PNAS RESEARCH ARTICLE | NEUROSCIENCE APPLIED MATHEMATICS OPEN ACCESS

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

Nikolai M. Chapochnikov[a,b,1], Cengiz Pehlevan[c,d,e], and Dmitri B. Chklovskii[a,f]

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received September 23, 2021; accepted May 8, 2023

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 formulate biologically plausible mechanistic models of the circuit. In particular, we consider a linear circuit model, for which we derive an exact theoretical solution, and a nonnegative circuit model, which we examine through simulations. The latter largely predicts the ORN → LN synaptic weights found in the connectome and demonstrates that they reflect correlations in ORN activity patterns. Furthermore, this model accounts for the relationship between ORN → LN and LN–LN synaptic counts and the emergence of different LN types. Functionally, we propose that LNs encode soft cluster memberships of ORN activity, and partially whiten and normalize the stimulus representations in ORNs through inhibitory feedback. Such a synaptic organization could, in principle, autonomously arise through Hebbian plasticity and would allow the circuit to adapt to different environments in an unsupervised manner. We thus uncover a general and potent circuit motif that can learn and extract significant input features and render stimulus representations more efficient. Finally, our study provides a unified framework for relating structure, activity, function, and learning in neural circuits and supports the conjecture that similarity-matching shapes the transformation of neural representations.

olfaction | connectome | encoding | clustering | normative approach

Technological advances in connectomics (1, 2) and neural population activity imaging (3) enable the anatomical and physiological characterization of neural circuits at unprecedented scales and detail. However, it remains unclear how to combine these datasets to advance our understanding of brain computation. To address this, we focus on the peripheral olfactory system of the first instar *Drosophila* larva—a small and genetically tractable circuit with available connectivity and activity imaging datasets (4, 5).

This circuit is an analogous but simpler version of the well-studied olfactory circuit in adult flies and vertebrates (6). It contains 21 olfactory receptor neurons (ORNs), each expressing a different receptor type (Fig. 1*A*). 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 only the presence or absence of a connection between two neurons, but also the number of synaptic contacts in parallel (4), which is an estimate of the connection strength (2, 7–9) (nonetheless, other factors like release probability and active zone properties also affect synaptic strength (10, 11)).

Previous studies examined the role of LNs in transforming the neural representation of odors from ORN somas to downstream projection neurons (PNs). In adult *Drosophila*, this circuit was suggested to perform gain control and divisive normalization (12, 13), 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 odor discrimination downstream (14, 15).

However, the underlying mechanistic principles of computation remain elusive. For example, while different types of LNs have different connectivity patterns with ORNs in the *Drosophila* larva (4), the role of different LN types, their multiplicity, and their specific connectivity is not yet understood. Furthermore, the peripheral olfactory circuit of adult *Drosophila* exhibits synaptic plasticity in response to changes in the olfactory environment (16–19), but the functional role of this 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,

Significance

The brain represents information with neural activity patterns. At the periphery, these patterns contain correlations, which are detrimental to stimulus discrimination. We study the peripheral olfactory circuit of the *Drosophila* larva, which preprocesses neural representations before relaying them downstream. A comprehensive understanding of this preprocessing is, however, lacking. We formulate a principle-driven framework based on similarity-matching and, using neural input activity, derive a circuit model that largely explains the biological circuit's synaptic organization. It also predicts that inhibitory neurons cluster odors and facilitate decorrelation and normalization of neural representations. If equipped with Hebbian synaptic plasticity, the circuit model autonomously adapts to different environments. Our work provides a comprehensive approach to deciphering the relationship between structure and function in neural circuits.

Author contributions: N.M.C., C.P., and D.B.C. designed research; C.P. and D.B.C. formulated the optimization problem; N.M.C. and C.P. performed theoretical derivations; N.M.C. wrote the code, analyzed the data, and performed numerical simulations; and N.M.C. wrote the paper with input from C.P. and D.B.C.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

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

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2117484120/-/DCSupplemental.

Published July 10, 2023.

