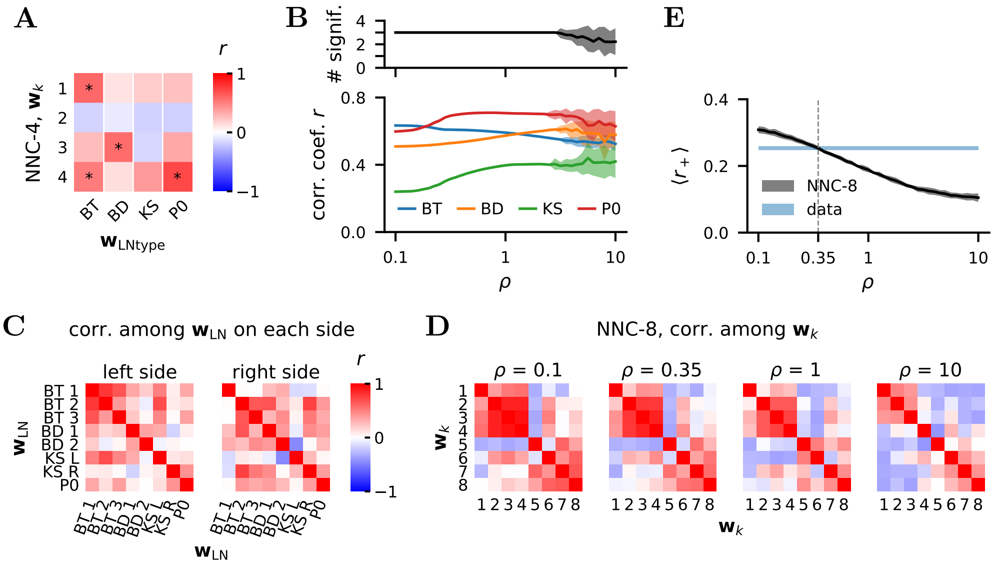

Fig. 4. Prediction of the connectivity with the NNC and emergence of LN types

A Correlation coefficient r between the four $\mathbf{w}_k$ from NNC-4 ($\rho = 1$) and the four $\mathbf{w}_{\text{LNtype}}$ (see also **Fig. S6C,D,F-H**). One-sided p-values were calculated by shuffling the entries of each $\mathbf{w}_{\text{LNtype}}$. 50,000 permutations used. *: significance with FDR at 5%.

B Bottom: maximum correlation coefficient (mean ± s.d.) of the four $\mathbf{w}_k$ from NNC-4 with the four $\mathbf{w}_{\text{LNtype}}$ for different values of ρ. Top: number of $\mathbf{w}_{\text{LNtype}}$ significantly correlated with at last one $\mathbf{w}_k$ from NNC-4 (FDR at 5%). 50 numerical simulations of NNC-4 for each value of ρ.

C Correlation between the $\mathbf{w}_{\text{LN}}$ on the left and right sides of the larva brain.

D Same as (**C**) for the eight $\mathbf{w}_k$ arising from NNC-8 and with $\rho = 0.1, 0.35, 1, 10$. $\mathbf{w}_k$ ordered with hierarchical clustering.

E Mean rectified correlation coefficient $\langle r_+ \rangle$ from (**C**) (blue band delimited by the value for left and right circuit) and from NNC-8 (black line, mean ± s.d.). One $\langle r_+ \rangle$ is obtained by averaging all the rectified values in a matrix in (**C**) or (**D**), excluding the diagonal. For the NNC-8 and a given value of ρ, we run 50 simulations. Each simulation can give rise to a different set of $\mathbf{w}_k$, we thus plot the mean ± s.d. of all the 50 $\langle r_+ \rangle$ for a given ρ.

soft-clustering membership coefficients of $\mathbf{x}^{(t)}$ (below).

Three of the four $\mathbf{w}_k$ align significantly with the $\mathbf{w}_{\text{LNtype}}$ (BT, BD, and P0, **Fig. 4A**). This result is robust for $\rho < 3.1$ (**Fig. 4B**): all numerical optimization converge to the same $\{\mathbf{y}^{(t)}\}$, $\{\mathbf{z}^{(t)}\}$, and $\{\mathbf{w}_k\}$ for the input $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and given ρ. This can partially be attributed to the non-negativity constraint in NNC, which removes an intrinsic symmetry of the LC model. Although $\mathbf{w}_{\text{KS}}$ is the least aligned to the found $\mathbf{w}_k$, NNC-5 has one $\mathbf{w}_k$ aligned with $\mathbf{w}_{\text{KS}}$ too (**Fig. S6H**). In summary, the ORN → LN connection weights predicted by the NNC model trained on ORN activity data $\{\mathbf{x}^{(t)}\}_{\text{data}}$ largely explain the $\mathbf{w}_{\text{LNtype}}$ of the connectome. Thus, several LNs are adapted to statistical features of these ORN activity patterns.

Emergence of LN groups in the NNC

In the connectome LNs are grouped by type and several $\mathbf{w}_{\text{LN}}$ are similar (**Fig. 1B-D, 4C**). Do LN groups naturally emerge in our model? In the LC, the $\{\mathbf{w}_k\}_{k=1\ldots K}$ spans a K-dimensional subspace (given enough independent dimensions in the input $\{\mathbf{x}^{(t)}\}$). All $\mathbf{w}_k$ are thus different. Therefore, in the LC, LN types emerge, but no similar LNs. In the NNC with small ρ, however, the objective function (18) leads to the symmetric non-negative matrix factorization (SNMF) objective function between $\{\mathbf{x}^{(t)}\}$ and $\{\mathbf{z}^{(t)}\}$ (**Supplementary Information**), which corresponds to a soft clustering of $\mathbf{x}^{(t)}$ by $\mathbf{z}^{(t)}$. Thus, each component in $\mathbf{z}^{(t)}$ discovers and encodes the presence of a sparse feature of $\mathbf{x}^{(t)}$ (Pehlevan & Chklovskii, 2015). In that case, when the number of significant sparse features in $\{\mathbf{x}^{(t)}\}$ is smaller than K, several components of $\mathbf{z}^{(t)}$ (i.e., LNs) encode a similar feature. Our simulations for NNC-8 ($K = 8$ as the number of LNs on each side of the larva) with $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and $\rho = 0.1$ indeed give rise to groups of similar $\mathbf{w}_k$ (**Fig. 4D**). Conversely, for larger ρ, the $\mathbf{w}_k$ become more decorrelated (**Fig. 4D**, $\rho = 10$). To study how the resemblance of the $\mathbf{w}_k$ changes with ρ, we calculated the average rectified correlation coefficient $\langle r_+ \rangle$ between all the $\mathbf{w}_k$ for different ρ (**Fig. 4D,E**). At $\rho = 0.35$, $\langle r_+ \rangle$ of the NNC-8 matched that of the connectome. This value of ρ should not, however, be interpreted as a "true" value for the actual biological circuit, because the true ORN activity patterns $\{\mathbf{x}^{(t)}\}$ that the larva experienced is unknown - in fact changing $\{\mathbf{x}^{(t)}\}$ and ρ are two independent means of influencing the model circuit synaptic weights. In summary, within reasonable parameter ranges, the NNC reproduces yet another property of the biological circuit: the emergence of LNs that can be grouped by type.

Relation between LN-LN and feedforward ORNs → LN connection weights

The ORN - LN circuit also contains inhibitory reciprocal LN - LN connections ($\mathbf{M} = \{m_{\text{LNi, LNj}}\}$, **Fig. 5A**) whose role is not fully understood. Our model predicts that $\mathbf{M}$ and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ are related thus (**Supplementary Information**):

$$\mathbf{M} \propto \sqrt{\mathbf{W}^\top \mathbf{W}} \tag{2}$$

Where $\top$ is the matrix transpose. This relationship is exact for the LC and approximate for the NNC. First, it predicts that the matrix $\mathbf{M}$ is symmetric, i.e., that the synaptic weight of LN$_i$ → LN$_j$ is equal to that of LN$_j$ → LN$_i$. This is indeed approximately true in the connectome, except for the P0, which inhibits KS, but is not strongly inhibited by them (**Fig. 5A**). Second, as predicted by the relationship (2), we find, in the connectome, a significant correlation between the entries of $\mathbf{M}$ and $\sqrt{\mathbf{W}^\top \mathbf{W}}$ for the left and right sides of the larva (excluding the diagonal entries, since the connectome does not provide the values corresponding to the diagonal of $\mathbf{M}$ of the model circuit) (**Fig. 5**). This suggests that the ORN-LN and LN-LN connections are meticulously co-organized to perform the circuit's function. Intuitively, LN-LN interaction could be interpreted as LNs competing with each other for activation. During circuit learning, without

LN-LN connections, all LNs would learn the same most significant direction of the input data. Thus, these lateral connections ensure that LNs span more than a single direction of the ORN activity space. After learning, LN-LN connections constitute an essential part of the computation (below, **Fig. S11**).

In summary, the NNC model accurately predicts several key features of the connectome: the $\mathbf{w}_{\mathrm{LNtype}}$ connection weights, the emergence of LN groups, and the relationship between ORNs → LN and LN - LN connections weights.

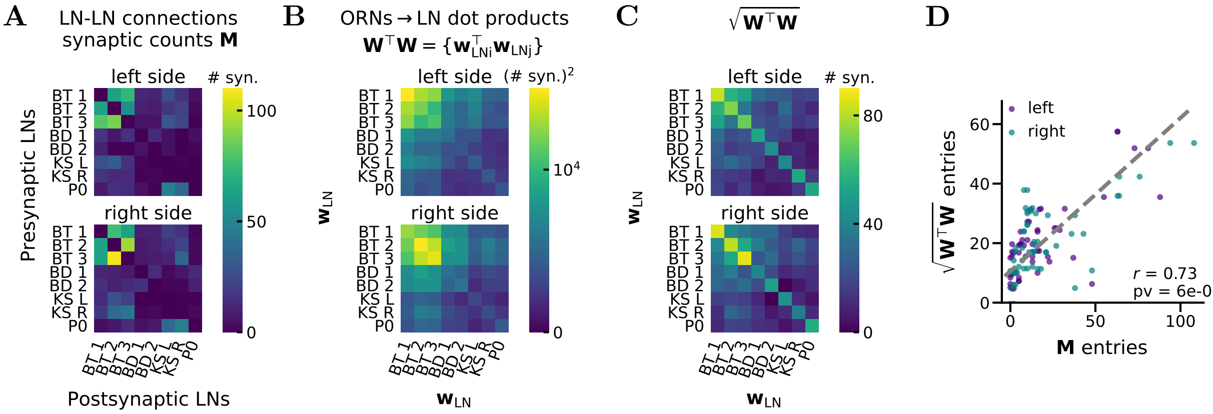

Fig. 5. Relation between LN-LN ($\mathbf{M}$) and ORNs → LN ($\mathbf{W}$) synaptic counts in the connectome reconstruction

A LN-LN connections synaptic counts $\mathbf{M}$ on the left and right sides of the larva.

B $\mathbf{W}^\top\mathbf{W}$ with $\mathbf{W} = [\mathbf{w}_{\mathrm{LN1}}, ..., \mathbf{w}_{\mathrm{LN8}}]$ on the left and right sides. Thus each entry is $\mathbf{w}_{\mathrm{LNi}}^\top \mathbf{w}_{\mathrm{LNj}}$, the scalar product between 2 ORNs →LN connection weight vectors $\mathbf{w}_{\mathrm{LN}}$.

C $\sqrt{\mathbf{W}^\top\mathbf{W}}$, i.e., the square root of the matrices in (**B**).

D Entries of $\mathbf{M}$ vs entries of $\sqrt{\mathbf{W}^\top\mathbf{W}}$, excluding the diagonal, for both sides. r: Pearson correlation coefficient. One-sided p-value calculated by shuffling the entries of each $\mathbf{w}_{\mathrm{LN}}$.

Computation in the LC: partial equalization of PCA variances in ORN axons and extraction of principal subspace by LNs

Next, we examine the computation performed by the LC model. The computation is implemented dynamically through the ORN - LN loop and converges exponentially to a steady state (equation (19)). Given inputs $\left\{\mathbf{x}^{(t)}\right\}$, we consider the twofold output of the circuit: the converged representations in ORN axons $\left\{\mathbf{y}^{(t)}\right\}$ and in LNs $\left\{\mathbf{z}^{(t)}\right\}$, both transmitted downstream. Although LNs are usually thought of only performing local computations, here LNs also project to several types of neuron like uni- and multi-glomerular PNs (Berck et al., 2016). Because the circuit is adapted to its input $\left\{\mathbf{x}^{(t)}\right\}$, the transformations from $\mathbf{x}^{(t)}$ to $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ are related to the statistics of $\left\{\mathbf{x}^{(t)}\right\}$ and are naturally expressed using the PCA directions $\{\mathbf{u}_i\}$ and variances $\left\{\sigma_{X,i}^2\right\}$ $(i = 1, ..., D)$ of uncentered $\left\{\mathbf{x}^{(t)}\right\}$. Formally, given the autocorrelation matrix $\boldsymbol{\Sigma}_X := \mathbf{E}\left[\mathbf{x}^{(t)}\mathbf{x}^{(t)\top}\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^{(t)}\mathbf{x}^{(t)\top} = \sum_{i=1}^{D}\sigma_{X,i}^2\mathbf{u}_i\mathbf{u}_i^\top = \mathbf{U}\boldsymbol{\Lambda}_X^2\mathbf{U}^\top$, $\sigma_{X,i}^2$ and $\mathbf{u}_i$ are the eigenvalues and eigenvectors of $\mathbf{\Sigma}_X$, respectively ($\sigma_{X,i}\sqrt{T} = s_{X,i}$ is also the i$^{\text{th}}$ singular value of $\left\{\mathbf{x}^{(t)}\right\}$), $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_D]$, and $\mathbf{\Lambda}_X = \text{diag}(\sigma_{X,1}, ..., \sigma_{X,D})$. We write the odor representations in ORN somas in this basis and find (**Supplementary Information**):

$$\mathbf{x}^{(t)} = \sum_{i=1}^{D} v_i^{(t)} \sigma_{X,i} \mathbf{u}_i \tag{3}$$

$$\mathbf{y}^{(t)} = \sum_{i=1}^{D} v_i^{(t)} \sigma_{Y,i} \mathbf{u}_i = \sum_{i=1}^{D} \frac{\sigma_{Y,i}}{\sigma_{X,i}} \mathbf{u}_i \mathbf{u}_i^\top \mathbf{x}^{(t)} \tag{4}$$

$$\mathbf{z}^{(t)} = \mathbf{Q} \sum_{i=1}^{K} v_i^{(t)} \frac{\rho}{\gamma} \sigma_{Y,i} \mathbf{u}_i = \mathbf{Q} \sum_{i=1}^{K} \frac{\rho}{\gamma} \frac{\sigma_{Y,i}}{\sigma_{X,i}} \mathbf{u}_i \mathbf{u}_i^\top \mathbf{x}^{(t)} \tag{5}$$

$$\text{with} \begin{cases} \sigma_{Y,i}\left(1 + \rho^2 \sigma_{Y,i}^2\right) = \sigma_{X,i} & 1 \le i \le K \qquad (6a) \\ \sigma_{Y,i} = \sigma_{X,i} & K+1 \le i \le D \qquad (6b) \end{cases}$$

where $v_i^{(t)} = \frac{1}{\sigma_{X,i}} \mathbf{u}_i^\top \mathbf{x}^{(t)}$ are the coefficients of $\mathbf{x}^{(t)}$ in the orthogonal basis $\{\sigma_{X,i}\mathbf{u}_i\}$ and $\mathbf{Q}$ is a $(K \times K)$ orthonormal (rotation) matrix and is a degree of freedom of the optimization.

On the dataset level, we find $\mathbf{\Sigma}_Y = \sum_{i=1}^{D} \sigma_{Y,i}^2 \mathbf{u}_i \mathbf{u}_i^\top = \mathbf{U}\mathbf{\Lambda}_Y^2\mathbf{U}^\top$ where $\mathbf{\Lambda}_Y = \text{diag}(\sigma_{Y,1}, ..., \sigma_{Y,D})$. Thus, the activity patterns in ORN axons $\left\{\mathbf{y}^{(t)}\right\}$ have the same principal directions $\{\mathbf{u}_i\}$ as $\left\{\mathbf{x}^{(t)}\right\}$ but with modified PCA variances (portrayed in **Fig. 6A,B** with $D = 2$ and $K = 1$). The variances of the last $D - K$ PCA directions of $\left\{\mathbf{x}^{(t)}\right\}$ remain unaltered in $\left\{\mathbf{y}^{(t)}\right\}$, whereas the variances of top K directions (as the number of LNs) are diminished according to equation (6a) (**Fig. 6C,D**), because LNs ($\left\{\mathbf{z}^{(t)}\right\}$) encode (a rotated version of) the top K principal subspace of $\left\{\mathbf{x}^{(t)}\right\}$ (equation (5)) and inhibit it in the ORN axons ($\left\{\mathbf{y}^{(t)}\right\}$). From the top K principal directions, those with relatively large variances are shrunken with a cubic root ($\sigma_{Y,i} \approx \sqrt[3]{\sigma_{X,i}/\rho^2}$), whereas those with relatively small variance remain virtually unchanged. Indeed, in the latter case, LNs are weakly activated and inhibition is almost inexistent.

For a LC with the same number of LNs as ORNs (i.e., $D = K$), this computation leads to a flatter spectrum of $\left\{\sigma_{Y,i}^2\right\}$ relatively to the one of $\left\{\sigma_{X,i}^2\right\}$, which can be quantified by the coefficient of variation, CV_σ (**Supplementary Information**). Although for $K < D$ only the top K principal direction are shrunken, in most cases it also leads to a decrease of CV_σ (see below).

This computation is a partial (Zero-phase) ZCA-whitening. By definition, a multivariate random variable $\mathbf{A}$ is white if its autocovariance matrix is proportional to the identity matrix: $\mathbf{E}\left[(\mathbf{A} - \mathbf{E}[\mathbf{A}])(\mathbf{A} - \mathbf{E}[\mathbf{A}])^\top\right] \propto \mathbf{I}$, which implies that all the PCA variances (i.e., eigenvalues of the autocovariance matrix) are equal. For the LC, the CV_σ of $\left\{\sigma_{Y,i}^2\right\}$ is smaller than the CV_σ of $\left\{\sigma_{X,i}^2\right\}$ (see also **Fig. 7E** below). Although these are formally the variances of the PCA on uncentered data, because the mean of $\left\{\mathbf{x}^{(t)}\right\}_{\text{data}}$ is close to $\mathbf{0}$, flattening the spectrum of $\left\{\sigma_i^2\right\}$ causes the flattening of the spectrum of the eigenvalues of the autocovariance matrix too, leading to partial whitening. Finally, since the principal directions of $\left\{\mathbf{y}^{(t)}\right\}$ and $\left\{\mathbf{x}^{(t)}\right\}$ are the same, the

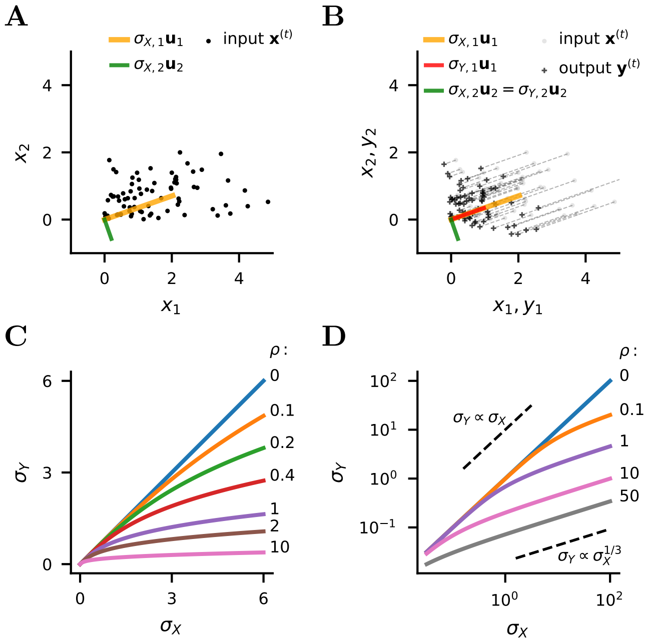

Fig. 6. Computation in the LC

A Example dataset $\{\mathbf{x}^{(t)}\}$ with $D = 2$ generated randomly from a zero-centered multivariate Gaussian and by removing points with negative coordinates. Depicted the PCA directions of $\{\mathbf{x}^{(t)}\}$ multiplied by the s.d. of that direction.

B Transformation from $\{\mathbf{x}^{(t)}\}$ to $\{\mathbf{y}^{(t)}\}$ by LC-1 ($K = 1$) with $\rho = 1$. Depicted the PCA directions of $\{\mathbf{x}^{(t)}\}$ and $\{\mathbf{y}^{(t)}\}$ multiplied by the s.d. of that direction.

C-D Transformation of the s.d. of PCA directions from $\{\mathbf{x}^{(t)}\}$ to $\{\mathbf{y}^{(t)}\}$ in the LC on linear and logarithmic axes.

transformation contains no rotation and is thus “zero-phase”, as ZCA-whitening.

LC and NNC computation on the ORN activity dataset

Finally, to elucidate the computation of this circuit on odor representations, we study the computation of the LC and the NNC on $\{\mathbf{x}^{(t)}\}_{\text{data}}$. We set the parameter regulating the strength of the inhibition $\rho = 2$ to distinctly portray the input-output transformation. Given the input of ORN activities $\{\mathbf{x}^{(t)}\}_{\text{data}}$, we calculate $\{\mathbf{y}^{(t)}\}$ and $\{\mathbf{z}^{(t)}\}$ with $K = 1$ and $K = 8$ using the analytical formula for the LC and by optimizing the objective function (18) for the NNC.

In the LC, LNs encode the top K principal subspace of $\{\mathbf{x}^{(t)}\}$ (above, **Fig. S7B**). In the NNC, the computation in LNs approximates SNMF for small ρ (**Supplementary Information**) which performs soft clustering and sparse feature discovery (Pehlevan & Chklovskii, 2015). LNs thus encode features of the odor representations in ORN (**Fig. S7C-G**), that are transmitted to downstream brain areas.

Next we show that in LC and NNC the transformation from $\{\mathbf{x}^{(t)}\}$ to $\{\mathbf{y}^{(t)}\}$ is a partial ZCA-whitening and a divisive normalization as reflected in the partial equalization of the PCA variances

(**Fig. 7E**), the decrease of channel (i.e., ORN) and pattern (i.e., neural representations of odors) correlations (**Fig. 7J-O, S9**), and the lack of rotation of the output (**Fig. S8E**). **Fig. 7A-C** shows the activity in ORN somas and the computed activity in ORN axons for LC-8 and NNC-8. The LC produces strongly negative values in $\{\mathbf{y}^{(t)}\}$, which might not be biologically plausible. We next compared the spectrum of $\left\{\sigma^2_{X,i}\right\}$ and $\left\{\sigma^2_{Y,i}\right\}$, since this characterizes whitening and the computation in LC affects this aspect (**Fig. 7D**). As expected, in the LC only the top K principal directions of the input are dampened. For the NNC, however, we find that all directions are dampened, even for $K = 1$. This can be attributed to the non-negativity constraint on the output $\{\mathbf{y}^{(t)}\}$ and $\{\mathbf{z}^{(t)}\}$ in NNC, which potentially affects all stimuli directions. We find a flattening of $\left\{\sigma^2_{Y,i}\right\}$ spectrum both in LN and NNC as seen in the smaller CV_σ (**Fig. 7E**) demonstrating that $\{\mathbf{y}^{(t)}\}$ is more white that $\{\mathbf{x}^{(t)}\}$. Changing the number of LNs does not affect the NNC as much as the LC. However, changing ρ greatly influences the strength of the dampening (**Fig. S10**). Although in the LC the principal directions of $\{\mathbf{x}^{(t)}\}$ and $\{\mathbf{y}^{(t)}\}$ remain the same, their order changes, because only a fraction of them are shrunken (**Fig. S8A,B**). For the NNC, however, there is only a slight mixing between principal directions of similar strength, but their order mainly remains (**Fig. S8C,D**).

As expected from a flatter $\{\sigma_{Y,i}\}$, we observe that channels and patterns are more decorrelated in the output $\{\mathbf{y}^{(t)}\}$ in the NNC (**Fig. 7J-O**) and in the LC (**Fig. S9**) than in the input, which is coherent with partial whitening. The strength of decorrelation increases with ρ (**Fig. S10**).

Next, we study the effect of the circuit computation on channel and pattern activity Euclidean norms, which reflect the total channel and total pattern activity. We find that both LC and NNC dampen the channels with strong norms and leave the weaker channels largely unaffected, thus decreasing the CV of channel norms (**Fig. 7F,G**). This allows the information to be more evenly distributed among channels, an important property of efficient coding. Similarly, the circuit partially equalizes the norms of activity patterns (**Fig. 7H,I**). This slightly removes the concentration information from the signal. These effects are similar to a divisive normalization-type computation, also reported in *Drosophila* (Carandini & Heeger, 2012; Olsen et al., 2010).

Finally, we aim at better understanding the role of LN-LN connections. We study the computations performed by the converged LC and NNC, with the off-diagonal elements in $\mathbf{M}$ set to 0 (**Fig. S11**). We find that this manipulation mixes the output principal direction in relation to the input and also increases the total level of inhibition. Thus, LN-LN connection helps to reduce the amount of rotation in the neural representation, regulate the amount of inhibition, and maintain the predicted computation.

In summary, the analysis of the LC and NNC predicts that the ORN-LN circuit performs the following computation on the odor representation in ORNs: it most strongly dampens the most prominent directions of the input dataset and thus flatten the PCA variance spectrum. This results in an output in ORN axons that is more white, decorrelated, and more equalized channels and patterns. This allows a more efficient neural representation and improves odor discrimination downstream.

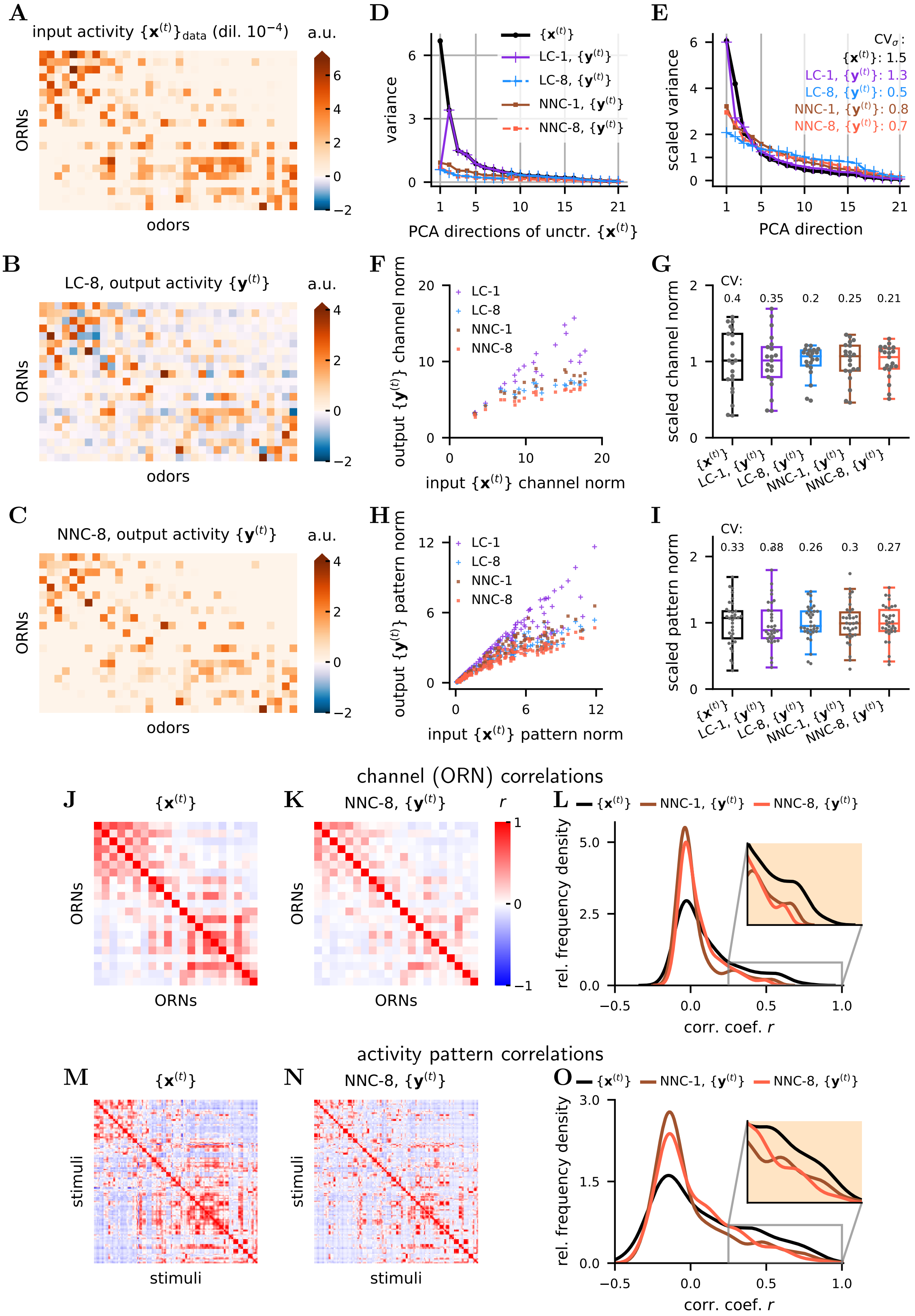

Fig. 7. Functional consequences of LC and NNC: partial whitening, normalization, decorrelation

(continued on next page)

Fig. 7. (continued)

A Input (ORN soma) activity patterns $\{\mathbf{x}^{(t)}\}_{\text{data}}$ for all odors at dilution 10^{-4}. Instead of $\Delta F/F_0$ as units of activity, we use arbitrary units (a.u.), which stand for appropriate activity units at the neurons level.

B Output $\{\mathbf{y}^{(t)}\}$ for the input of (**A**) for the LC-8.

C Same as (**B**) for NNC-8.

D Variances of $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and $\{\mathbf{y}^{(t)}\}$ in the principal directions of uncentered $\{\mathbf{x}^{(t)}\}$.

E PCA variances of $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and $\{\mathbf{y}^{(t)}\}$, scaled by their mean. $\{\mathbf{y}^{(t)}\}$ has a smaller span of variances than $\{\mathbf{x}^{(t)}\}$. See **Fig. S8** for the relation between the principal directions of $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and $\{\mathbf{y}^{(t)}\}$.

F Euclidean norm of the 21 channels in output $\{\mathbf{y}^{(t)}\}$ (ORN axons) vs in the input $\{\mathbf{x}^{(t)}\}_{\text{data}}$ (ORN somas).

G Box plot of the channel norms scaled by their mean, CV on top.

H Euclidean norm of the 170 activity pattern in output $\{\mathbf{y}^{(t)}\}$ vs in the input $\{\mathbf{x}^{(t)}\}_{\text{data}}$.

I Box plot of the activity patters norms (only for dilution 10^{-4}) scaled by their mean, CV on top.

J ORNs correlations in the input $\{\mathbf{x}^{(t)}\}_{\text{data}}$.

K ORNs correlations in the output $\{\mathbf{y}^{(t)}\}$ of the NNC with $K = 8$.

L Histogram for the channel correlation coefficients from (**J**-**K**), excluding the diagonal (n=210).

M Activity vector (i.e., pattern) correlation in $\{\mathbf{x}^{(t)}\}_{\text{data}}$.

N Activity vector correlation in $\{\mathbf{y}^{(t)}\}$ of NNC-8.

O Histogram for the pattern correlation coefficients from (**M**-**N**), only for dilution 10^{-4} (n=561) (see also **Fig. S9**). $\rho = 2$ in the whole figure.

Discussion

Combining the *Drosophila* larva olfactory circuit connectome, ORN activity data, and a new normative model, we advance the understanding of sensory computation and adaptation, quantitatively link ORN activity statistics, functional data and connectome, and make testable predictions. Our work uncovers and characterizes a simple and potent neural circuit architecture capable of adaptive data preprocessing and feature extraction, which, as an independent computational unit, could arise in other brain areas and be useful for machine learning and signal processing. Finally, our normative approach provides a general framework to understand circuit computation (Bahroun et al., 2019; Golkar et al., 2020) and could be applied to more connectomes (Eichler et al., 2017; Scheffer et al., 2020).

Circuit computation, partial ZCA-whitening, and divisive normalization

We propose that the circuit's effect on neural odor representation in ORNs correspond to partial ZCA-whitening and divisive normalization (DN) (**Fig. 6, 7**). Such computations, which reduce correlations originating from the sensory system and the environment, have appeared in efficient coding and redundancy reduction theories (Atick & Redlich, 1992; Barlow, 1961; Carandini & Heeger, 2012; Linsker, 1988; Plumbley, 1993; Simoncelli & Olshausen, 2001). Partial whitening is indeed a solution for mutual information maximization in the presence of input noise (Atick & Redlich, 1990). In this circuit too, we suggest that a pure whitening transformation might not be desirable, as it could lead to noise amplification. Thus, keeping low-variance signal directions of the input unchanged and damping larger ones might accord with mutual information maximization. Our conclusions are in line with reports of pattern decorrelation and/or whitening in the olfactory system in zebrafish (Friedrich, 2013; Friedrich & Laurent, 2001; Friedrich & Wiechert, 2014; Wanner & Friedrich, 2020) and mice (Giridhar et al., 2011; Gschwend et al., 2015).

Infinitely many whitening transformations exist - indeed, a rotated white signal remains white. ZCA-whitening, where the output is not rotated relatively to the input, might be advantageous over other flavors of whitening because it is the optimal whitening transform that minimizes the distance between the original and the whitened signal (Kessy et al., 2018). Since inputs (i.e., spike rates) are non-negative, this property of ZCA-whitening will reduce the amount of negative deviations and lessen the distortion of the computation that arises from the non-negative constraint on neural activity.

On the other hand, the computation in our model also resembles DN, a ubiquitous computation in the brain (Carandini & Heeger, 2012) which was suggested for the analogous circuit in the adult *Drosophila* (Olsen et al., 2010; Olsen & Wilson, 2008). In its simplest form, DN is defined as $Y_j = \gamma\frac{X_j^n}{\sigma^n+\sum_k X_k^n}$, where Y_j is the response of the neuron j, X_i is the driving input of the neuron i, and γ, σ, and n are positive parameters. DN captures two effects of neuronal and circuit computation: (1) the saturation of a neural response with increasing input up to a maximum spiking rate γ, which mainly arises from neuron's biophysical properties; (2) dampening of the response of a given neuron when other neurons also receive input, usually originating from lateral inhibition (but see Sato et al., 2016). In our model, aspect (1) of DN is absent, but could readily be implemented with a saturating non-linearity. However, signatures of (2) are especially apparent in the saturation of the pattern output norm for increasing input norm (**Fig. 7H**). This saturation occurs because inputs with higher norms correspond to inputs at higher odor concentrations and with a higher number of active ORNs. Because such input directions are more statistically significant in our dataset, these stimuli that are more strongly dampened by LNs (which encode those directions) than those with few ORNs active. Thus, our model presents a possible linear implementation of a crucial aspect of DN, which in itself is a nonlinear operation.

The basic form of DN equalizes the channels and performs channel decorrelation, but not pattern decorrelation (Friedrich & Wiechert, 2014; Olsen et al., 2010; Wanner & Friedrich, 2020), which appears in our model. However, a modified version of DN, which includes different coefficients for the driving inputs in the denominator (Westrick et al., 2016), performs pattern decorrelation too, as seen in our circuit. The proposed neural implementations of DN usually require a multiplication by the feedback (Heeger, 1992; Westrick et al., 2016), which might not be as biologically realistic as our circuit implementation.

Several neural architectures similar to ours have been proposed to learn to decorrelate channels, perform DN, or learn sparse representations in an unsupervised manner (Atick & Redlich, 1993; King et al., 2013; Koulakov & Rinberg, 2011; Olshausen & Field, 1997; Pehlevan & Chklovskii, 2015, 2016; Westrick et al., 2016; Wick et al., 2010; M. Zhu & Rozell, 2015). These studies, however, either do not have an objective function, or have a different circuit architecture or synaptic learning rules.

Roles of LNs

LNs form a significant part of the neural populations in the brain, have multiple crucial computational functions, and have extremely diverse morphologies and excitabilities (Chou et al., 2010; Hattori et al., 2017). We propose a dual role for LNs in this olfactory circuit: altering the odor representation in ORNs and extracting ORN activity features, which can be used downstream (Berck et al., 2016). In the olfactory system of *Drosophila* and zebrafish, LNs perform multiple roles like gain control, normalization of odor representations, pattern and channel decorrelation (Friedrich, 2013; Friedrich & Wiechert, 2014; Olsen et al., 2010; Olsen & Wilson, 2008; Wanner & Friedrich, 2020; P. Zhu et al., 2013), roles that are in line with our results. Also, in *Drosophila* the LN population expands the temporal bandwidth of synaptic transmission and temporally tune PN responses (Kim et al., 2015; Nagel et al., 2014; Nagel & Wilson, 2016), which was not addressed here.

In topographically organized circuits such as visual periphery or auditory cortex, several LN types uniformly tile the topographic space and each LN type has its own role and selectivity (e.g., in the retina (Masland, 2012)). In non-topographically organized networks, however, the organization and selectivity of LNs is still a matter of research and controversy (Chou et al., 2010; Hong & Wilson, 2015). We have included 4 LN types in the studied subcircuit (**Fig. 1**). Several LN types contains multiple copies of LNs, with similar connection weights, and thus presumably similar roles. In the LC model, the K LNs span a K-dimensional subspace of activity, thus each LN has a different connectivity and would form a type of its own. In the NNC model, large ρ lead to different LNs, whereas smaller ρ lead to the formation of LN groups (**Fig. 4C-E**). Thus based on our study and the different connectivity patterns of LNs in the connectome (Berck et al., 2016), we suggest that in the *Drosophila* larva LN types extract different features of ORN activity and are thus differently activated in response to different input directions (and glomeruli) and also different ORNs are differently inhibited by different LNs. This seems at odds with the results of Hong and Wilson, 2015 who found that the activation of the LN population appears invariant to odor identity. However, the latter study imaged several LNs simultaneously and thus might have missed the selectivity of individual LNs.

What are the features being extracted by LNs? The Broad Trio, whose connection weight vector aligns to the first PCA direction of ORN activity and to a $\mathbf{w}$ of the NNC model (**Fig. 3, 4A,B**), could potentially encode the mean ORN activity, and thus be related to the global odor concentration (Asahina et al., 2009). Other LNs, whose connectivity aligns with the $\mathbf{w}$ of the NNC model, might encode features of odors, like aromatic vs long carbon chain (Si et al., 2019), or specific information influencing larva behavior (Berck et al., 2016). What is the function of multiple "copies" of LNs within each type? Firstly, LNs might differentiate further as the larva grows, and as the circuit continues learning. Secondly, several LNs might help expand the dynamical range of a single LN.

The connectome reveals that the circuit also includes LN-LN connections, which arise naturally in our approach. We suggest that LN-LN connections constitute a crucial part of learning and LN differentiation, as well as performing partial ZCA-whitening and normalization. Our model also correctly predicted how LN-LN connections co-organize with the ORN-LN connections (**Fig. 5**). To our knowledge, the role of LN-LN connections and their relationship to ORN-LN connections has not been addressed previously in such circuits.

In summary, our study highlights the significance of the different ORN-LN and LN-LN connection strengths and argues that LNs are minutely selective and organized to extract features and render the representation of odors more efficient.

Learning and ORN activity statistics

Using ORN activity dataset (Si et al., 2019), our NNC model could predict to a large extent the connection weight vectors found in the connectome (**Fig. 4A-B**). This suggests that the circuit is adapted to ORN activity patterns (**Fig. 2, 3, 4**). How could the connectivity prediction be successful, when the ORN activity dataset was mainly chosen to uniformly and broadly activate all

ORNs and not to match the true larva odor environment, in terms of odor identity, frequency, and intensity? One possibility is that, given an ORN activity dataset large enough, certain generic correlations between ORNs always appear, giving rise to the same robust features in the connectivity. These correlations could be caused by intrinsic chemical properties of ORN receptors. Moreover, the exact odor statistics would also alter the connection weights, but to a lesser extent than the former effect. Thus, given an activity dataset closely mimicking the larva natural odor environment, the model predictions of the connectome might further improve.

Are those synaptic weights learned during the animal lifetime or are they encoded genetically, i.e., "learned" over an evolutionary time span? A genetic origin is undoubtedly present, given that several LNs types (e.g., Keystone and Picky) differ by their connectivity to specific neurons outside the studied circuit and seem to be linked to different hard-wired animal behaviors (Berck et al., 2016). Additionally, several studies reveal that glomeruli sizes (and thus ORN-LN or ORN-PN synaptic weights) or activity vary depending on the environment where the *Drosophila* grows up (Arenas et al., 2012; Das et al., 2011; Devaud et al., 2001; Sachse et al., 2007; Sudhakaran et al., 2012). This feature would equip the circuit with a potent mechanism to adapt to evolving natural environment. Additionally, synaptic count and innervation variability arises for *Drosophila* brought up in similar environments (Chou et al., 2010; Tobin et al., 2017), indicating the potential imprecision of the development and/or learning. Resolving connectomes of larva raised in different odor environments, probing the synaptic plasticity present in the network, and recording ORN responses to the full ensemble of odors present in its environment would help clarify the influence of learning and of genetics.

In conclusion, our work uncovers a canonical circuit model that could robustly adapt to different environments in an unsupervised manner, while maintaining the critical computations of partial whitening, normalization, and feature extraction. Our comprehensive normative approach, which contains only one effective parameter, predicted the structural organization based on input activity, and found in the connectome the signatures of circuit function and adaptation to ORN pattern statistics. Such an approach could provide important insights into more complicated adaptive neural circuits, whose structural and activity data is becoming available.

Methods

ORN activity

We use the average maximal Ca^{2+} $\Delta F/F_0$ responses among trials for the activity data as in Si et al., 2019. For the ORN 85c in response to 2-heptanone, and for the ORN 22c in response to methyl salicylate, we only have responses to dilutions $\leq 10^{-7}$. Because the ORN responses are very similar for dilutions 10^{-7} and 10^{-8} and are already saturated (for this cell we have responses down to dilutions of 10^{-11}), we set the missing response for dilutions 10^{-6}, 10^{-5} and 10^{-4} as the response for 10^{-7}.

RCF distribution of correlation coefficient and significance testing

Given a vector $\mathbf{a} \in \mathbb{R}^D$, we define the mean $\bar{a}$, the centered vector $\mathbf{a}_c$, and the centered normalized vector $\widehat{\mathbf{a}}$:

$$\bar{a} := \frac{1}{D}\sum_{i=1}^{D} a_i \tag{7}$$

$$\mathbf{a}_c := \mathbf{a} - \bar{a} \tag{8}$$

$$\widehat{\mathbf{a}} := \frac{\mathbf{a}_c}{||\mathbf{a}_c||} \tag{9}$$

We call $\widehat{\mathbf{w}} \in \mathbb{R}^D$ the centered and normalized ORNs → LN synaptic weight vector $\mathbf{w}$. Similarly, we define $\widehat{\mathbf{X}} \in \mathbb{R}^{D\times T}$ the centered and normalized ORN activity $\mathbf{X}_{\text{data}} = \left[\mathbf{x}^{(1)}, ..., \mathbf{x}^{(T)}\right]$, where each column vector is centered and normalized.

Each row of the matrix of correlation coefficients depicted in **Fig. 2E** is given by $\mathbf{c} := \widehat{\mathbf{w}}_{\text{LNtype}}^{\top}\widehat{\mathbf{X}}$. $\mathbf{c}$ is used to calculate the true relative cumulative frequency (RCF) of correlation coefficients in **Fig. 2G**: $\mathrm{RCF}_c(x) := \frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{[-1,x]}(c_i)$, where $\mathbf{1}_A(y)$ is the indicator function of a given set A.

We define the random variables $\mathbf{w}'$, $\mathbf{c}'$ and RCF'. $\mathbf{w}'$ is generated by shuffling the entries of a connectivity vector $\widehat{\mathbf{w}}$:

$$w_i' := w_{\sigma(i)} \tag{10}$$

$$\mathbf{c}' := \widehat{\mathbf{w}}'^{\top}\widehat{\mathbf{X}} \tag{11}$$

$$RCF_c'(x) := \frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{[-1,x]}(c_i') \tag{12}$$

Where $\sigma(i)$ is a random permutation operator. We define $\overline{RCF}'(x)$ (**Fig. 2G**, black line) as the mean $RCF'(x)$ arising from all RCFs that come from shuffled $\widehat{\mathbf{w}}$. Next, we define, the maximum negative deviation δ' random variable as:

$$\delta' := \max_x \left[\overline{RCF}'(x) - RCF'(x) \right] \tag{13}$$

Finally, we define p-value $= \Pr(\delta' \geq \delta_{true})$. The p-value is thus the proportion of RCFs generated with random shuffling of entries of $\widehat{\mathbf{w}}$ that deviate from the mean RCF more than the true RCF.

Numerically, these calculations were done by binning the RCF function into 0.02 bins and generating 10000 instances of shuffled $\widehat{\mathbf{w}}$.

Number of aligned dimensions between two subspaces

Given a Hilbert space of dimension D, we define Ω - a measure of dissimilarity between 2 subspaces $\mathbf{S}_A$ and $\mathbf{S}_B$ generated by the matrices of linearly independent K_A and K_B column vectors: $\mathbf{A} \in \mathbb{R}^{D \times K_A}$ and $\mathbf{B} \in \mathbb{R}^{D \times K_B}$:

$$\begin{aligned}
\Omega :=& \|\mathbf{P}_A - \mathbf{P}_B\|_F^2 && (14)\\
=& \operatorname{Tr}\left[\mathbf{P}_A^2\right] + \operatorname{Tr}\left[\mathbf{P}_B^2\right] - 2\operatorname{Tr}\left[\mathbf{P}_A\mathbf{P}_B\right] = \dim\left[\mathbf{S}_A\right] + \dim\left[\mathbf{S}_B\right] - 2\operatorname{Tr}\left[\mathbf{P}_A\mathbf{P}_B\right] && (15)\\
=& K_A + K_B - 2\operatorname{Tr}\left[\mathbf{P}_A\mathbf{P}_B\right] && (16)
\end{aligned}$$

Where $\mathbf{P}_A, \mathbf{P}_B \in \mathbb{R}^{D \times D}$ are the projectors onto the subspaces S_A and S_B, respectively, F stands for the Frobenius norm, Tr is the matrix trace, and $K_X = \dim(\mathbf{S}_X)$ is the dimensionality of a subspace S_X. We assume $K_A + K_B \leq D$. We have that $|K_A - K_B| \leq \Omega \leq K_A + K_B$. The projection matrix can be obtained thus $\mathbf{P}_A = \mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$, or via QR factorization: $\mathbf{QR} = \mathbf{A}$, $\mathbf{P}_A = \mathbf{QQ}^\top$.

Intuitively, for two very similar subspaces, the projection $\mathbf{P}_A v$ of an arbitrary vector v onto S_A will be very similar to the projection $\mathbf{P}_B v$ vector v onto S_B, thus $\mathbf{P}_A v \approx \mathbf{P}_B v$ and Ω will be small. Conversely, if the subspaces are very different, the projections $\mathbf{P}_A v$ and $\mathbf{P}_B v$ will also be different and Ω will be large.

We now define the more intuitive measure:

$$\Gamma := \left(K_A + K_B - \Omega\right)/2 \tag{17}$$

which is a proxy of the number of aligned dimensions in the two subspaces. Indeed $0 \leq \Gamma \leq \min(K_A, K_B)$. For 2 perpendicular subspaces, $\Gamma = 0$ and for 2 fully aligned subspaces $\Gamma = \min(K_A, K_B)$.

In the main text we have $\mathbf{A} = [\mathbf{w}_{\mathrm{BT}}, \mathbf{w}_{\mathrm{BD}}, \mathbf{w}_{\mathrm{KS}}, \mathbf{w}_{\mathrm{P0}}]$ and $\mathbf{B}$ is the matrix with the top 5 PCA loading vectors of $\{\mathbf{x}^{(t)}\}$ as columns, $K_A = \dim\left[\mathbf{S}_A\right] = 4$, $K_B = \dim\left[\mathbf{S}_B\right] = 5$ and $D = 21$.

Objective function for the ORN-LN circuit

We choose a normative-theoretical approach to study the ORN-LN circuit. It has the advantage of providing analytical expressions describing different aspects of the computation and the circuit architecture. Studying the circuit's computation is then equivalent to studying the optimum of a cost function.

We first define the following variables: an input $\mathbf{X} = \left[\mathbf{x}^{(1)}, ..., \mathbf{x}^{(T)}\right]$ of T samples, and outputs $\mathbf{Y} = \left[\mathbf{y}^{(1)}, ..., \mathbf{y}^{(T)}\right]$, $\mathbf{Z} = \left[\mathbf{z}^{(1)}, ..., \mathbf{z}^{(T)}\right]$. $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ are D-dimensional vectors, whereas $\mathbf{z}^{(t)}$ are K-dimensional. $\mathbf{x}^{(t)}$, $\mathbf{y}^{(t)}$, and $\mathbf{z}^{(t)}$ represent the activity of ORN somas (i.e., the inputs), ORN axons and K LNs, respectively. We postulate the following similarity-based objective function (e.g., Pehlevan et al., 2018), which links the steady state activity of the outputs to that of the input:

$$\mathcal{L} = \min_{\mathbf{Y}\geq 0} \max_{\mathbf{Z}\geq 0} \frac{1}{T^2}\left(\frac{T}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 - \frac{\rho^2}{4}\left\|\mathbf{Y}^\top\mathbf{Y} - \frac{\gamma^2}{\rho^2}\mathbf{Z}^\top\mathbf{Z}\right\|_F^2 + \frac{\rho^2}{4}\left\|\mathbf{Y}^\top\mathbf{Y}\right\|_F^2\right) \tag{18}$$

Intuitively this objective function drives the activity of the ORN axons $\mathbf{Y}$ to be close to the activity of ORN somas $\mathbf{X}$ through the term $\|\mathbf{X}-\mathbf{Y}\|_F^2$, it aligns the similarity between the activity of ORN axons and LNs through the term $\left\|\mathbf{Y}^\top\mathbf{Y} - \frac{\gamma^2}{\rho^2}\mathbf{Z}^\top\mathbf{Z}\right\|_F^2$, and finally puts a 4th order penalty on the norm of $\mathbf{Y}$ through the term $\left\|\mathbf{Y}^\top\mathbf{Y}\right\|_F^2$. ρ and γ are two parameters. Scaling ρ is related to the strength of the dampening in $\mathbf{Y}$ and affects both the optima of $\mathbf{Y}$ and $\mathbf{Z}$. Changing γ only scales $\mathbf{Z}$, without affecting $\mathbf{Y}$. Since γ does not fundamentally change the computation, we set $\gamma = 1$ in the whole paper.

We consider two objective functions. One without the non-negativity constraints on $\mathbf{Y}$ and $\mathbf{Z}$, representing the Linear Circuit (LC) model, and one with the non-negativity constrains as in equation (18), representing the Non-Negative Circuit (NNC) model. Non-negativity constraints account for the fact that neural activity is usually non-negative, or at least not symmetric in the negative and positive directions.

In order to map the objective function to a neural circuit (**Supplementary Information**), we first introduce two auxiliary matrices $\mathbf{W} = \frac{1}{T}\mathbf{Y}\mathbf{Z}^\top$ and $\mathbf{M} = \frac{1}{T}\mathbf{Z}\mathbf{Z}^\top$, which naturally map onto ORNs - LNs and LNs - LNs synaptic weights, respectively. The objective function is thus optimized over the variables $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{W}$, and $\mathbf{M}$. We then consider the objective function in the "online setting". In this situation one $\mathbf{x}^{(t)}$ is presented at a time, the optimal $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ are found with the current $\mathbf{W}$ and $\mathbf{M}$, and subsequently the $\mathbf{W}$ and $\mathbf{M}$ are updated. The optimal $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ are found with gradient descent/ascent equations, which also correspond to the ORN-LN neural dynamics equations ((19) for the LC or (20) for the NNC). The gradient descent/ascent steps on $\mathbf{W}$ and $\mathbf{M}$ correspond to the Hebbian learning update rules equation (21).

Circuit neural dynamics

When optimized online, the objective function (18) without the non-negativity constraints gives rise to the following differential equations describing the LC, whose steady state solutions correspond to the optima for $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ (**Supplementary Information**). These equations naturally map onto the ORN-LN neural circuit dynamics (dropping the sample index t for simplicity of notation):

$$\begin{cases} \tau_y \dfrac{d\mathbf{y}(\tau)}{d\tau} = -\mathbf{y}(\tau) - \gamma^2 \mathbf{W}\mathbf{z}(\tau) + \mathbf{x} \\ \tau_z \dfrac{d\mathbf{z}(\tau)}{d\tau} = -\mathbf{M}\mathbf{z}(\tau) + \rho^2/\gamma^2 \mathbf{W}^\top \mathbf{y}(\tau) \end{cases} \tag{19}$$

Where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are D, D, and K-dimensional vectors, and represent the activity (e.g., spiking rate) of the ORN somas, ORN axons, and LNs, respectively. τ_y and τ_z are neural time constants, τ is the local time evolution (not to be confused with the t sample index). The elements of the $D \times K$ matrices $\rho^2/\gamma^2\mathbf{W}$ and $\gamma^2\mathbf{W}$ contain the synaptic weights of the feedforward ORNs → LN and feedback LN → ORNs connections, respectively. Thus, the feedforward connection vectors are proportional to the feedback vectors, with a scaling factor ρ^2/γ^4. This assumption is reasonable considering the connectivity data (**Fig. S1, S2B**). Off-diagonal elements of the $K \times K$ matrix $\mathbf{M}$ contain the weights of LN - LN inhibitory connections, whereas the diagonal elements are related to the LNs leak. In the absence of LN activity and at steady state, the equations satisfy $\mathbf{y} = \mathbf{x}$, i.e., ORN soma and axonal activities are identical. In the absence of input (i.e., $\mathbf{x} = 0$) both $\mathbf{y}$ and $\mathbf{z}$ decay exponentially to $\mathbf{0}$, because of the terms $-\mathbf{y}(\tau)$ and $-M_{i,i}z_i(\tau)$, respectively. In summary, these equations effectively model the ORN-LN circuit dynamics by implementing that (1) the ORN axonal activity is driven by the input in ORN somas $\mathbf{x}$ and inhibited by the feedback from the LNs thought the term $-\gamma^2\mathbf{W}\mathbf{z}(\tau)$ and (2) LN activity is driven by the activity in ORN axonal terminals by $\rho^2/\gamma^2\mathbf{W}^\top\mathbf{y}(\tau)$ and inhibited by LNs through the term $-\mathbf{M}\mathbf{z}(\tau)$. ρ and γ are two parameters. In fact, a general system of differential equations describing this circuit architecture can be reduced to having just two parameters (**Supplementary Information**). Scaling ρ affects both the steady state solution of $\mathbf{y}$ and $\mathbf{z}$, whereas scaling γ only scales $\mathbf{z}$. Note that changing ρ in the objective function, will also give rise to different optimal $\mathbf{W}$ and $\mathbf{M}$.

When optimized online, the objective function (18) with the non-negativity constraints gives rise to the following equations describing the NNC:

$$\begin{cases} \mathbf{y}(\tau+1) = \max\Big[\mathbf{0},\ \mathbf{y}(\tau) + \epsilon(\tau)\big(-\mathbf{y}(\tau) - \gamma^2\mathbf{W}\mathbf{z}(\tau) + \mathbf{x}\big)\Big] \\ \mathbf{z}(\tau+1) = \max\Big[\mathbf{0},\ \mathbf{z}(\tau) + \epsilon(\tau)\big(-\mathbf{M}\mathbf{z}(\tau) + \rho^2/\gamma^2\mathbf{W}^\top\mathbf{y}(\tau)\big)\Big] \end{cases} \tag{20}$$

Where $\epsilon(\tau)$ is the step size parameter and the max is performed component wise. Here τ is a discrete time variable. These equations can be seen as the equivalent to equations (19), but also satisfying constraints on the activity, such as $y_i(\tau) \geq 0$, $z_i(\tau) \geq 0$, $\forall \tau, i$. Such constraints are implemented by formulating circuit dynamics in discrete time and using a projected gradient descent.

We call LC-K the linear circuit implemented by (19) and NNC-K the non-negative circuit implemented by (20), with K LNs. The actual biological circuit might exhibit a behavior somewhere between the LC and NNC. For the circuit studied here, we have $D = 21$ (number of ORNs), and $K = 8$ (number of LNs on one side of the larva) or $K = 4$ (number of LN types) or $K = 1$ (to build intuition).

Mathematical description of synaptic plasticity

When the objective function (18) is optimized online, we obtain the following updates for $\mathbf{W}$ and $\mathbf{M}$ after each presentation of a sample $\mathbf{x}^{(t)}$ and convergence to optimal $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$:

$$\begin{aligned}\mathbf{W}^{(t+1)} &= \mathbf{W}^{(t)} + \epsilon_1(t)\left(\mathbf{y}^{(t)}\mathbf{z}^{(t)\top} - \mathbf{W}^{(t)}\right)\\ \mathbf{M}^{(t+1)} &= \mathbf{M}^{(t)} + \epsilon_2(t)\left(\mathbf{z}^{(t)}\mathbf{z}^{(t)\top} - \mathbf{M}^{(t)}\right)\end{aligned} \tag{21}$$

Where $\epsilon_i(t)$ are learning rates. We assume that the ORN soma activation $\mathbf{x}^{(t)}$ in present long enough so that $\mathbf{y}^{(t)}(\tau)$ and $\mathbf{z}^{(t)}(\tau)$ reach steady state values. These equations represent Hebbian plasticity in $\mathbf{W}$ and $\mathbf{M}$, which is a form of correlative unsupervised learning. This is justified by (1) the adaptation of the connectivity to statistics of the ORN activity found in our data, (2) the presence of activity-dependent plasticity in *Drosophila* (Arenas et al., 2012; Das et al., 2011; Devaud et al., 2001; Sachse et al., 2007; Sudhakaran et al., 2012), and (3) that glomeruli activity is best explained with glomerulus-glomerulus inhibitory connectivity that is proportional to the correlation between glomeruli (Linster et al., 2005). These equations (21) set the diagonal values of $\mathbf{M}$ by analogy to the off-diagonal ones.

With appropriate learning rates, these synaptic update rules lead to:

$$\mathbf{W} \to \mathbf{E}\left[\bar{\mathbf{y}}\bar{\mathbf{z}}^\top\right], \quad \mathbf{M} \to \mathbf{E}\left[\bar{\mathbf{z}}\bar{\mathbf{z}}^\top\right] \tag{22}$$

Such $\mathbf{W}$ and $\mathbf{M}$ could potentially arise either over evolutionary time scales, or during the animal lifetime. In summary, based on the postulated objective function (18), we derived neural dynamics equation (equations (19) for LC, (20) for NNC) which map onto the ORN-LN circuit and biologically plausible Hebbian synaptic plasticity rules (equations (21)). This fully specifies the circuit, its synaptic weights, and its input-output relationship.

Numerical simulation of the LC offline

For the LC, we have the theoretical solution, so numerical simulations are not necessary to obtain $\mathbf{Y}$. Also, there is a degeneracy in the solutions of $\mathbf{Z}$, $\mathbf{W}$, and $\mathbf{M}$. However, to confirm the theoretical results, we did simulate the LC too. For that, we used the following equation, where the cost function depends on $\mathbf{Z}$ only (**Supplementary Information**, equation (S49), with $\gamma = 1$):

$$\mathcal{L} = \min_{\mathbf{Z}} \frac{1}{T^2} \operatorname{Tr}\left[\frac{T}{2}\mathbf{X}^\top\mathbf{X}\left(\mathbf{I}_T + \frac{1}{T}\mathbf{Z}^\top\mathbf{Z}\right)^{-1} + \frac{1}{4\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right] \tag{23}$$

We used an algorithm similar to Kuang et al., 2012.

Algorithm 1 Finding the minimum of
$f(\mathbf{Z}) = \operatorname{Tr}\left[\frac{T}{2}\mathbf{X}^\top\mathbf{X}\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{T} + \mathbf{I}_T\right)^{-1} + \frac{1}{4\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right]$

```
Objective: find $\mathbf{Z} \in \mathbb{R}^{K\times T}$ that minimizes $f(\mathbf{Z})$.
Inputs:
$\mathbf{X} \in \mathbb{R}^{D\times T}$
$K > 0$: the number of dimensions of $\mathbf{Z}$
$\rho > 0$: a constant encoding the strength of the inhibition by the LNs
$0 < \sigma < 1$: acceptance parameter (usually 0.1)
$\alpha_0 > 0$: initial gradient step coefficient (usually 1)
$0 < \beta < 1$: reduction factor (usually 0.1)
$0 < \mu \ll 1$: tolerance parameter (usually $\approx 10^{-6}$)
$n_{cycle} \approx 500$: number of steps after which one decreases the value of $\alpha_0$
Initialize:
$\mathbf{Z}_{new} \in \mathbb{R}^{K\times N} \sim \mathcal{N}(0, \mathrm{s.d.}(\mathbf{X})/100)$
$i \leftarrow 1$
Iterate:
repeat
    $\mathbf{Z} \leftarrow \mathbf{Z}_{new}$
    $\alpha = \alpha_0$
    repeat
        $\mathbf{Z}_{new} = \mathbf{Z} - \alpha\nabla f(\mathbf{Z})$    ▷ Find a potential new $\mathbf{Z}$ through a gradient descent step
        $\widehat{\Delta f} = \sigma \cdot \mathrm{sum}[\nabla f(\mathbf{Z}) \odot (\mathbf{Z}_{new} - \mathbf{Z})]$    ▷ Acceptable decrease in $f$ (negative number)
        $\Delta f = f(\mathbf{Z}_{new}) - f(\mathbf{Z})$    ▷ True decrease in $f$ (negative number)
        $\alpha \leftarrow \beta\alpha$    ▷ Decrease the gradient descent step size for the next iteration, if it occurs
    until $\Delta f < \widehat{\Delta f}$    ▷ Exit loop if the true decrease in $f$ is larger than the acceptable one
    if $i \mod n_{cycle} = 0$ then    ▷ Every $n_{cycle}$, decrease the initial step size $\alpha_0$ by $\beta$
        $\alpha_0 \leftarrow \beta\alpha_0$
    end if
    $i \leftarrow i + 1$
until $|f(\mathbf{Z}) - f(\mathbf{Z}_{new})|/|f(\mathbf{Z})| < \mu$
Output: $\mathbf{Z}_{new}$
```

Where $\odot$ is an element-wise multiplication and the "sum" adds all the elements of the matrix. In the inner repeat loop of the algorithm, it can happen that because of limited numerical precision, no α is small enough to make a decrease in f (i.e., satisfy the condition $\Delta f < \widehat{\Delta f}$), in that case the inner and outer repeat loops stop and the current $\mathbf{Z}$ (not $\mathbf{Z}_{new}$) is outputted.

$\nabla f(\mathbf{Z})$ is given by:

$$\mathbf{B} := \left(\mathbf{Z}^\top \mathbf{Z}/T + \mathbf{I}\right)^{-1} \tag{24}$$

$$\nabla f(\mathbf{Z}) = -\mathbf{Z}\mathbf{B}\mathbf{X}\mathbf{X}^\top \mathbf{B} + \mathbf{Z}\mathbf{Z}^\top \mathbf{Z}/\rho^2 \tag{25}$$

Finally, the expression for $\mathbf{Y}$ is (**Supplementary Information**, equation (S48)):

$$\mathbf{Y} = \mathbf{X}\left(\mathbf{I}_T + \frac{1}{T}\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \tag{26}$$

Numerical simulation of the NNC offline

For the NNC, we do not have the analytical expressions of $\mathbf{Y}$ and $\mathbf{Z}$. To minimize the objective function, we perform alternating gradient descent/ascent steps on $\mathbf{Y}$ and $\mathbf{Z}$, respectively. We start from the expanded expression of the objective function (18) (with $\gamma = 1$):

$$\mathcal{L} = \min_{\mathbf{Y}\geq 0} \max_{\mathbf{Z}\geq 0} \frac{1}{T^2} \operatorname{Tr}\left[-T\mathbf{X}^\top \mathbf{Y} + \frac{T}{2}\mathbf{Y}^\top \mathbf{Y} + \frac{1}{2}\mathbf{Y}^\top \mathbf{Y}\mathbf{Z}^\top \mathbf{Z} - \frac{1}{4\rho^2}\mathbf{Z}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{Z}\right] \tag{27}$$

Algorithm 2 Finding the minimum in $\mathbf{Y}$ and maximum in $\mathbf{Z}$ of $f(\mathbf{Y},\mathbf{Z}) = \mathrm{Tr}\left[-T\mathbf{X}^\top\mathbf{Y} + \frac{T}{2}\mathbf{Y}^\top\mathbf{Y} + \frac{1}{2}\mathbf{Y}^\top\mathbf{Y}\mathbf{Z}^\top\mathbf{Z} - \frac{1}{4\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right]$

```
**Objective**: find $\mathbf{Y} \in \mathbb{R}_+^{D\times T}$ and $\mathbf{Z} \in \mathbb{R}_+^{K\times T}$ that optimize $\min_{\mathbf{Y}} \max_{\mathbf{Z}} f(\mathbf{Y},\mathbf{Z})$.
**Inputs**:
$\mathbf{X} \in \mathbb{R}^{D\times T}$
$K > 0$: the number of dimensions of $\mathbf{Z}$
$\rho > 0$: a constant encoding the strength of the inhibition by the LNs
$0 < \sigma < 1$: acceptance parameter (usually 0.1)
$\alpha_0 > 0$: initial gradient step coefficient (usually 1)
$0 < \beta < 1$: reduction factor (usually 0.1)
$0 < \mu \ll 1$: tolerance parameter (usually $\approx 10^{-6}$)
$n_{cycle} \approx 500$: number of steps after which one decreases the value of $\alpha_0$
**Initialize**:
$\mathbf{Y}_{new} \in \mathbb{R}_+^{D\times N} \sim \mathrm{abs}[\mathcal{N}(0, \mathrm{s.d.}(\mathbf{X})/100)]$
$\mathbf{Z}_{new} \in \mathbb{R}_+^{K\times N} \sim \mathrm{abs}[\mathcal{N}(0, \mathrm{s.d.}(\mathbf{X})/100)]$
$i \leftarrow 1$
**Iterate**:
**repeat**
    $\mathbf{Y} \leftarrow \mathbf{Y}_{new}$
    $\mathbf{Z} \leftarrow \mathbf{Z}_{new}$
    $\alpha = \alpha_0$
    **repeat**
        $\mathbf{Y}_{new} = [\mathbf{Y} - \alpha\nabla_{\mathbf{Y}} f(\mathbf{Y},\mathbf{Z})]^+$ ▷ Find a potential new $\mathbf{Y}$ through a gradient descent step
        $\widehat{\Delta f} = \sigma \cdot \mathrm{sum}[\nabla_{\mathbf{Y}} f(\mathbf{Y},\mathbf{Z}) \odot (\mathbf{Y}_{new} - \mathbf{Y})]$ ▷ Acceptable decrease in $f$ (negative number)
        $\Delta f = f(\mathbf{Y}_{new},\mathbf{Z}) - f(\mathbf{Y},\mathbf{Z})$    ▷ True decrease in $f$ (negative number)
        $\alpha \leftarrow \beta\alpha$    ▷ Decrease the gradient descent step size for the next iteration, if it occurs
    **until** $\Delta f < \widehat{\Delta f}$    ▷ Exit loop if the true decrease in $f$ is larger than the acceptable one
    $\alpha = \alpha_0$
    **repeat**
        $\mathbf{Z}_{new} = [\mathbf{Z} + \alpha\nabla_{\mathbf{Z}} f(\mathbf{Y}_{new},\mathbf{Z})]^+$ ▷ find a potential new $\mathbf{Z}$ through a gradient ascend step
        $\widehat{\Delta f} = \sigma \cdot \mathrm{sum}[\nabla_{\mathbf{Z}} f(\mathbf{Y}_{new},\mathbf{Z}) \odot (\mathbf{Z}_{new} - \mathbf{Z})]$ ▷ Acceptable increase in $f$ (positive number)
        $\Delta f = f(\mathbf{Y}_{new},\mathbf{Z}_{new}) - f(\mathbf{Y}_{new},\mathbf{Z})$    ▷ True increase in $f$ (positive number)
        $\alpha \leftarrow \beta\alpha$    ▷ Decrease the ascent descent step size for the next iteration, if it occurs
    **until** $\Delta f > \widehat{\Delta f}$    ▷ Exit loop if the true increase in $f$ is larger than the acceptable one
    **if** $i \mod n_{cycle} = 0$ **then**    ▷ Every $n_{cycle}$, decrease the initial step size $\alpha_0$ by $\beta$
        $\alpha_0 \leftarrow \beta\alpha_0$
    **end if**
    $i \leftarrow i + 1$
**until** $|f(\mathbf{Y},\mathbf{Z}) - f(\mathbf{Y}_{new},\mathbf{Z})|/|f(\mathbf{Y},\mathbf{Z})| < \mu$ and $|f(\mathbf{Y}_{new},\mathbf{Z}) - f(\mathbf{Y}_{new},\mathbf{Z}_{new})|/|f(\mathbf{Y}_{new},\mathbf{Z})| < \mu$
**Output**: $\mathbf{Y}_{new}, \mathbf{Z}_{new}$
```

In the case of the LC, the same algorithm holds, with all the rectifications $[.]^+$ removed from the algorithm and the "abs" removed from the initiation. If in either of the inner repeat loops, no α is small enough to make a decrease/increase in f (i.e., satisfy the condition $\Delta f < \widehat{\Delta f}$ or $\Delta f > \widehat{\Delta f}$), the iterations stop and the current $\mathbf{Y}$ and $\mathbf{Z}$ are the output of the algorithm.

The gradients of $f(\mathbf{Y}, \mathbf{Z})$ are:

$$\nabla_{\mathbf{Y}} f(\mathbf{Y}, \mathbf{Z}) = -T(\mathbf{X} - \mathbf{Y}) + \mathbf{Y}\mathbf{Z}^\top\mathbf{Z} \tag{28}$$

$$\nabla_{\mathbf{Z}} f(\mathbf{Y}, \mathbf{Z}) = \mathbf{Z}\mathbf{Y}^\top\mathbf{Y} - \mathbf{Z}\mathbf{Z}^\top\mathbf{Z}/\rho^2 \tag{29}$$

Numerical simulation of the circuits online

For **Fig. S11**, we simulated the circuit dynamics for a given $\mathbf{W}$, $\mathbf{M}$, and $\mathbf{X}$. For that purpose, to find $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$, we performed gradient descent steps based on the discretized equations (19) for the LC or equation (20) for the NNC.

Data and code availability

All data in this study is published in Berck et al., 2016; Si et al., 2019 and is accessible online: https://github.com/samuellab/Larval-ORN, https://doi.org/10.7554/eLife.14859.019, https://doi.org/10.7554/eLife.14859.020.

All the code used in this study is available here: https://github.com/chapochn/ORN-LN_circuit

References

Aimon, S., Katsuki, T., Jia, T., Grosenick, L., Broxton, M., Deisseroth, K., Sejnowski, T. J., & Greenspan, R. J. (2019). Fast near-whole–brain imaging in adult drosophila during responses to stimuli and behavior. *PLOS Biology*, *17*(2), e2006732.

Arenas, A., Giurfa, M., Sandoz, J. C., Hourcade, B., Devaud, J. M., & Farina, W. M. (2012). Early olfactory experience induces structural changes in the primary olfactory center of an insect brain. *European Journal of Neuroscience*, *35*(5), 682–690.

Asahina, K., Louis, M., Piccinotti, S., & Vosshall, L. B. (2009). A circuit supporting concentration-invariant odor perception in drosophila. *Journal of Biology*, *8*(1), 9.

Atick, J. J., & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation*, *2*(3), 308–320.

Atick, J. J., & Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation*, *4*(2), 196–210.

Atick, J. J., & Redlich, A. N. (1993). Convergent algorithm for sensory receptive field development. *Neural Computation*, *5*(1), 45–60.

Bahroun, Y., Chklovskii, D., & Sengupta, A. (2019). A similarity-preserving network trained on transformed images recapitulates salient features of the fly motion detection circuit. *Advances in Neural Information Processing Systems*, *32*.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–34). The MIT Press.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Berck, M. E., Khandelwal, A., Claus, L., Hernandez-Nunez, L., Si, G., Tabone, C. J., Li, F., Truman, J. W., Fetter, R. D., Louis, M., Samuel, A. D., & Cardona, A. (2016). The wiring diagram of a glomerular olfactory system. *eLife*, *5*, e14859.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 1–12.

Chou, Y.-H., Spletter, M. L., Yaksi, E., Leong, J. C. S., Wilson, R. I., & Luo, L. (2010). Diversity and wiring variability of olfactory local interneurons in the *Drosophila* antennal lobe. *Nature Neuroscience*, *13*(4), 439–449.

Das, S., Sadanandappa, M. K., Dervan, A., Larkin, A., Lee, J. A., Sudhakaran, I. P., Priya, R., Heidari, R., Holohan, E. E., Pimentel, A., Gandhi, A., Ito, K., Sanyal, S., Wang, J. W., Rodrigues, V., & Ramaswami, M. (2011). Plasticity of local GABAergic interneurons drives olfactory habituation. *Proceedings of the National Academy of Sciences*, *108*(36), E646–E654.

Devaud, J.-M., Acebes, A., & Ferrús, A. (2001). Odor exposure causes central adaptation and morphological changes in selected olfactory glomeruli in drosophila. *Journal of Neuroscience*, *21*(16), 6274–6282.

Eichler, K., Li, F., Litwin-Kumar, A., Park, Y., Andrade, I., Schneider-Mizell, C. M., Saumweber, T., Huser, A., Eschbach, C., Gerber, B., Fetter, R. D., Truman, J. W., Priebe, C. E., Abbott, L. F., Thum, A. S., Zlatic, M., & Cardona, A. (2017). The complete connectome of a learning and memory centre in an insect brain. *Nature*, *548*(7666), 175–182.

Friedrich, R. W. (2013). Neuronal computations in the olfactory system of zebrafish. *Annual Review of Neuroscience*, *36*(1), 383–402.

Friedrich, R. W., & Laurent, G. (2001). Dynamic optimization of odor representations by slow temporal patterning of mitral cell activity. *Science*, *291*(5505), 889–894.

Friedrich, R. W., & Wiechert, M. T. (2014). Neuronal circuits and computations: Pattern decorrelation in the olfactory bulb. *FEBS Letters*, *588*(15), 2504–2513.

Giridhar, S., Doiron, B., & Urban, N. N. (2011). Timescale-dependent shaping of correlation by olfactory bulb lateral inhibition. *Proceedings of the National Academy of Sciences*, *108*(14), 5843–5848.

Golkar, S., Lipshutz, D., Bahroun, Y., Sengupta, A., & Chklovskii, D. (2020). A simple normative network approximates local non-hebbian learning in the cortex. *Advances in Neural Information Processing Systems*, *33*, 7283–7295.

Gschwend, O., Abraham, N. M., Lagier, S., Begnaud, F., Rodriguez, I., & Carleton, A. (2015). Neuronal pattern separation in the olfactory bulb improves odor discrimination learning. *Nature Neuroscience*, *18*(10), 1474–1482.

Hattori, R., Kuchibhotla, K. V., Froemke, R. C., & Komiyama, T. (2017). Functions and dysfunctions of neocortical inhibitory neuron subtypes. *Nature Neuroscience*, *20*(9), 1199–1208.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197.

Holderith, N., Lorincz, A., Katona, G., Rózsa, B., Kulik, A., Watanabe, M., & Nusser, Z. (2012). Release probability of hippocampal glutamatergic terminals scales with the size of the active zone. *Nature Neuroscience*, *15*(7), 988–997.

Hong, E. J., & Wilson, R. I. (2015). Simultaneous encoding of odors by channels with diverse sensitivity to inhibition. *Neuron*, *85*(3), 573–589.

Kessy, A., Lewin, A., & Strimmer, K. (2018). Optimal whitening and decorrelation. *The American Statistician*, *72*(4), 309–314.

Kim, A. J., Lazar, A. A., & Slutskiy, Y. B. (2015). Projection neurons in drosophila antennal lobes signal the acceleration of odor concentrations. *eLife*, *4*, e06651.

King, P. D., Zylberberg, J., & DeWeese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1. *Journal of Neuroscience*, *33*(13), 5475–5485.

Koulakov, A. A., & Rinberg, D. (2011). Sparse incomplete representations: A potential role of olfactory granule cells. *Neuron*, *72*(1), 124–136.

Kuang, D., Park, H., & Ding, C. (2012). Symmetric nonnegative matrix factorization for graph clustering. *International Conference on Data Mining*, 494–505.

Laurent, G. (2002). Olfactory network dynamics and the coding of multidimensional signals. *Nature Reviews Neuroscience*, *3*(11), 884–895.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*(3), 105–117.

Linster, C., Sachse, S., & Galizia, C. G. (2005). Computational modeling suggests that response properties rather than spatial position determine connectivity between olfactory glomeruli. *Journal of Neurophysiology*, *93*(6), 3410–3417.

Masland, R. H. (2012). The neuronal organization of the retina. *Neuron*, *76*(2), 266–280.

Nagel, K. I., Hong, E. J., & Wilson, R. I. (2014). Synaptic and circuit mechanisms promoting broadband transmission of olfactory stimulus dynamics. *Nature Neuroscience*, *18*(1), nn.3895.

Nagel, K. I., & Wilson, R. I. (2016). Mechanisms underlying population response dynamics in inhibitory interneurons of the drosophila antennal lobe. *Journal of Neuroscience*, *36*(15), 4325–4338.

Olsen, S. R., Bhandawat, V., & Wilson, R. I. (2010). Divisive normalization in olfactory population codes. *Neuron*, *66*(2), 287–299.

Olsen, S. R., & Wilson, R. I. (2008). Lateral presynaptic inhibition mediates gain control in an olfactory circuit. *Nature*, *452*(7190), 956–960.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, *37*(23), 3311–3325.

Pehlevan, C., & Chklovskii, D. B. (2015). A hebbian/anti-hebbian network derived from online nonnegative matrix factorization can cluster and discover sparse features. *Conference Record - Asilomar Conference on Signals, Systems and Computers*, *2015-April*, 769–775.

Pehlevan, C., & Chklovskii, D. B. (2016). Optimization theory of hebbian/anti-hebbian networks for PCA and whitening. *2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015*, 1458–1465.

Pehlevan, C., Sengupta, A., & Chklovskii, D. B. (2018). Why do similarity matching objectives lead to hebbian/anti-hebbian networks? *Neural computation*, *30*(1), 84–124.

Plumbley, M. D. (1993). A hebbian/anti-hebbian network which optimizes information capacity by orthonormalizing the principal subspace. *in Proc. IEE Conf. on Artificial Neural Networks*, 86–90.

Sachse, S., Rueckert, E., Keller, A., Okada, R., Tanaka, N. K., Ito, K., & Vosshall, L. B. (2007). Activity-dependent plasticity in an olfactory circuit. *Neuron*, *56*(5), 838–850.

Sato, T. K., Haider, B., Häusser, M., & Carandini, M. (2016). An excitatory basis for divisive normalization in visual cortex. *Nature Neuroscience*, *19*(4), 568–570.

Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S.-y., Hayworth, K. J., Huang, G. B., Shinomiya, K., Maitlin-Shepard, J., Berg, S., Clements, J., Hubbard, P. M., Katz, W. T., Umayam, L., Zhao, T., Ackerman, D., Blakely, T., Bogovic, J., Dolafi, T., … Plaza, S. M. (2020). A connectome and analysis of the adult drosophila central brain (E. Marder, M. B. Eisen, J. Pipkin, & C. Q. Doe, Eds.). *eLife*, *9*, e57443.

Si, G., Kanwal, J. K., Hu, Y., Tabone, C. J., Baron, J., Berck, M., Vignoud, G., & Samuel, A. D. T. (2019). Structured odorant response patterns across a complete olfactory receptor neuron population. *Neuron*, *101*(5), 950–962.e7.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.

Sudhakaran, I. P., Holohan, E. E., Osman, S., Rodrigues, V., VijayRaghavan, K., & Ramaswami, M. (2012). Plasticity of recurrent inhibition in the drosophila antennal lobe. *Journal of Neuroscience*, *32*(21), 7225–7231.

Takemura, S.-y., Bharioke, A., Lu, Z., Nern, A., Vitaladevuni, S. N., Rivlin, P. K., Katz, W. T., Olbris, D. J., Plaza, S. M., Winston, P., Zhao, T., Horne, J. A., Fetter, R. D., Takemura, S., Blazek, K., Chang, L.-A., Ogundeyi, O., Saunders, M. a., Shapiro, V., … Chklovskii, D. B. (2013). A visual motion detection circuit suggested by drosophila connectomics. *Nature*, *500*(7461), 175–181.

Tobin, W. F., Wilson, R. I., & Lee, W.-C. A. (2017). Wiring variations that enable and constrain neural computation in a sensory microcircuit (L. Luo, Ed.). *eLife*, *6*, e24838.

Wanner, A. A., & Friedrich, R. W. (2020). Whitening of odor representations by the wiring diagram of the olfactory bulb. *Nature Neuroscience*, 1–10.

Westrick, Z. M., Heeger, D. J., & Landy, M. S. (2016). Pattern adaptation and normalization reweighting. *Journal of Neuroscience*, *36*(38), 9805–9816.

Wick, S. D., Wiechert, M. T., Friedrich, R. W., & Riecke, H. (2010). Pattern orthogonalization via channel decorrelation by adaptive networks. *Journal of Computational Neuroscience*, *28*(1), 29–45.

Wilson, R. I. (2013). Early olfactory processing in drosophila: Mechanisms and principles. *Annual Review of Neuroscience*, *36*(1), 217–241.

Zhu, M., & Rozell, C. J. (2015). Modeling inhibitory interneurons in efficient sensory coding models. *PLOS Computational Biology*, *11*(7), e1004353.

Zhu, P., Frank, T., & Friedrich, R. W. (2013). Equalization of odor representations by a network of electrically coupled inhibitory interneurons. *Nature Neuroscience*, *16*(11), 1678–1686.

Acknowledgments

We thank Aravinthan D.T. Samuel, Jacob Baron, Guangwei Si, Thomas Frank, Victor Minden, Anirvan Sengupta, Eftychios A. Pnevmatikakis, Shiva GhaaniFarashahi, and the Neuroscience Group at the Flatiron Institute for discussions and/or comments on the manuscript.

Author contributions

All authors designed the study. C.P. and D.B.C. formulated the objective function. N.M.C. and C.P. performed theoretical derivations. N.M.C. wrote the computer code, analyzed the data, performed numerical simulations, and prepared the original draft. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Figures

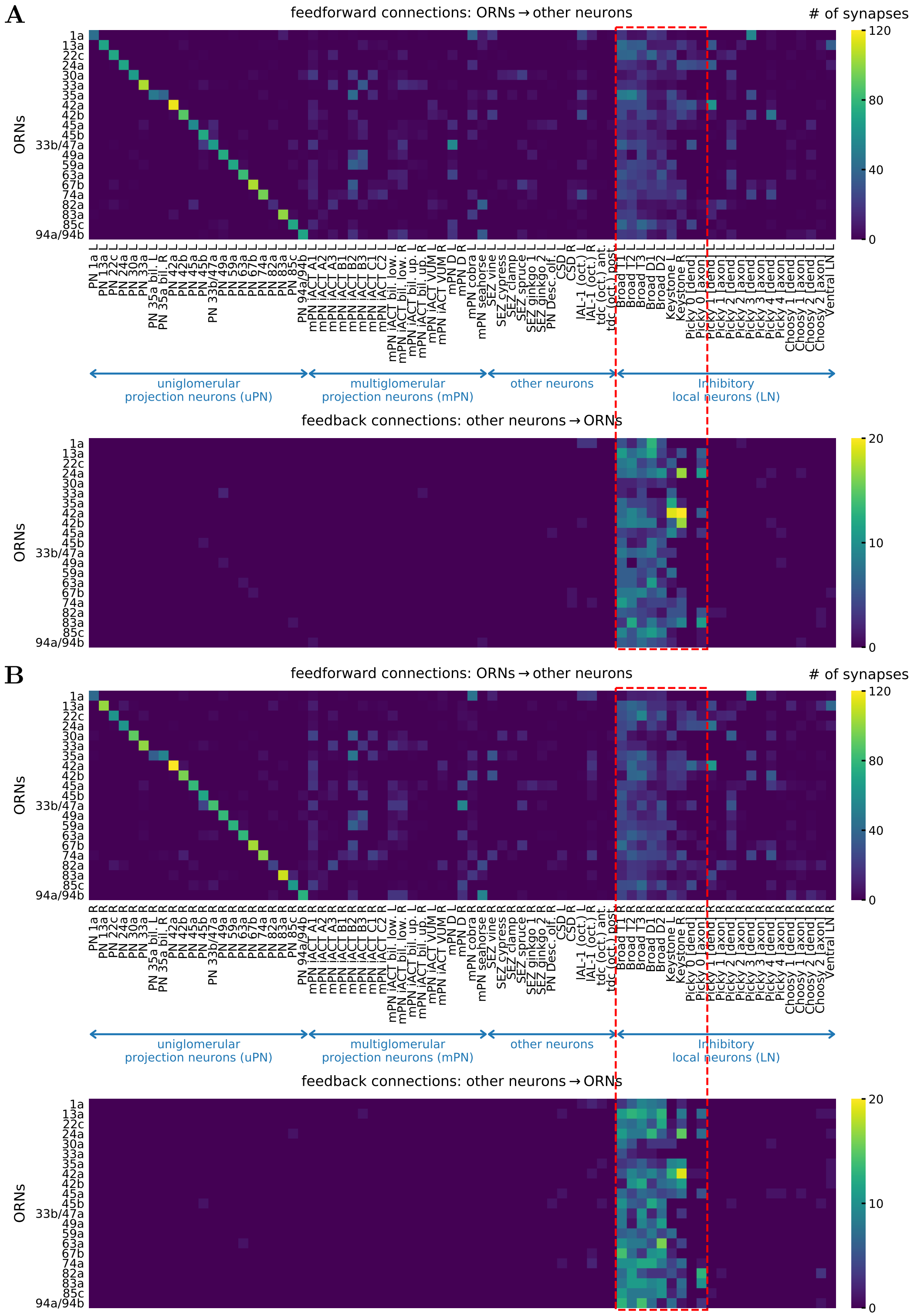

Fig. S1. Full ORN connectivity and circuit selection

A Heat map of the ORNs ↔ LN feedforward and feedback connections on the left side of the *Drosophila* larva. We focus on the neurons, that synapse bidirectionally with ORNs (inside the red dashed rectangle): Broad Trios, Broad Duets, Keystones, and Picky 0. These neurons are all LNs.

B Same as (**A**) for the right side.

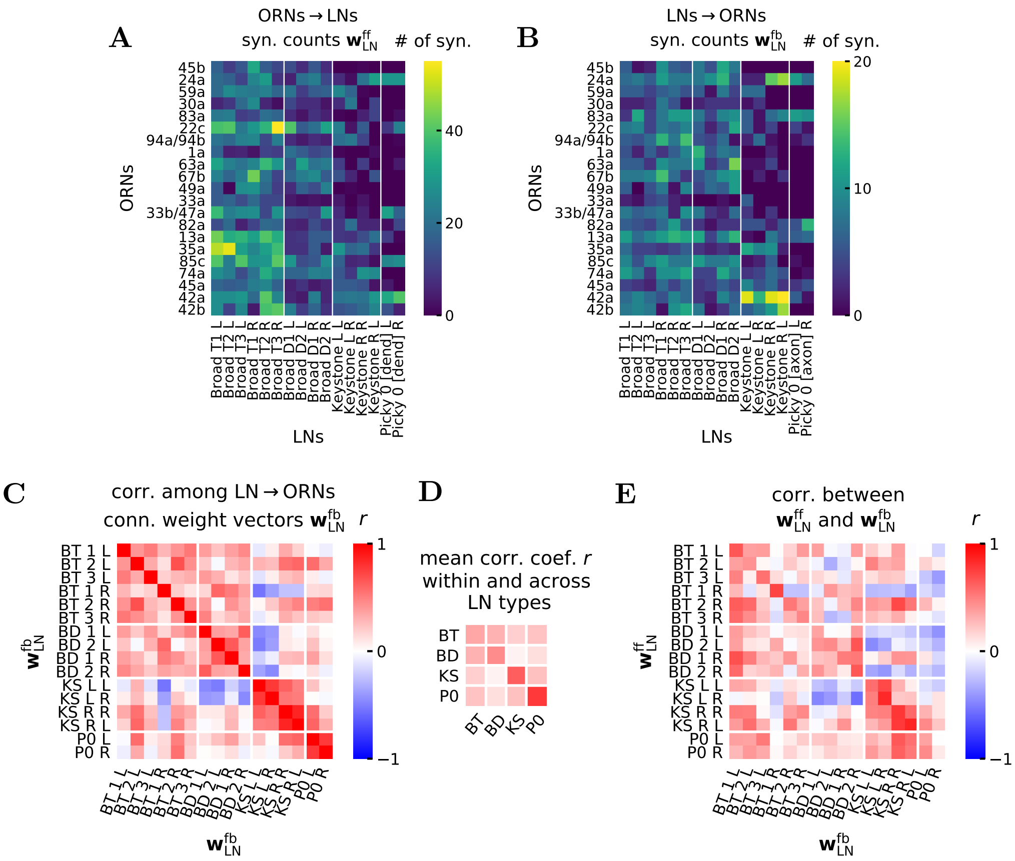

Fig. S2. ORN-LN connectivity, comparison feedforward with feedback

A ORNs → LNs feedforward connections weights $\mathbf{w}^{\mathrm{ff}}_{\mathrm{LN}}$ on both left and right sides of the antennal lobe with the chosen LNs, ordered by LN class. The vectors $\mathbf{w}^{\mathrm{ff}}_{\mathrm{LN}}$ correspond to the columns of the depicted matrix.

B LN → ORNs feedback connections weights $\mathbf{w}^{\mathrm{fb}}_{\mathrm{LN}}$ on both left and right sides of the antennal lobe with the chosen LNs, ordered by LN class. The vectors $\mathbf{w}^{\mathrm{fb}}_{\mathrm{LN}}$ correspond to the columns of the depicted matrix.

C Correlation coefficients between feedback LN → ORNs connection weight vectors $\mathbf{w}^{\mathrm{fb}}_{\mathrm{LN}}$.

D Average rectified correlation coefficient $\langle r_+ \rangle$ $(r_+ := \max[0, r])$ between LN types calculated by averaging the rectified values from (**C**) in each rectangle with white border, excluding the diagonal entries of the full matrix. The average correlation coefficient within a class is larger than the correlation coefficient across classes.

E Correlation coefficients between feedforward ORNs → LN $\mathbf{w}^{\mathrm{ff}}_{\mathrm{LN}}$ and feedback LN → ORNs $\mathbf{w}^{\mathrm{fb}}_{\mathrm{LN}}$ connection weight vectors. The Picky 0 LN is the only LN that has a separation between axonal and dendritic terminals. For the feedforward ORNs → LN connections, we only include in the connection weight vector the synapses onto the Picky 0 dendrite, and for the LN → ORNs connection, we only count the synapses from the Picky 0 axon.

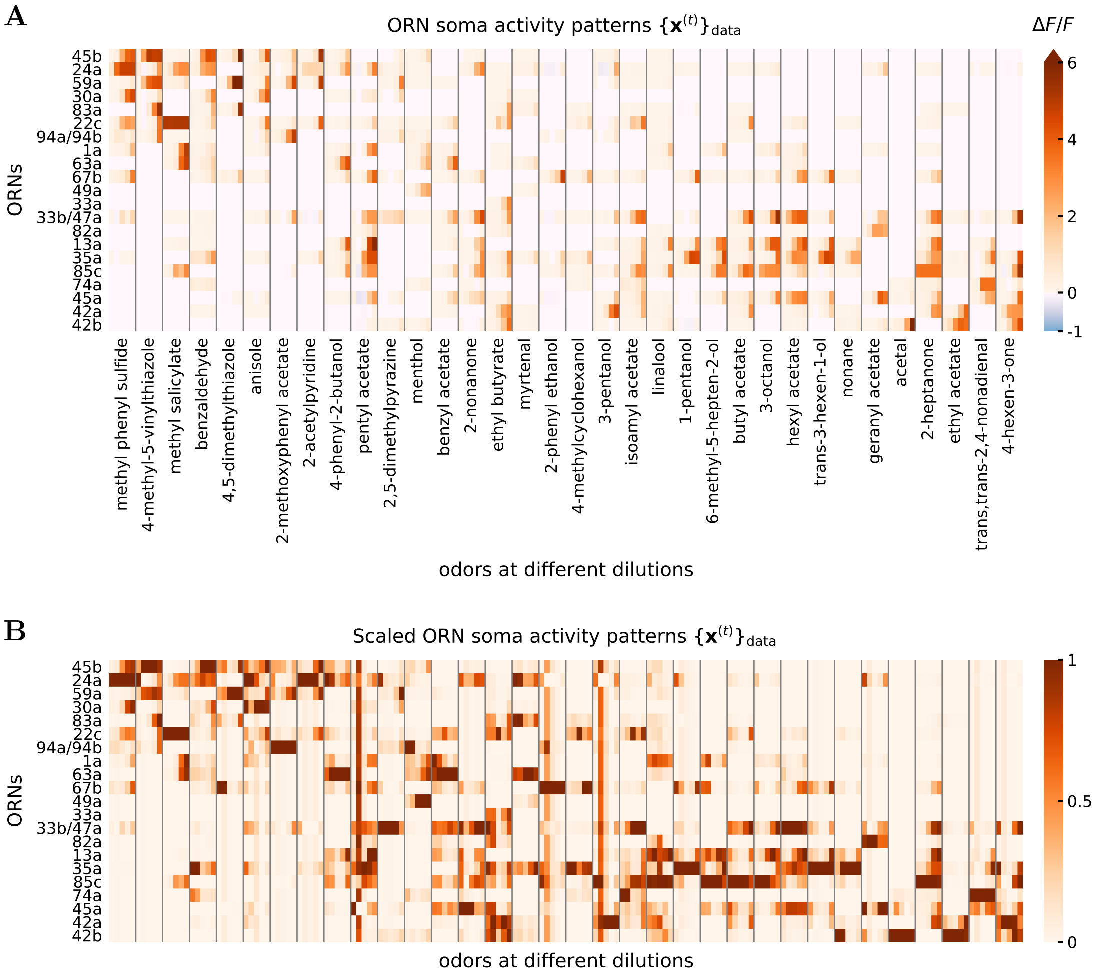

Fig. S3. ORN soma activity from Si et al., 2019

A ORN soma activity patterns $\left\{\mathbf{x}^{(t)}\right\}_{\text{data}}$ in response to 34 odors at 5 dilutions acquired through Ca^{2+} imaging. Different odors are separated by vertical gray lines. For each odor, there are 5 columns corresponding to 5 dilutions: $10^{-8}, ..., 10^{-4}$. The odors and ORNs are ordered by the value of the second singular vectors of the left and right SVD matrices of this activity data, after centering and normalizing. This data is obtained by averaging the maximum responses of several trials to the same odor and dilution (as in Si et al., 2019).

B Same as (**A**), with each $\mathbf{x}^{(t)}$ scaled between 0 and 1 to better portray the patterns.

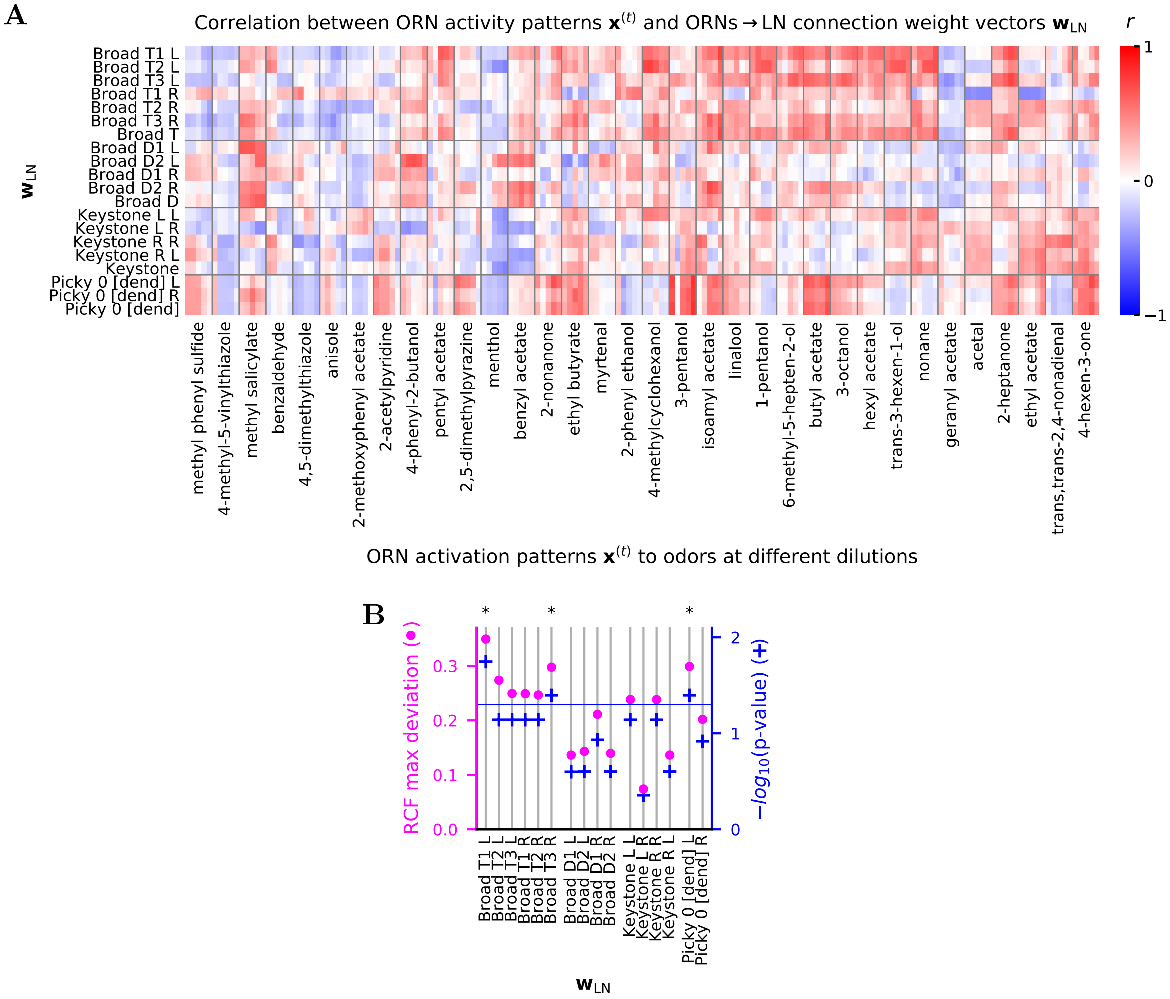

Fig. S4. Alignment of activity patterns $\mathrm{x}^{(t)}$ in ORNs and ORNs → LN connectivity weight vectors w_{LN}
A Same as **Fig. 2E**, for all the w_{LN} and with all the odors labeled. Same odor order.
B Same as **Fig. 2I**, for all w_{LN}.

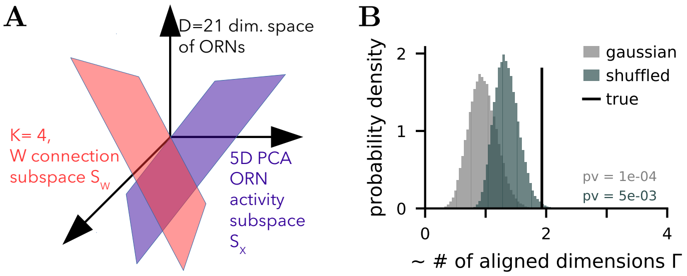

Fig. S5. Activity and connectivity subspace alignment

A Scheme representing the comparison of the 4-dimensional connectivity (S_W) and 5-dimensional activity (S_X) subspaces in 21 dimensions $(D = 21$, dimensionality of the ORN space).

B Number of aligned dimensions Γ between the 2 subspaces of (**A**) in the data (true, $\Gamma = 1.9$), from randomly shuffling the connectivity vector entries (shuffled, mean $\Gamma = 1.3$) and from random normal vectors (Gaussian, mean $\Gamma = 1$). pv: one-sided p-value.

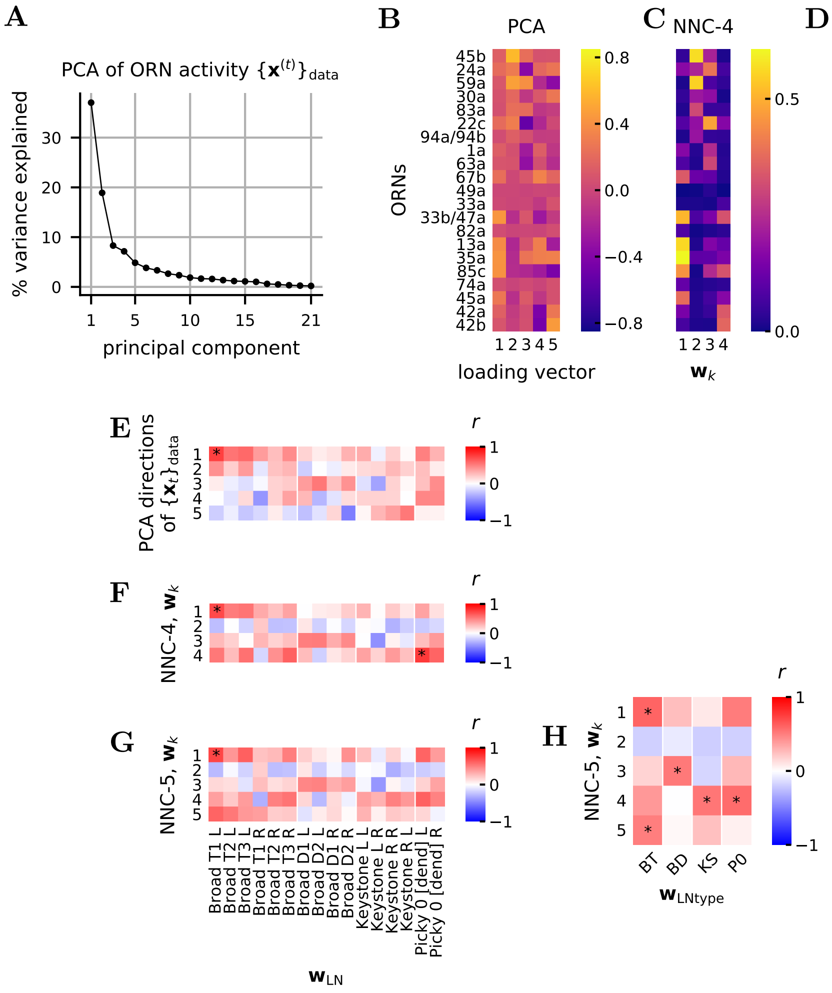

Fig. S6. Activity and connectivity

A Percentage of the variance of the ORN activity patters $\left\{\mathbf{x}^{(t)}\right\}_{\text{data}}$ explained by the uncentered PCA. The top 4 and 5 PCA directions explain 71% and 76% of the variance, respectively.

B First 5 PCA loading vectors of $\left\{\mathbf{x}^{(t)}\right\}_{\text{data}}$.

C-D $\mathbf{w}_k$ from NNC with $K = 4, 5$ and $\rho = 1$, ordered to resemble the PCA ordering.

E Same as **Fig. 3C** with all $\mathbf{w}_{\text{LN}}$.

F Same as (**E**), with $\mathbf{w}_k$ from NNC-4 instead of PCA loading vectors.

G Same as (**F**), for NNC-5. The small number of significant points in (**E**-**G**) results from the higher number of hypothesis tests, which decreases the adjusted p-values in the FDR multi-hypothesis testing framework.

H Same as **Fig. 4A**, for NNC-5.

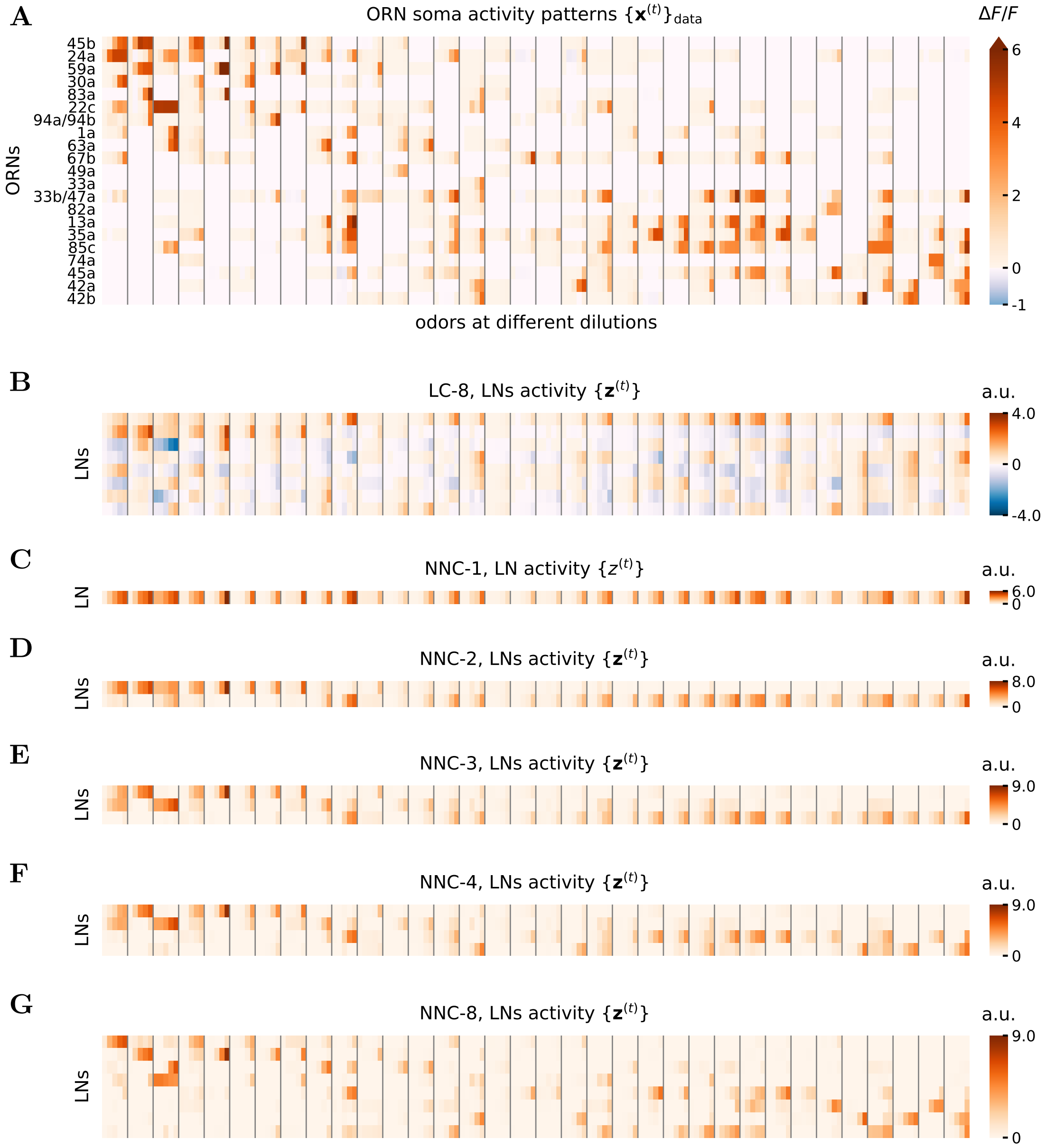

Fig. S7. Activity of LNs $\{\mathbf{z}^{(t)}\}$ in the NNC and LC

A ORN soma activity patterns $\{\mathbf{x}^{(t)}\}_{\text{data}}$ as in **Fig. S3A**.

B Activity in the LNs $\{\mathbf{z}^{(t)}\}$ for the LC-8. Stimuli are aligned to the panel above. As mentioned in the text, $\{\mathbf{z}^{(t)}\}$ is undetermined up to an orthogonal matrix $\mathbf{U}_Z$. Here we set $\mathbf{U}_Z = \mathbf{I}_K$, i.e., identity matrix. For LC-K, the response in LNs correspond to the first K row of this matrix, multiplied by any $K \times K$ orthogonal matrix on the left. Thus, the matrix depicted in this plot shows the potential activity in LNs for any LC-K with $K \leq 8$.

C $\{z_t\}$ for the NNC-1. The activity of the LN approximately follows the total activity.

D $\{\mathbf{z}^{(t)}\}$ for the NNC-2. One can see that the 2 LNs roughly clusters the sets of odors into those activating the top ORNs and those activating the lower ORNs.

E-**G** $\{\mathbf{z}^{(t)}\}$ for the NNC with $K = 3, 4, 8$. One observes a more sophisticated clustering of the data. As more LNs are added, LN activity increases in sparsity. The activity in the LNs for the NNC is more sparse than for the LC.

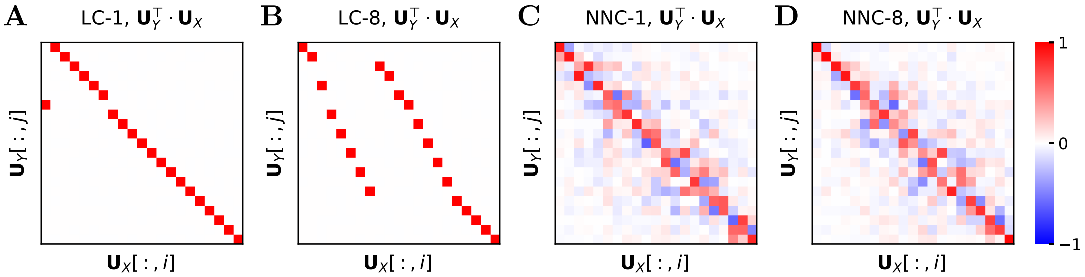

Fig. S8. Input vs output principal directions in LC and NNC

A-D Scalar product between principal directions of uncentered $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and $\{\mathbf{y}^{(t)}\}$ for the LC and NNC for $K = 1, 8$. For the LC the identity of the principal directions in conserved, only their order change. For the NNC, the principal directions are slightly mixed, but conserve the approximate ordering.

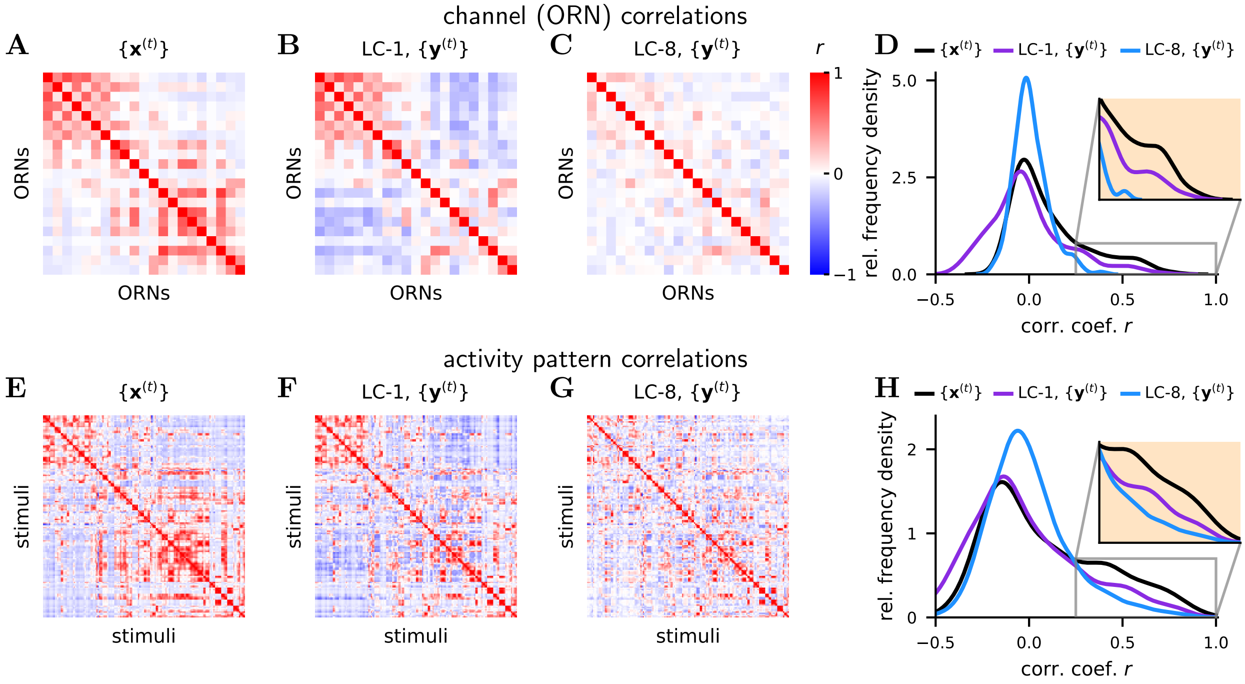

Fig. S9. Decorrelation in the LC

A-H Same as **Fig. 7J-O** for the LC.

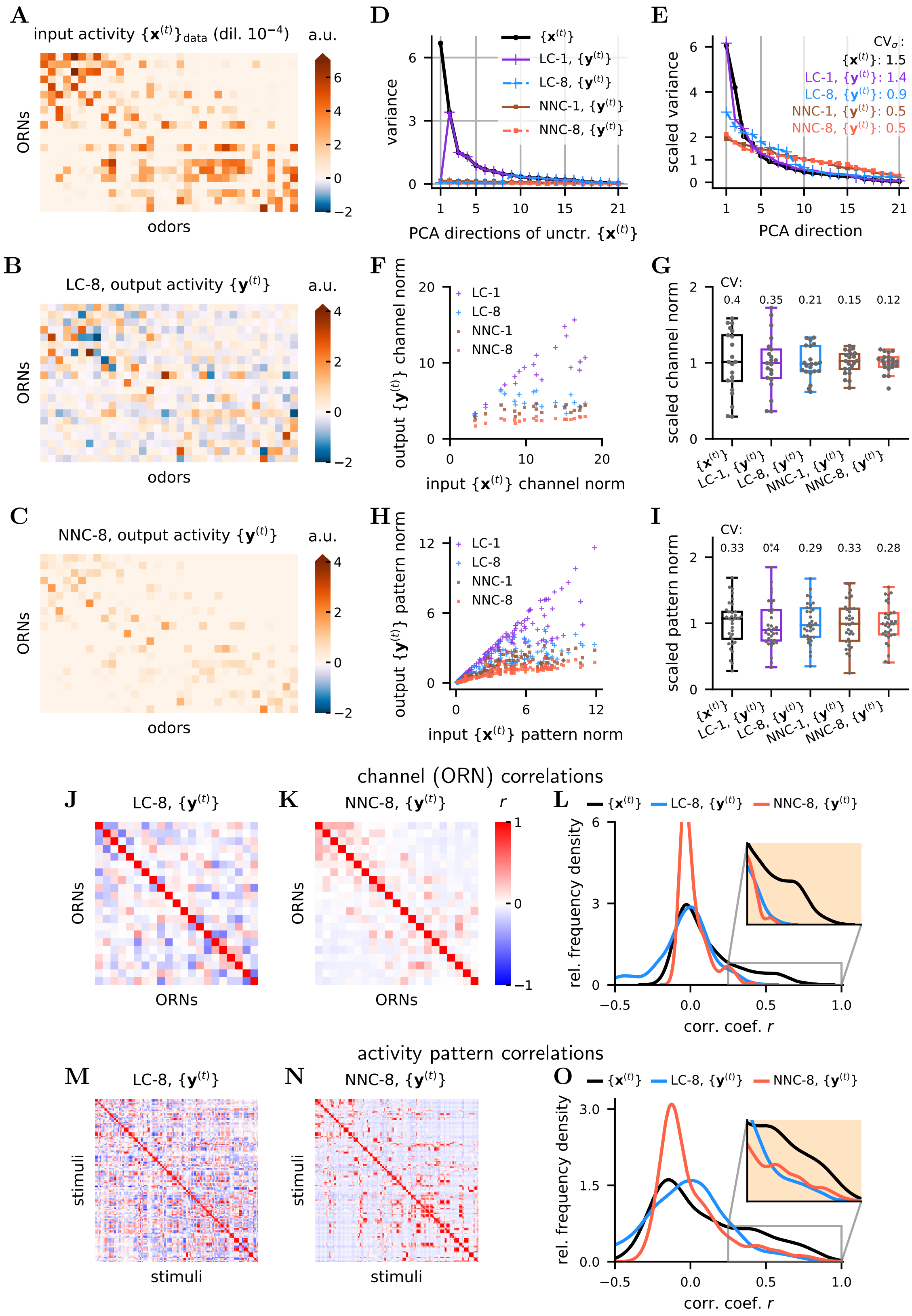

Fig. S10. Input transformation by LC and NNC with $\rho = 10$

Same as **Fig. 7** for $\rho = 10$. Note the even stronger dampening, flattening, and decorrelation.

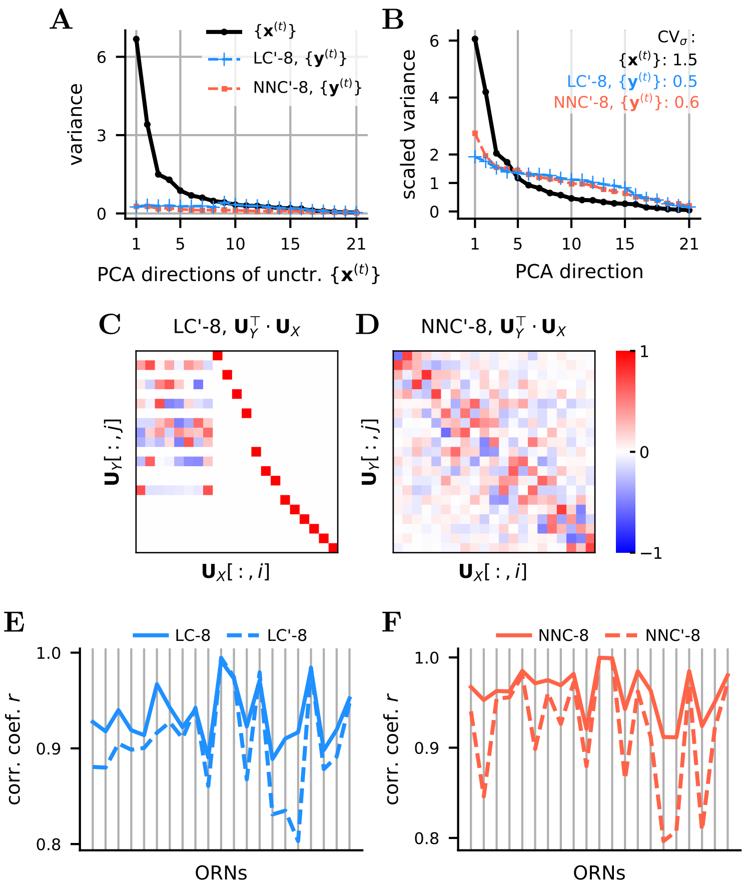

Fig. S11. Effect of removing off-diagonal entries in M for LC and NNC

A-**B** Same as **Fig. 7D,E** for the trained LC and NNC on $\{\mathbf{x}^{(t)}\}_{\text{data}}$, where the off-diagonal values of M are set to 0 (LC' and NNC'). Note that the values in LC' in (**A**) do not monotonically decrease as in LC.

C-**D** Same as **Fig. S8** for LC'-8 and NNC'-8. Note the increased mixture between the principal directions of $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and $\{\mathbf{y}^{(t)}\}$.

E Correlation between the input $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and output $\{\mathbf{y}^{(t)}\}$ for each channel (i.e., ORN) for LC-8 and LC'-8. Note that in the LC-8, the output of each channel is more strongly correlated to its own input for the LC-8 than for the LC'-8.

F Same as (**E**) for NNC-8 and NNC'-8.

Supplementary Information

In this supplement, we prove statements made in the results and methods sections:

Section 1: we describe the objective function from equation (18), show the equivalence of scaling $\mathbf{X}$ and ρ (section 1.1) and show the resemblance of this circuit's objective function with a whitening objective function (section 1.2).

Section 2: we show that the objective function (18), when optimized online with or without non-negativity constraints, lead to the circuit dynamics (19) or (20), respectively, and to Hebbian learning rules (21). We then show the steady state solution to which the circuit dynamics equations (19) converge and show that the steady state is stable (section 2.2).

Section 3: we show that a general system of differential equations describing the circuit contains two effective parameters and can be reduced to the form found in the main text in equation (19).

Section 4: we analyze computation in LC and prove equations (4), (5), (6a), and (6b) in the main text from the main text. These results are proven in two ways (sections 4.1 and 4.2). Section 4.3 discusses limiting cases of the computation for small and large values of ρ, and show the relation of NNC to SNMF (symmetric non-negative matrix factorization).

Section 5: we prove the relationship between $\mathbf{W}$ and $\mathbf{M}$, equation (2) in the main text.

Section 6: we prove the relationship between $\mathbf{W}$ and $\mathbf{X}$, equation (1) in the main text.

Section 7: we prove that the CV of singular values in $\mathbf{Y}$ is smaller than in $\mathbf{X}$ for the LC when $K = D$.

1 Objective function

We postulate the following minimax objective function:

$$\mathcal{L} = \min_{\mathbf{Y}} \max_{\mathbf{Z}} \frac{1}{T^2}\left(\frac{T}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 - \frac{\rho^2}{4u^2}\left\|\mathbf{Y}^\top\mathbf{Y} - \frac{\gamma^2}{\rho^2}\mathbf{Z}^\top\mathbf{Z}\right\|_F^2 + \frac{\rho^2}{4u^2}\left\|\mathbf{Y}^\top\mathbf{Y}\right\|_F^2\right) \tag{S1}$$

Which can be expanded thus:

$$\mathcal{L} = \min_{\mathbf{Y}} \max_{\mathbf{Z}} \frac{1}{T^2}\operatorname{Tr}\left[-T\mathbf{X}^\top\mathbf{Y} + \frac{T}{2}\mathbf{Y}^\top\mathbf{Y} + \frac{\gamma^2}{2u^2}\mathbf{Y}^\top\mathbf{Y}\mathbf{Z}^\top\mathbf{Z} - \frac{\gamma^4}{4u^2\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right] \tag{S2}$$

Where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{D\times T}$, $\mathbf{Z} \in \mathbb{R}^{K\times T}$ with D the number of ORNs (21 for this olfactory circuit), K the number of LNs, T the number of data (sample) points, ρ a positive unitless constant, u a unit with the physical dimension as $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ (e.g., spikes $\cdot\, s^{-1}$) (dropped for simplicity in the main text). $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ represent the activity of ORN somas, ORN axons, and LNs, respectively. We can interpret $\mathbf{X}$ as all the discretized activity of ORNs up to a certain point in their lifetime.

Optimizing objective function (S2) leads to the linear circuit (LC) model. Adding the non-negativity constraints on $\mathbf{Y}$ and $\mathbf{Z}$ leads to the non-negative circuit (NNC) model.

1.1 Equivalence of scaling $\mathbf{X}$ and ρ

Here, we show that scaling $\mathbf{X}$ is equivalent to scaling ρ in terms of circuit computation. It is easy to see that the transformation $\mathbf{X} \to a\mathbf{X}$, $\mathbf{Y} \to a\mathbf{Y}$ and $\rho \to \rho/a$ (for $a \neq 0$) leaves the objective function unaffected, i.e., this transformation is a symmetry of the optimization. Indeed:

$$\mathcal{L} = \min_{\mathbf{Y}} \max_{\mathbf{Z}} \frac{1}{T^2} \operatorname{Tr}\left[-T\mathbf{X}^\top\mathbf{Y} + \frac{T}{2}\mathbf{Y}^\top\mathbf{Y} + \frac{\gamma^2}{2u^2}\mathbf{Y}^\top\mathbf{Y}\mathbf{Z}^\top\mathbf{Z} - \frac{\gamma^4}{4u^2\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right] \tag{S3}$$

$$= \min_{\mathbf{Y}} \max_{\mathbf{Z}} \frac{1}{T^2} \operatorname{Tr}\left[-Ta^2\mathbf{X}^\top\mathbf{Y} + \frac{T}{2}a^2\mathbf{Y}^\top\mathbf{Y} + \frac{a^2\gamma^2}{2u^2}\mathbf{Y}^\top\mathbf{Y}\mathbf{Z}^\top\mathbf{Z} - \frac{a^2\gamma^4}{4u^2\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right] \tag{S4}$$

Let us explore the consequence of this symmetry. The output $\mathbf{Y}$ of the optimization is a function of $\mathbf{X}$ and ρ, thus we can define a function f such that: $\mathbf{Y} = f(\mathbf{X}, \rho)$:

$$\mathbf{Y} = f(\mathbf{X}, \rho) = \arg\min_{\mathbf{Y}} \max_{\mathbf{Z}} \frac{1}{T^2} \operatorname{Tr}\left[-T\mathbf{X}^\top\mathbf{Y} + \frac{T}{2}\mathbf{Y}^\top\mathbf{Y} + \frac{\gamma^2}{2u^2}\mathbf{Y}^\top\mathbf{Y}\mathbf{Z}^\top\mathbf{Z} - \frac{\gamma^4}{4u^2\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right] \tag{S5}$$

The symmetry implies:

$$\mathbf{Y} = f(\mathbf{X}, \rho) \Leftrightarrow a\mathbf{Y} = f(a\mathbf{X}, \rho/a) \tag{S6}$$

Thus:

$$f(\mathbf{X}, \rho) = \frac{1}{a} f(a\mathbf{X}, \rho/a) \quad \text{and also} \quad f(a\mathbf{X}, \rho) = af(\mathbf{X}, a\rho) \tag{S7}$$

This means performing an optimization with an input $a\mathbf{X}$, is equivalent to doing the optimization with input $\mathbf{X}$ and parameter $a\rho$, and finally multiplying the obtained $\mathbf{Y}$ by a.

It is worth noting though, that for a circuit with fixed $\mathbf{W}$ and $\mathbf{M}$, scaling an input $\mathbf{x}$ by a factor a, simply scales the output $\mathbf{y}$ by the same factor a, since it is a linear transformation, at least for the circuit without the non-negative constraints.

1.2 Limiting case and relation to whitening

For the case when $D = K$, the optimum for $\mathbf{Z}$ is $\mathbf{Z} = \frac{\rho}{\gamma}\mathbf{Y}$ and thus the middle term of the objective function (S1) drops, with and without non-negativity constraints on $\mathbf{Y}$ and $\mathbf{Z}$. The objective function becomes:

$$\mathcal{L} = \min_{\mathbf{Y}} \frac{1}{T^2}\left(\frac{T}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2 + \frac{\rho^2}{4u^2}\left\|\mathbf{Y}^\top\mathbf{Y}\right\|_F^2\right) \tag{S8}$$

This objective function closely resembles the whitening objective function:

$$\begin{aligned}
\mathcal{L} &= \min_{\mathbf{Y}} \|\mathbf{X}-\mathbf{Y}\|_F^2 + \lambda \left\|\mathbf{Y}\mathbf{Y}^\top - \alpha^2 \mathbf{I}_D\right\|_F^2 && \text{(S9)}\\
&= \min_{\mathbf{Y}} \|\mathbf{X}-\mathbf{Y}\|_F^2 + \lambda \operatorname{Tr}\left[\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top - 2\alpha^2\mathbf{Y}\mathbf{Y}^\top + \alpha^4\mathbf{I}_D\right] && \text{(S10)}\\
&= \min_{\mathbf{Y}} \|\mathbf{X}-\mathbf{Y}\|_F^2 + \lambda \operatorname{Tr}\left[\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{Y} - 2\alpha^2\mathbf{Y}^\top\mathbf{Y}\right] && \text{(S11)}\\
&= \min_{\mathbf{Y}} \|\mathbf{X}-\mathbf{Y}\|_F^2 - 2\alpha^2\lambda\,\|\mathbf{Y}\|_F^2 + \lambda \left\|\mathbf{Y}^\top\mathbf{Y}\right\|_F^2 && \text{(S12)}
\end{aligned}$$

For a fixed α, increasing λ will eventually lead to perfect whitening. The singular values of $\mathbf{Y}$ will then all be equal to α, and the left and right singular vectors will be the same as those of $\mathbf{X}$.

2 Online solution

We show that the online algorithm to optimize the objective function (S2) can be mapped onto the architecture and neural dynamics of the olfactory neural circuit (**Fig. 1A**) with Hebbian plasticity. To find the online solution, we first introduce the unitless variables $\mathbf{W} \in \mathbb{R}^{D\times K}$ and $\mathbf{M} \in \mathbb{R}^{K\times K}$:

$$\mathbf{W} = \frac{1}{Tu^2}\mathbf{Y}\mathbf{Z}^\top, \quad \mathbf{M} = \frac{1}{Tu^2}\mathbf{Z}\mathbf{Z}^\top \tag{S13}$$

and perform the Hubbard-Stratonovich transform of (S2):

$$\begin{aligned}
\mathcal{L} = \min_{\mathbf{Y}}\max_{\mathbf{Z}}\max_{\mathbf{W}}\min_{\mathbf{M}} \frac{1}{T}\operatorname{Tr}\left[-\mathbf{X}^\top\mathbf{Y} + \frac{1}{2}\mathbf{Y}^\top\mathbf{Y} + \gamma^2\mathbf{Y}^\top\mathbf{W}\mathbf{Z} - \frac{\gamma^4}{2\rho^2}\mathbf{Z}^\top\mathbf{M}\mathbf{Z}\right]&\\
- \frac{u^2\gamma^2}{2}\operatorname{Tr}\left[\mathbf{W}^\top\mathbf{W}\right] + \frac{u^2\gamma^4}{4\rho^2}\operatorname{Tr}\left[\mathbf{M}^\top\mathbf{M}\right]& \qquad \text{(S14)}
\end{aligned}$$

We then rewrite (S14) in vector notation, with each sample point written out separately, and invert the order of min max (Pehlevan et al., 2018):

$$\begin{aligned}
\mathcal{L} = \max_{\mathbf{W}}\min_{\mathbf{M}}\min_{\{\mathbf{y}^{(t)}\}}\max_{\{\mathbf{z}^{(t)}\}} \frac{1}{T}\sum_{t=1}^{T}\left(-\mathbf{x}^{(t)\top}\mathbf{y}^{(t)} + \frac{1}{2}\mathbf{y}^{(t)\top}\mathbf{y}^{(t)} + \gamma^2\mathbf{y}^{(t)\top}\mathbf{W}\mathbf{z}^{(t)} - \frac{\gamma^4}{2\rho^2}\mathbf{z}^{(t)\top}\mathbf{M}\mathbf{z}^{(t)}\right)&\\
- \frac{u^2\gamma^2}{2}\operatorname{Tr}\left[\mathbf{W}^\top\mathbf{W}\right] + \frac{u^2\gamma^4}{4\rho^2}\operatorname{Tr}\left[\mathbf{M}^\top\mathbf{M}\right]& \qquad \text{(S15)}
\end{aligned}$$

Next we perform the optimization for each variable separately: $\mathbf{y}^{(t)}$, $\mathbf{z}^{(t)}$, $\mathbf{W}$, and $\mathbf{M}$. We consider the following case, which corresponds to the "online setting" for this objective function and alternate the optimization in $\{\mathbf{y}^{(t)}, \mathbf{z}^{(t)}\}$ and in $\{\mathbf{W}, \mathbf{M}\}$: as a new sample (i.e., stimulus, input) $\mathbf{x}^{(t)}$ arrives, we find the values of $\mathbf{z}^{(t)}$ and $\mathbf{y}^{(t)}$ with the current values $\mathbf{W}^{(t)}$ and $\mathbf{M}^{(t)}$ and update $\mathbf{W}^{(t)}$ and $\mathbf{M}^{(t)}$ to $\mathbf{W}^{(t+1)}$ and $\mathbf{M}^{(t+1)}$ before the arrival of the next sample $\mathbf{x}^{(t+1)}$. Biologically, this can be seen as first a convergence of neural spiking rates or neural electrical potential encoded through the variables $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$, and second a synaptic weight update based on those steady state activity values. At a given sample index t, the minimum in $\mathbf{y}^{(t)}$ and the maximum in $\mathbf{z}^{(t)}$ can be found by taking a derivative of (S15) with respect to $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$, respectively:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{y}^{(t)}} &= \frac{1}{T}\left(-\mathbf{x}^{(t)} + \mathbf{y}^{(t)} + \gamma^2 \mathbf{W}^{(t)}\mathbf{z}^{(t)}\right)\\ \frac{\partial \mathcal{L}}{\partial \mathbf{z}^{(t)}} &= \frac{1}{T}\left(\gamma^2 \mathbf{W}^{(t)\top}\mathbf{y}^{(t)} - \frac{\gamma^4}{\rho^2}\mathbf{M}^{(t)}\mathbf{z}^{(t)}\right)\end{aligned} \tag{S16}$$

The minimum in $\mathbf{y}^{(t)}$ and the maximum in $\mathbf{z}^{(t)}$ can be reached by a gradient descent and ascent, respectively. We can thus write a system of differential equations whose steady state correspond to the optimum:

$$\begin{cases}\tau_y \dfrac{d\mathbf{y}^{(t)}(\tau)}{d\tau} & = & -\mathbf{y}^{(t)}(\tau) & - & \gamma^2 \mathbf{W}^{(t)}\mathbf{z}^{(t)}(\tau) & + & \mathbf{x}^{(t)}\\ \tau_z \dfrac{d\mathbf{z}^{(t)}(\tau)}{d\tau} & = & -\mathbf{M}^{(t)}\mathbf{z}^{(t)}(\tau) & + & \rho^2/\gamma^2 \mathbf{W}^{(t)\top}\mathbf{y}^{(t)}(\tau) & & \end{cases} \tag{S17}$$

Where τ is the local time evolution variable. We rearranged the parameters so that the equation form is the same as in equations (19), which does not change the final steady state of the equations. Thus, we obtained equations to find the optima $\bar{\mathbf{y}}^{(t)}$ and $\bar{\mathbf{z}}^{(t)}$ of the objective function. As explained in the main text, these question can directly be mapped onto the dynamics of the ORN-LN neural circuit.

Next, we derived the updates for the variables $\mathbf{W}$ and $\mathbf{M}$. By construction, the offline solution for $\mathbf{W}$ and $\mathbf{M}$ is given by (S13). Online - we compute a new $\mathbf{W}^{(t)}$ and $\mathbf{M}^{(t)}$ after each sample $\mathbf{x}^{(t)}$ is presented and $\bar{\mathbf{y}}^{(t)}$ and $\bar{\mathbf{z}}^{(t)}$ are found. The gradient descent (respectively ascent) steps on these variables give the following updates (e.g., Pehlevan et al., 2018):

$$\begin{aligned}\mathbf{W}^{(t+1)} &= \mathbf{W}^{(t)} + \eta^{(t)}\left(\frac{\bar{\mathbf{z}}^{(t)}\bar{\mathbf{y}}^{(t)\top}}{u^2} - \mathbf{W}^{(t)}\right)\\ \mathbf{M}^{(t+1)} &= \mathbf{M}^{(t)} + \frac{\eta^{(t)}}{2\rho^2\nu}\left(\frac{\bar{\mathbf{z}}^{(t)}\bar{\mathbf{z}}^{(t)\top}}{u^2} - \mathbf{M}^{(t)}\right)\end{aligned} \tag{S18}$$

where $\eta^{(t)}$ and ν are parameters of the gradient descent/ascent, and where $\bar{\mathbf{y}}^{(t)}$ and $\bar{\mathbf{z}}^{(t)}$ are the steady states solutions of equations (S17) for given $\mathbf{W}^{(t)}$ and $\mathbf{M}^{(t)}$. This indeed corresponds to a local Hebbian synaptic update rules. Choosing $\eta^{(t)}$ and ν appropriately will lead to equation (21) from the main text.

2.1 Circuit equations for the NNC

In the case of the NNC, where we have objective function (18) instead of (S2), we get equation (S15) with non-negativity constraints:

$$\mathcal{L} = \max_{\mathbf{W}} \min_{\mathbf{M}} \min_{\{\mathbf{y}^{(t)} \geq 0\}} \max_{\{\mathbf{z}^{(t)} \geq 0\}} \frac{1}{T} \sum_{t=1}^{T} \left(-\mathbf{x}^{(t)\top}\mathbf{y}^{(t)} + \frac{1}{2}\mathbf{y}^{(t)\top}\mathbf{y}^{(t)} + \gamma^2 \mathbf{y}^{(t)\top}\mathbf{W}\mathbf{z}^{(t)} - \frac{\gamma^4}{2\rho^2}\mathbf{z}^{(t)\top}\mathbf{M}\mathbf{z}^{(t)} \right) - \frac{u^2\gamma^2}{2} \operatorname{Tr}\left[\mathbf{W}^\top\mathbf{W}\right] + \frac{u^2\gamma^4}{4\rho^2} \operatorname{Tr}\left[\mathbf{M}^\top\mathbf{M}\right] \quad \text{(S19)}$$

Here too, we perform the optimization for each variable separately: $\mathbf{y}^{(t)}$, $\mathbf{z}^{(t)}$, $\mathbf{W}$, and $\mathbf{M}$. However, because of the non-negativity constraints, the optima for $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ are not to be found at where the derivatives (S16) are zeros. We can, however, reach the optima by a projected gradient descent:

$$\begin{cases} \mathbf{y}^{(t)}(\tau+1) = \max\left[\mathbf{0},\ \mathbf{y}^{(t)}(\tau) + \epsilon(\tau)\big(-\mathbf{y}^{(t)}(\tau) - \gamma^2\mathbf{W}\mathbf{z}^{(t)}(\tau) + \mathbf{x}^{(t)}\big)\right] \\ \mathbf{z}^{(t)}(\tau+1) = \max\left[\mathbf{0},\ \mathbf{z}^{(t)}(\tau) + \epsilon(\tau)\big(-\mathbf{M}\mathbf{z}^{(t)}(\tau) + \rho^2/\gamma^2\mathbf{W}^\top\mathbf{y}^{(t)}(\tau)\big)\right] \end{cases} \quad \text{(S20)}$$

where the max is performed component-wise. For the NNC, the updates on $\mathbf{W}^{(t)}$ and $\mathbf{M}^{(t)}$ (equations (S18)) remain the same as for the LC.

2.2 Steady state solution of the circuit dynamical equations for the LC and stability

We can directly find the steady state solution of the circuit dynamics equations (S17) of the LC by setting the derivatives to 0. For $\mathbf{M}$ invertible, the steady state is (after dropping the index (t) for simplicity of notation):

$$\begin{cases} \bar{\mathbf{y}} = (\mathbf{I}_D + \rho^2\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^\top)^{-1}\mathbf{x} \\ \bar{\mathbf{z}} = \rho^2/\gamma^2\mathbf{M}^{-1}\mathbf{W}^\top\bar{\mathbf{y}} \end{cases} \quad \text{(S21)}$$

As mentioned above, the steady state for $\mathbf{y}$ does not depend on γ, whereas $\mathbf{z}$ does depend on γ. Note that the transformation from $\mathbf{x}$ to $\bar{\mathbf{y}}$ is symmetric: indeed, writing $\bar{\mathbf{y}} = \mathbf{F}\mathbf{x}$, we have that $\mathbf{F} = \mathbf{F}^\top$. This means that the transformation is diagonalizable. It will be shown below that this basis in which the transformation is diagonal corresponds to the PCA basis of $\mathbf{X}$.

Here we show that the fix point of equations (S17) is stable if $\mathbf{W}$ is maximum rank and $\mathbf{M}$ positive definite. We first rewrite the dynamical system:

$$\begin{bmatrix} \tau_y d\mathbf{y}(\tau)/d\tau \\ \tau_y d\mathbf{z}(\tau)/d\tau \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{I}_D & \gamma^2\mathbf{W} \\ -\rho^2/\gamma^2\mathbf{W}^\top & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{y}(\tau) \\ \mathbf{z}(\tau) \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix} - \mathbf{A} \begin{bmatrix} \mathbf{y}(\tau) \\ \mathbf{z}(\tau) \end{bmatrix} \quad \text{(S22)}$$

This system has a unique stable fix point if and only if $\mathbf{A}$ has only positive eigenvalues. To investigate under which conditions this is the case, we write the eigenvalue equations for $\mathbf{A}$:

$$\begin{bmatrix} \mathbf{I}_D & \gamma^2\mathbf{W} \\ -\rho^2/\gamma^2\mathbf{W}^\top & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \tag{S23}$$

$$\begin{cases} \mathbf{y} + \gamma^2\mathbf{W}\mathbf{z} = \lambda\mathbf{y} \\ -\rho^2/\gamma^2\mathbf{W}^\top\mathbf{y} + \mathbf{M}\mathbf{z} = \lambda\mathbf{z} \end{cases} \tag{S24}$$

$$\begin{cases} \gamma^2\mathbf{W}\mathbf{z} = (\lambda-1)\mathbf{y} \\ \rho^2/\gamma^2\mathbf{W}^\top\mathbf{y} = (\mathbf{M}-\lambda)\mathbf{z} \end{cases} \tag{S25}$$

We consider the case when $\lambda \neq 1$, as we are interested to see if λ could potentially be negative.

$$\mathbf{y} = (\lambda-1)^{-1}\gamma^2\mathbf{W}\mathbf{z} \tag{S26}$$

$$\implies \rho^2\mathbf{W}^\top\mathbf{W}\mathbf{z} = (\lambda-1)(\mathbf{M}-\lambda)\mathbf{z} \tag{S27}$$

$\mathbf{W}^\top\mathbf{W} \in \mathbb{R}^{K\times K}$ is a positive semi-definite matrix, it is positive definite if $\mathbf{W}$ is maximum rank (i.e., rank K). Assuming that $\mathbf{W}$ is full rank, the matrix $\mathbf{W}^\top\mathbf{W}$ on the left-hand side of the equation has only positive eigenvalues. The above equation does not have any solution $\mathbf{z} \neq \mathbf{0}$ for $\lambda < 0$ if $\mathbf{M}$ is positive definite (which is true when constructed as the autocorrelator of $\mathbf{z}$). Thus, $\mathbf{W}$ full rank and $\mathbf{M}$ positive definite are sufficient conditions for the dynamical system to always converges to a stable fix point.

3 Circuit dynamics equations contains two effective parameters

Here we show that, in its general form, the system of differential equation describing the olfactory circuit has just two effective parameters and can be reduced to equation (19) (or (20)) from the main text. Without lack of generality the system of differential equations yields:

$$\begin{cases} \tau_1 \dfrac{d\mathbf{y}(\tau)}{d\tau} = -a\mathbf{y}(\tau) - b\mathbf{W}_1\mathbf{z}(\tau) + a\mathbf{x} \\ \tau_2 \dfrac{d\mathbf{z}(\tau)}{d\tau} = -c\mathbf{M}\mathbf{z}(\tau) + d\mathbf{W}_2^\top\mathbf{y}(\tau) \end{cases} \tag{S28}$$

Where we imposed that $\mathbf{x} = \mathbf{y}$ in the case of no LN activity (i.e., $\mathbf{z} = \mathbf{0}$), that $a > 0$, $b > 0$, $c > 0$, $d > 0$, and that all ORNs have similar response properties (i.e., same coefficient in front of each x_i and y_i). To extract the effective parameters, we compute the steady state solution of equations (S28) by setting the derivatives to zero. We find, for invertible $\mathbf{M}$:

$$\begin{cases} \bar{\mathbf{y}} = \left(\mathbf{I}_D + \dfrac{bd}{ac}\mathbf{W}_1\mathbf{M}^{-1}\mathbf{W}_2^\top\right)^{-1}\mathbf{x} \\ \bar{\mathbf{z}} = \dfrac{d}{c}\mathbf{M}^{-1}\mathbf{W}_2^\top\bar{\mathbf{y}} \end{cases} \tag{S29}$$

This shows that we only have two degrees of freedom: $\frac{bd}{ac}$ and $\frac{d}{c}$. We define $\rho^2 := \frac{bd}{ac}$ and $\gamma^2 := \frac{c}{d}\rho^2 = \frac{b}{a}$. This gives us:

$$\begin{cases} \bar{\mathbf{y}} = \left(\mathbf{I}_D + \rho^2 \mathbf{W}_1 \mathbf{M}^{-1} \mathbf{W}_2^\top\right)^{-1} \mathbf{x} \\ \bar{\mathbf{z}} = \rho^2/\gamma^2 \mathbf{M}^{-1} \mathbf{W}_2^\top \mathbf{y} \end{cases} \tag{S30}$$

Now replacing these definitions into the original equations (S28) we get:

$$\begin{cases} \tau_1/a \dfrac{d\mathbf{y}(\tau)}{d\tau} &= -\mathbf{y}(\tau) &- &\gamma^2 \mathbf{W}_1 \mathbf{z}(\tau) &+ &\mathbf{x} \\ \tau_2/c \dfrac{d\mathbf{z}(\tau)}{d\tau} &= -\mathbf{M}\mathbf{z}(\tau) &+ &\rho^2/\gamma^2 \mathbf{W}_2^\top \mathbf{y}(\tau) \end{cases} \tag{S31}$$

By setting $\tau_y := \tau_1/a$, $\tau_z := \tau_2/c$ we obtain equation (19) from the main text (when $\mathbf{W}_1 = \mathbf{W}_2$):

$$\begin{cases} \tau_y \dfrac{d\mathbf{y}(\tau)}{d\tau} &= -\mathbf{y}(\tau) &- &\gamma^2 \mathbf{W}_1 \mathbf{z}(\tau) &+ &\mathbf{x} \\ \tau_z \dfrac{d\mathbf{z}(\tau)}{d\tau} &= -\mathbf{M}\mathbf{z}(\tau) &+ &\rho^2/\gamma^2 \mathbf{W}_2^\top \mathbf{y}(\tau) \end{cases} \tag{S32}$$

Thus, scaling $\mathbf{x}$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{M}$ is equivalent to controlling just two effective parameter γ and ρ. Scaling τ_y and τ_z does not influence the steady state solutions.

Increasing ρ increases the weight of feedforward connection, making the LN activity and the feedback inhibition stronger. Increasing γ simultaneously increases the feedback connection strength and decreases the feedforward connection strength. Changing γ influences the steady state solution of $\mathbf{z}$ but not $\mathbf{y}$. Thus, a manifold of circuits lead to the same steady state output $\mathbf{y}$. In addition, the same differential equations can be implemented by different circuits. For example, multiplying a differential equation by a parameter does not alter the final steady state, but gives yet another implementation to the circuit as a scaling of the synaptic weights and of the time constant.

4 Circuit computation

To understand the computation performed by the olfactory circuit, we analyzed the optimization done by the objective function (S2), which corresponds to the linear circuit (LC). We use the singular value decomposition (SVD) for $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$: $\mathbf{X} = \mathbf{U}_X \tilde{\mathbf{S}}_X \mathbf{V}_X^\top$, $\mathbf{Y} = \mathbf{U}_Y \tilde{\mathbf{S}}_Y \mathbf{V}_Y^\top$, $\mathbf{Z} = \mathbf{U}_Z \tilde{\mathbf{S}}_Z \mathbf{V}_Z^\top$, with the following convention: $\mathbf{U}_X, \mathbf{U}_Y \in \mathbb{R}^{D\times D}$, $\mathbf{U}_Z \in \mathbb{R}^{K\times K}$, $\mathbf{V}_X, \mathbf{V}_Y, \mathbf{V}_Z \in \mathbb{R}^{T\times T}$, $\tilde{\mathbf{S}}_X, \tilde{\mathbf{S}}_Y \in \mathbb{R}^{D\times T}$, $\tilde{\mathbf{S}}_Z \in \mathbb{R}^{K\times T}$, $\tilde{\mathbf{S}}_X, \tilde{\mathbf{S}}_Y, \tilde{\mathbf{S}}_Z$ only have values on the diagonal. We call $\mathbf{S} \in \mathbb{R}^{T\times T}$ the diagonal square matrix corresponding to the rectangular matrix $\tilde{\mathbf{S}}$, with padded zeros. Only the first D columns in $\mathbf{V}_X$ and $\mathbf{V}_Y$ and the first K in $\mathbf{V}_Z$ contain relevant information about $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, respectively. The left singular vectors $\mathbf{U}_X$, $\mathbf{U}_Y$, and $\mathbf{U}_Z$ are also the principal directions of the uncentered PCA of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, respectively. Whereas the values on the diagonal of $\tilde{\mathbf{S}}_X$, $\tilde{\mathbf{S}}_Y$, and $\tilde{\mathbf{S}}_Z$ are the square root of the variances of the corresponding PCA directions.

In the following, using two approaches, we prove that:

$$\mathbf{Y} = \mathbf{U}_X \tilde{\mathbf{S}}_Y \mathbf{V}_X^\top = \mathbf{U}_X \tilde{\mathbf{S}}_Y \tilde{\mathbf{S}}_X^+ \mathbf{U}_X^\top \mathbf{X} \tag{S33}$$

$$\mathbf{Z} = \rho/\gamma \mathbf{U}_Z \tilde{\mathbf{S}}_{Y|K} \mathbf{V}_X^\top = \rho/\gamma \mathbf{U}_Z \tilde{\mathbf{S}}_{Y|K} \tilde{\mathbf{S}}_X^+ \mathbf{U}_X^\top \mathbf{X} \tag{S34}$$

$$\text{with} \begin{cases} s_{Y,i}\left(1 + \dfrac{\rho^2}{u^2 T} s_{Y,i}^2\right) = s_{X,i} & 1 \le i \le K \quad \text{(S35a)} \\ s_{Y,i} = s_{X,i} & K+1 \le i \le D \quad \text{(S35b)} \\ \mathbf{U}_Z\text{: a degree of freedom} & \quad \text{(S35c)} \end{cases}$$

where $\mathbf{A}^+$ the Moore-Penrose pseudo-inverse of $\mathbf{A}$ and $\tilde{\mathbf{S}}_{Y|K}$ is the matrix with the first K columns of $\tilde{\mathbf{S}}_Y$. This proves the relations (4), (5), (6a), (6b) in the main text.

In other words, writing $\mathbf{Y} = \mathbf{F}\mathbf{X}$, we have that $\mathbf{F} = \mathbf{F}^\top = \mathbf{U}_X \tilde{\mathbf{S}}_Y \tilde{\mathbf{S}}_X^+ \mathbf{U}_X^\top$, $\mathbf{S}_Y \mathbf{S}_X^+$ being a diagonal matrix. This signifies that the linear transformation $\mathbf{F}$ does not perform any rotation of the input.

This explicit expressions for s_Y and s_Z are:

$$\begin{aligned} s_Y &= \frac{1}{\rho}\left(\frac{\sqrt{12T^3u^6 + 81T^2u^4\rho^2 s_X^2} + 9Tu^2\rho s_X}{18}\right)^{\frac{1}{3}} - \frac{1}{\rho}\left(\frac{\frac{2}{3}T^3u^6}{\sqrt{12T^3u^6 + 81T^2u^4\rho^2 s_X^2} + 9Tu^2\rho s_X}\right)^{\frac{1}{3}} \\ s_Z &= \frac{\rho}{\gamma} s_Y \end{aligned} \tag{S36}$$

The behavior of s_Y is such:

$$s_Y \approx \begin{cases} s_X & s_X \ll \frac{\sqrt{T}u}{\rho} \quad \text{(S37a)} \\ \sqrt[3]{\dfrac{Tu^2}{\rho^2} s_X} & s_X \gg \frac{\sqrt{T}u}{\rho} \quad \text{(S37b)} \end{cases}$$

Note that because $\mathbf{Z}$ only appears as $\mathbf{Z}^\top\mathbf{Z}$ in the objective function (S2), $\mathbf{U}_Z$ is a degree of freedom of the optimization. Thus, for $\{\mathbf{Y},\ \mathbf{Z},\ \mathbf{W},\ \mathbf{M}\}$ a solution of the optimization, $\{\mathbf{Y},\ \mathbf{QZ},\ \mathbf{WQ}^\top,\ \mathbf{QMQ}^\top\}$ is a solution as well, where $\mathbf{Q} \in \mathbb{R}^{K\times K}$ is an orthogonal matrix. Consequently, there is a manifold of $\mathbf{W}$, $\mathbf{M}$, and $\mathbf{Z}$ that satisfies the optimization for the LC.

4.1 Approach 1

In this approach, we first perform the minimization in $\mathbf{Z}$. Based on the similarity matching objective function results (Pehlevan et al., 2018), we know in the linear case that the right singular vectors of $\mathbf{Y}$ and $\mathbf{Z}$ are equal, and thus $\mathbf{V}_Y = \mathbf{V}_Z$. We also know that the top K singular values of $\mathbf{Y}$ and $\gamma/\rho\mathbf{Z}$ are equal ($\mathbf{Z}$ is K-dimensional, thus all other singular values of $\mathbf{Z}$ are 0), and thus

$s_{Z,i} = \rho/\gamma s_{Y,i}$. The similarity matching term becomes:

$$\begin{aligned}
\left\|\mathbf{Y}^\top\mathbf{Y} - \frac{\gamma^2}{\rho^2}\mathbf{Z}^\top\mathbf{Z}\right\|_F^2 &= \left\|\mathbf{V}_Y\mathbf{S}_Y^2\mathbf{V}_Y^\top - \frac{\gamma^2}{\rho^2}\mathbf{V}_Z\mathbf{S}_Z^2\mathbf{V}_Z^\top\right\|_F^2 && \text{(S38)}\\
&= \left\|\mathbf{V}_Y\left(\mathbf{S}_Y^2 - \frac{\gamma^2}{\rho^2}\mathbf{S}_Z^2\right)\mathbf{V}_Y^\top\right\|_F^2 && \text{(S39)}\\
&= \operatorname{Tr}\left[\left(\mathbf{S}_Y^2 - \frac{\gamma^2}{\rho^2}\mathbf{S}_Z^2\right)^2\right] && \text{(S40)}\\
&= \sum_{i=K+1}^{D} s_{Y,i}^4 && \text{(S41)}
\end{aligned}$$

And thus $\mathbf{U}_Z$ does not appear in the optimization and is a free parameter. Also $\left\|\mathbf{Y}^\top\mathbf{Y}\right\|_F^2 = \sum_{i=1}^{D} s_{Y,i}^4$. Thus, the objective function (S1) becomes:

$$\begin{aligned}
\mathcal{L} &= \min_{\mathbf{Y}} \frac{1}{T^2}\left(\operatorname{Tr}\left[-T\mathbf{X}^\top\mathbf{Y} + \frac{T}{2}\mathbf{Y}^\top\mathbf{Y}\right] - \frac{\rho^2}{4u^2}\sum_{i=K+1}^{D} s_{Y,i}^4 + \frac{\rho^2}{4u^2}\sum_{i=1}^{D} s_{Y,i}^4\right) && \text{(S42)}\\
&= \min_{\mathbf{Y}} \frac{1}{T^2}\left(\operatorname{Tr}\left[-T\mathbf{X}^\top\mathbf{Y}\right] + \frac{T}{2}\sum_{i=1}^{D} s_{Y,i}^2 + \frac{\rho^2}{4u^2}\sum_{i=1}^{K} s_{Y,i}^4\right) && \text{(S43)}
\end{aligned}$$

Thus there is a fourth order penalty on the first K singular values of $\mathbf{Y}$.

We now replace the remaining $\mathbf{X}$ and $\mathbf{Y}$ by their SVD:

$$\mathcal{L} = \min_{\mathbf{Y}} \frac{1}{T^2}\left(\operatorname{Tr}\left[-T\mathbf{V}_X\tilde{\mathbf{S}}_X\mathbf{U}_X^\top\mathbf{U}_Y\tilde{\mathbf{S}}_Y\mathbf{V}_Y\right] + \frac{T}{2}\sum_{i=1}^{D} s_{Y,i}^2 + \frac{\rho^2}{4u^2}\sum_{i=1}^{K} s_{Y,i}^4\right) \tag{S44}$$

Based on von Neumann trace inequality, given a fixed $\tilde{\mathbf{S}}_Y$, the trace term is minimized when $\mathbf{U}_Y = \mathbf{U}_X$ and $\mathbf{V}_Y = \mathbf{V}_X$. We are thus left with:

$$\mathcal{L} = \min_{\{s_{Y,i}\}} \frac{1}{T^2}\left(-T\sum_{i=1}^{D} s_{X,i}s_{Y,i} + \frac{T}{2}\sum_{i=1}^{D} s_{Y,i}^2 + \frac{\rho^2}{4u^2}\sum_{i=1}^{K} s_{Y,i}^4\right) \tag{S45}$$

Each $s_{Y,i}$ can be optimized independently. We take the derivative of (S45) with respect to $s_{Y,i}$ and equate it to 0. For $i > K$, we have $s_{Y,i} = s_{X,i}$. For $i \leq K$:

$$\begin{aligned}
-Ts_{X,i} + Ts_{Y,i} + \frac{\rho^2}{u^2}s_{Y,i}^3 &= 0 && \text{(S46)}\\
s_{X,i} &= s_{Y,i}\left(1 + \frac{\rho^2}{Tu^2}s_{Y,i}^2\right) && \text{(S47)}
\end{aligned}$$

This end the derivation.

4.2 Approach 2

We first find the stationary point of the objective function (S2) in $\mathbf{Y}$ by taking the partial derivative of $\mathcal{L}$ with respect to $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{X}\left(\mathbf{I}_T + \frac{\gamma^2}{Tu^2}\mathbf{Z}^\top\mathbf{Z}\right)^{-1} \tag{S48}$$

where $\mathbf{I}_T$ is the identity matrix of dimension T and replace this solution for $\mathbf{Y}$ into the objective function $\mathcal{L}$:

$$\mathcal{L} = \min_{\mathbf{Z}} \frac{1}{T^2}\operatorname{Tr}\left[\frac{T}{2}\mathbf{X}^\top\mathbf{X}\left(\mathbf{I}_T + \frac{\gamma^2}{Tu^2}\mathbf{Z}^\top\mathbf{Z}\right)^{-1} + \frac{\gamma^4}{4u^2\rho^2}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\right] \tag{S49}$$

Next we replace $\mathbf{X}$ and $\mathbf{Z}$ by their SVD, use the property of the trace $\operatorname{Tr}(\mathbf{AB}) = \operatorname{Tr}(\mathbf{BA})$ and the property of orthogonal matrices $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$:

$$\mathcal{L} = \min_{\mathbf{Z}} \frac{1}{T^2}\operatorname{Tr}\left[\frac{T}{2}\mathbf{V}_X\mathbf{S}_X^2\mathbf{V}_X^\top\left(\mathbf{I}_T + \frac{\gamma^2}{Tu^2}\mathbf{V}_Z\mathbf{S}_Z^2\mathbf{V}_Z^\top\right)^{-1} + \frac{\gamma^4}{4u^2\rho^2}\mathbf{S}_Z^4\right] \tag{S50}$$

$$= \min_{\mathbf{Z}} \frac{1}{T^2}\operatorname{Tr}\left[\frac{T}{2}\mathbf{V}_X\mathbf{S}_X^2\mathbf{V}_X^\top\left(\frac{\mathbf{V}_Z(Tu^2\mathbf{I}_T + \gamma^2\mathbf{S}_Z^2)\mathbf{V}_Z^\top}{Tu^2}\right)^{-1} + \frac{\gamma^4}{4u^2\rho^2}\mathbf{S}_Z^4\right] \tag{S51}$$

$$= \min_{\mathbf{Z}} \operatorname{Tr}\left[\frac{1}{2}\mathbf{V}_X\mathbf{S}_X^2\mathbf{V}_X^\top\mathbf{V}_Z(Tu^2\mathbf{I}_T + \gamma^2\mathbf{S}_Z^2)^{-1}\mathbf{V}_Z^\top + \frac{\gamma^4}{4T^2u^4\rho^2}\mathbf{S}_Z^4\right] \tag{S52}$$

Since $\mathbf{U}_Z$ does not appear in the minimization, it is a free parameter, i.e., it can be any orthogonal matrix. For fixed $\mathbf{S}_Z$, only the first term in the trace needs to be minimized. One can show that the optimal $\mathbf{V}_Z$ is $\mathbf{V}_Z = \mathbf{V}_X$: based on von Neumann trace inequality, we know that $\operatorname{Tr}[\mathbf{AB}] \geq \sum_i^N a_i b_{N-i+1}$ where a_i and b_i are the ordered singular values of $\mathbf{A}$ and $\mathbf{B}$, respectively. Thus, choosing $\mathbf{V}_Z = \mathbf{V}_X$ will give us the lower bound of that inequality. Indeed:

$$\begin{aligned}
&\operatorname{Tr}\left[\mathbf{V}_X\mathbf{S}_X^2\mathbf{V}_X^\top\mathbf{V}_Z(Tu^2\mathbf{I}_T + \gamma^2\mathbf{S}_Z^2)^{-1}\mathbf{V}_Z^\top\right] \\
&= \operatorname{Tr}\left[\mathbf{S}_X^2(Tu^2\mathbf{I}_T + \gamma^2\mathbf{S}_Z^2)^{-1}\right] \\
&= \sum_i^T s_{X,i}^2\frac{1}{Tu^2 + \gamma^2 s_{Z,i}^2}
\end{aligned} \tag{S53}$$

Where $s_{X,i}$ and $s_{Z,i}$ are the values on the diagonal of $\mathbf{S}_X$ and $\mathbf{S}_Z$, respectively. Thus, the highest singular values of $\mathbf{V}_X\mathbf{S}_X^2\mathbf{V}_X^\top$ match the lowest singular values of $\mathbf{V}_Z\left(Tu^2\mathbf{I}_T + \gamma^2\mathbf{S}_Z^2\right)^{-1}\mathbf{V}_Z^\top$, giving us the lower bound of the von Neumann inequality. Equation (S52) can now be simplified to:

$$\mathcal{L} = \min_{\{s_{Z,i}\}} \sum_i^T\left(\frac{1}{2}s_{X,i}^2\frac{1}{Tu^2 + \gamma^2 s_{Z,i}^2} + \frac{\gamma^4}{4T^2u^4\rho^2}s_{Z,i}^4\right) \tag{S54}$$

Each $s_{Z,i}$ can be minimized independently. By construction of SVD, we already have that $s_{Z,i} = 0$ for $i > K$. We thus consider $1 \leq i \leq K$. To simplify notation, we drop the index i. We take the derivative of (S54) with respect to $s_{Z,i}$ and equate it to 0:

$$\frac{\partial \mathcal{L}}{\partial s_Z} = 0 \tag{S55}$$

$$-\frac{\gamma^2}{(Tu^2+\gamma^2 s_Z^2)^2} s_X^2 s_Z + \frac{\gamma^4}{T^2u^4\rho^2} s_Z^3 = 0 \tag{S56}$$

$$s_X^2 = \frac{\gamma^2}{\rho^2}\frac{(Tu^2+\gamma^2 s_Z^2)^2}{T^2u^4} s_Z^2 \tag{S57}$$

Which leads to, considering that singular values are positive:

$$s_X = \frac{\gamma}{\rho} s_Z \left(1 + \frac{\gamma^2}{Tu^2} s_Z^2\right) \tag{S58}$$

We can now use the obtained solution for $\mathbf{Z}$ to find the solution for $\mathbf{Y}$. We replace $\mathbf{X}$ and $\mathbf{Z}$ by their SVD in relation (S48) and use that $\mathbf{V}_X = \mathbf{V}_Z$:

$$\mathbf{Y} = \mathbf{U}_Y \tilde{\mathbf{S}}_Y \mathbf{V}_Y^\top = \mathbf{X}\left(\mathbf{I}_T + \frac{\gamma^2}{Tu^2}\mathbf{Z}^\top\mathbf{Z}\right)^{-1} \tag{S59}$$

$$= \mathbf{U}_X \tilde{\mathbf{S}}_X \mathbf{V}_X^\top \left(\mathbf{I}_T + \frac{\gamma^2}{Tu^2}\mathbf{V}_X \mathbf{S}_Z^2 \mathbf{V}_X^\top\right)^{-1} \tag{S60}$$

$$= \mathbf{U}_X \tilde{\mathbf{S}}_X \mathbf{V}_X^\top \mathbf{V}_X \left(\mathbf{I}_T + \frac{\gamma^2}{Tu^2}\mathbf{S}_Z^2\right)^{-1} \mathbf{V}_X^\top \tag{S61}$$

$$\mathbf{U}_Y \tilde{\mathbf{S}}_Y \mathbf{V}_Y^\top = \mathbf{U}_X \tilde{\mathbf{S}}_X \left(\mathbf{I}_T + \frac{\gamma^2}{Tu^2}\mathbf{S}_Z^2\right)^{-1} \mathbf{V}_X^\top \tag{S62}$$

Equating the SVD terms on the left and right sides we obtain $\mathbf{U}_Y = \mathbf{U}_X$ and $\mathbf{V}_X = \mathbf{V}_Y$ and

$$s_{Y,i} = s_{X,i}\left(1 + \frac{\gamma^2}{Tu^2} s_{Z,i}^2\right)^{-1} \tag{S63}$$

Thus, for $i > K$, we have $s_{Y,i} = s_{X,i}$ (since $s_{Z,i} = 0$), whereas for $i \leq K$: $s_{Y,i} = \frac{\gamma}{\rho} s_{Z,i}$ (using relation (S58) to replace s_X). The relation analogous to (S58) is:

$$s_X = s_Y\left(1 + \frac{\rho^2}{Tu^2} s_Y^2\right) \tag{S64}$$

This ends the derivation.

4.3 Effect of ρ and relation to SNMF

Having the expression for the output $\mathbf{Y}$, we can now describe the effect of ρ on the computation. For $\rho \to 0$, $\mathbf{Z} \to 0$, leading to $\mathbf{X} = \mathbf{Y}$, which means that the output is equal to the input and no inhibition is taking place. On the other hand, for $\rho \to \infty$, the lowest $D - K$ singular values of $\mathbf{Y}$ remain the same, whereas top K drop to 0, i.e., the top K singular values are totally suppressed.

To better understand the behavior of the circuit for small ρ we do a first order expansion in ρ of $\mathbf{Y}$ around $\mathbf{X}$, i.e., $\mathbf{Y} = \mathbf{X} + \rho\boldsymbol{\Xi}$. Replacing this expression for $\mathbf{Y}$ in the objective function (S2), and keeping only the leading terms in ρ, the objective function becomes:

$$\mathcal{L} = \min_{Z} \left\| \gamma^2 \mathbf{Z}^\top \mathbf{Z} - \rho^2 \mathbf{X}^\top \mathbf{X} \right\|_F^2 \tag{S65}$$

Which corresponds to the basic similarity matching objective function (Pehlevan et al., 2018).

For the non-negative objective function, for small ρ we get $\mathbf{Y} = [\mathbf{X}]_+$ and the objective function simplifies to

$$\mathcal{L} = \min_{\mathbf{Z} \geq 0} \left\| \gamma^2 \mathbf{Z}^\top \mathbf{Z} - \rho^2 [\mathbf{X}]_+^\top [\mathbf{X}]_+ \right\|_F^2 \tag{S66}$$

Which corresponds to the symmetric non-negative matrix factorization (SNMF) objective function and can also be implemented online by a neural circuit (Pehlevan & Chklovskii, 2015).

5 Relationship between $\mathbf{W}$ and $\mathbf{M}$

Here we prove the relationship $\rho^2/\gamma^2 \mathbf{W}\mathbf{W}^\top = \mathbf{M}^2$ for the LC.

One way to obtain this relationship is to start from the circuit dynamics (equations (S17)). The steady state for $\bar{\mathbf{z}}^{(t)}$ is:

$$\rho^2/\gamma^2 \mathbf{W} \bar{\mathbf{y}}^{(t)} = \mathbf{M} \bar{\mathbf{z}}^{(t)} \tag{S67}$$

Multiplying by $\bar{\mathbf{z}}^{(t)\top}$ on both sides, taking the average over all samples t, and using the definition of $\mathbf{W}$ and $\mathbf{M}$ (equation (S13)):

$$\rho^2/\gamma^2 \mathbf{W} \mathbf{E} \left[\bar{\mathbf{y}}^{(t)} \bar{\mathbf{z}}^{(t)\top} \right] / u^2 = \mathbf{M} \mathbf{E} \left[\bar{\mathbf{z}}^{(t)} \bar{\mathbf{z}}^{(t)\top} \right] / u^2 \tag{S68}$$

$$\rho^2/\gamma^2 \mathbf{W}\mathbf{W}^\top = \mathbf{M}^2 \tag{S69}$$

An alternative approach to find the above relationship is to use the definition of $\mathbf{W}$ and $\mathbf{M}$ (equation (S13)) and the SVD decomposition of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. We write out $\mathbf{W}$ and $\mathbf{M}$:

$$\mathbf{W} = \frac{1}{Tu^2} \mathbf{Y}\mathbf{Z}^\top = \frac{1}{Tu^2} \mathbf{U}_Y \tilde{\mathbf{S}}_Y \mathbf{V}_Y^\top \mathbf{V}_Z \tilde{\mathbf{S}}_Z^\top \mathbf{U}_Z^\top = \frac{1}{Tu^2} \mathbf{U}_X \tilde{\mathbf{S}}_Y \tilde{\mathbf{S}}_Z^\top \mathbf{U}_Z^\top = \frac{\gamma}{Tu^2 \rho} \mathbf{U}_{X|K} \hat{\mathbf{S}}_Z^2 \mathbf{U}_Z^\top \tag{S70}$$

$$\mathbf{M} = \frac{1}{Tu^2} \mathbf{Z}\mathbf{Z}^\top = \frac{1}{Tu^2} \mathbf{U}_Z \tilde{\mathbf{S}}_Z \mathbf{V}_Z^\top \mathbf{V}_Z \tilde{\mathbf{S}}_Z^\top \mathbf{U}_Z^\top = \frac{1}{Tu^2} \mathbf{U}_Z \hat{\mathbf{S}}_Z^2 \mathbf{U}_Z^\top \tag{S71}$$

Where we used that $\mathbf{V}_X = \mathbf{V}_Y = \mathbf{V}_Z$ and $\mathbf{U}_X = \mathbf{U}_Y$ are orthogonal matrices and that $s_{Y,i} = \frac{\gamma}{\rho} s_{Z,i}$ for $i \leq K$ and $s_{Z,i} = 0$ for $i > K$. We call $\hat{\mathbf{S}}_Z \in \mathbb{R}^{K \times K}$ the small square submatrix of the rectangular matrix $\mathbf{S}_Z \in \mathbb{R}^{K\times N}$. $\mathbf{U}_{X|K} \in \mathbb{R}^{D\times K}$ is the submatrix with the first K columns of $\mathbf{U}_X$. Thus:

$$\mathbf{W}^\top \mathbf{W} = \frac{\gamma^2}{T^2 u^4 \rho^2} \mathbf{U}_Z \hat{\mathbf{S}}_Z^2 \mathbf{U}_{X|K}^\top \mathbf{U}_{X|K} \hat{\mathbf{S}}_Z^2 \mathbf{U}_Z^\top \tag{S72}$$

$$= \frac{\gamma^2}{T^2 u^4 \rho^2} \mathbf{U}_Z \hat{\mathbf{S}}_Z^4 \mathbf{U}_Z^\top = \frac{\gamma^2}{\rho^2} \mathbf{M}^2 \tag{S73}$$

Taking the square root on both sides gives the relationship (2) in the results section.

6 Relationship between the statistics of ORN activity and ORN-LN connectivity

Based on the expressions for $\mathbf{W}$ and $\mathbf{M}$ (equations (S70) and (S71)) we can write $\mathbf{W}$ as:

$$\mathbf{W} = \frac{\gamma}{Tu^2\rho} \mathbf{U}_{X|K} \hat{\mathbf{S}}_Z^2 \mathbf{U}_Z^\top = \frac{\gamma}{Tu^2\rho} \mathbf{U}_{X|K} \mathbf{U}_Z^\top \mathbf{U}_Z \hat{\mathbf{S}}_Z^2 \mathbf{U}_Z^\top = \frac{\gamma}{\rho} \mathbf{U}_{X|K} \mathbf{U}_Z^\top \mathbf{M} \tag{S74}$$

Where we used that $\mathbf{U}_Z^\top \mathbf{U}_Z = \mathbf{I}_K$. Where $\mathbf{U}_{X|K} \in \mathbb{R}^{D\times K}$ is the submatrix with the first K columns of $\mathbf{U}_X$. As stated above, $\mathbf{U}_Z$ is a free parameter and could be any orthogonal matrix.

In the case of a single LN, $\mathbf{W}$ is a column vector and corresponds to the first left eigenvector of $\mathbf{X}$. For multiple LNs, the column vectors of $\mathbf{W}$ span the same subspace as the top K loading vectors of $\mathbf{X}$, $\mathbf{U}_{X|K}$. However, because of the multiplication on the right by $\mathbf{U}_Z^\top \mathbf{M}$, the connections vectors do not necessarily correspond to specific PCA directions and are not orthogonal, but only span the top K-dimensional PCA subspace. Thus, this relation above gives us the relationship between the left eigenvectors of $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{M}$.

7 Decrease of the spread of the spectrum of singular values

Here we show that the coefficient of variation (CV, i.e., the spread) of singular values is smaller at the ORN output (axons) than at the input (somas) in the LC model with the number of ORNs equal to the number of LN. In that case, we have $s_X = s_Y \left(1 + \frac{\rho^2}{T} s_Y^2\right)$. As we have shown, for small s_X, we have $s_Y \approx s_X$ and for large s_X, we have $s_Y \approx \left(T/\rho^2 s_X\right)^{1/3}$. We call X a positive random variable. We will show that for a $0 < \alpha < 1$, $\mathrm{CV}(X) \geq \mathrm{CV}(X^\alpha)$, which mimics the case we have.

$$\begin{aligned}
& \mathrm{CV}(X) \geq \mathrm{CV}(X^\alpha) && \text{(S75)} \\
\Leftrightarrow\ & \frac{\sigma_X}{\mathbf{E}\left[X\right]} \geq \frac{\sigma_{X^\alpha}}{\mathbf{E}\left[X^\alpha\right]} && \text{(S76)} \\
\Leftrightarrow\ & \frac{\sigma_X^2}{\mathbf{E}\left[X\right]^2} \geq \frac{\sigma_{X^\alpha}^2}{\mathbf{E}\left[X^\alpha\right]^2} && \text{(S77)} \\
\Leftrightarrow\ & \frac{\mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2}{\mathbf{E}\left[X\right]^2} \geq \frac{\mathbf{E}\left[X^{2\alpha}\right] - \mathbf{E}\left[X^\alpha\right]^2}{\mathbf{E}\left[X^\alpha\right]^2} && \text{(S78)} \\
\Leftrightarrow\ & \frac{\mathbf{E}\left[X^2\right]}{\mathbf{E}\left[X\right]^2} \geq \frac{\mathbf{E}\left[X^{2\alpha}\right]}{\mathbf{E}\left[X^\alpha\right]^2} && \text{(S79)}
\end{aligned}$$

The last inequality can be proven by using Hölder's inequality twice. First:

$$\left(\mathbf{E}\left[X^2\right]\right)^{\frac{1-\alpha}{2-\alpha}} \left(\mathbf{E}\left[X^\alpha\right]\right)^{\frac{1}{2-\alpha}} \geq \mathbf{E}\left[X\right] \tag{S80}$$

which leads to:

$$\frac{\mathbf{E}\left[X^2\right]}{\mathbf{E}\left[X\right]^2} \geq \frac{\left(\mathbf{E}\left[X^2\right]\right)^{\frac{\alpha}{2-\alpha}}}{\left(\mathbf{E}\left[X^\alpha\right]\right)^{\frac{2}{2-\alpha}}} \tag{S81}$$

and second:

$$\left(\mathbf{E}\left[X^2\right]\right)^{\frac{\alpha}{2-\alpha}} \left(\mathbf{E}\left[X^\alpha\right]\right)^{\frac{2-2\alpha}{2-\alpha}} \geq \mathbf{E}\left[X^{2\alpha}\right] \tag{S82}$$

which leads to:

$$\frac{\left(\mathbf{E}\left[X^2\right]\right)^{\frac{\alpha}{2-\alpha}}}{\left(\mathbf{E}\left[X^\alpha\right]\right)^{\frac{2}{2-\alpha}}} \geq \frac{\mathbf{E}\left[X^{2\alpha}\right]}{\mathbf{E}\left[X^\alpha\right]^2} \tag{S83}$$

Combining inequalities (S81) and (S83) proves inequality (S79) and ends the proof.