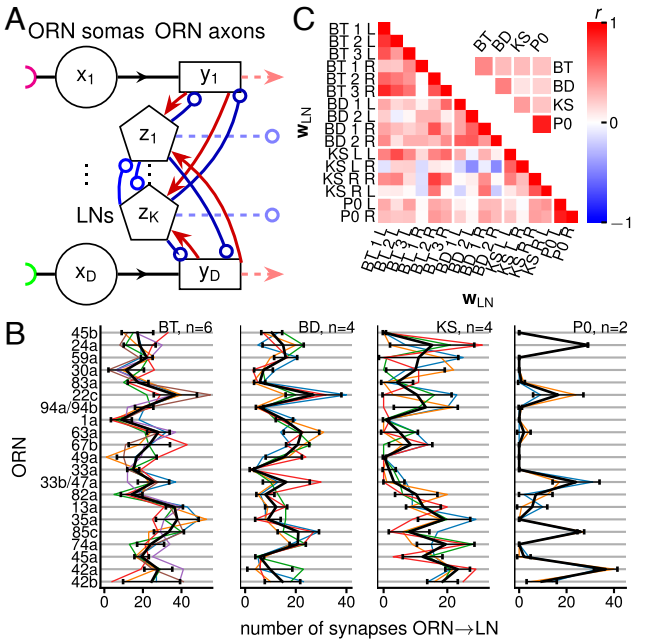

Fig. 1. Circuit connectivity and LN types. (*A*) ORN-LN circuit diagram. x_i, y_i, z_i: activity each ORN soma (circle), axonal terminal (rectangle), and LN (pentagon). Each ORN is depicted as a two-compartment unit with a soma and an axon. Half-circles: different types of chemical receptors. Red lines with arrowheads, blue lines with open circles: excitatory and inhibitory connections. LNs reciprocally connect with ORN axons and between themselves. ORN axons and LNs synapse onto neurons downstream (dashed lines). (*B*) Feedforward ORNs → LN synaptic count vectors, $\mathbf{w}_{LN}$ (colored lines), and average feedforward ORNs → LN_{type} synaptic count vectors, $\mathbf{w}_{LNtype}$ (black lines, mean ± SD) for each LN type (*SI Appendix*, Fig. S2*A*). (*C*) Correlation coefficients r between all $\mathbf{w}_{LN}$. L, R: left and right side of the Drosophila larva. The numerical indices of BT and BD are arbitrary, and there is no correspondence between the left and right side indices. Although BT 1 R is of the same type as other BT, its connection vector has a correlation of 0 with other BT in the connectome data. Inset: Mean rectified correlation coefficient $\bar{r}_+$ ($r_+ := \max[0, r]$, i.e., negative values are set to 0) between LN types calculated by averaging the rectified values in each region delimited by a white border, excluding the diagonal entries of the full matrix.

activity data, and learning. Our contribution is fivefold: 1) We find that the vectors of the number of synapses between ORNs and LNs reflect features of the independently acquired ORN activity pattern dataset (Figs. 2 and 3). 2) Building upon the normative similarity-matching framework (20, 21), we develop an optimization problem solvable by a biologically realistic circuit model with the same architecture as the ORN-LN circuit. 3) The model, driven by the ORN activity dataset, largely predicts the following observations in the structural dataset (Figs. 3 and 4): the ORNs → LN synaptic weights, the emergence of LN groups, and the relationship between feedforward ORN → LN and lateral LN–LN connections. 4) Using our model, we characterize the circuit computation (Figs. 5 and 6), and propose that LNs play a dual role in rendering the neural representation of odors in ORNs more efficient and extracting useful features that are transmitted downstream. 5) We show that the synaptic weights that enable this computation can, in principle, be learned in an unsupervised manner via Hebbian plasticity. Note that, given the connectome (4) originates from a 6-h-old first instar *Drosophila* larva, new synaptic contact formation can take longer than 6 h (11), and no study has yet demonstrated activity-dependent plasticity in the larval ORN-LN circuit, it is unknown whether the observed synaptic counts in this connectome could result from activity-dependent synaptic plasticity.

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 allow 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 PCA variances, which enables a more efficient stimulus encoding (22). In fact, 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 (23, 24) and divisive normalization (25). In essence, we uncover an elegant neural circuit motif that can extract features and perform two critical computations. If endowed with Hebbian plasticity, the circuit can also adapt and perform its functions in different stimulus environments. Thus, we present a framework that allows us 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 toward a more integrated understanding of neural circuits.

The results are organized as follows. First, we show that the connectome is adapted to ORN activity patterns. Second, we propose a normative approach leading to two circuit models: a linear circuit (LC) model, and a nonnegative circuit (NNC) model. Third, we show that the NNC reproduces key structural observations. Finally, we describe the computations performed by the LC and NNC in general and on the ORN activity dataset in particular.

Results

ORN-LN Circuit. ORNs in the *Drosophila* larva carry odor information from the antennas to the antennal lobe, where they synapse onto LNs and PNs. There, olfactory information is reformatted and transferred through ORN axons and LNs to PNs. LNs, which synapse bidirectionally with ORN axons and PN dendrites, strongly contribute to the reformatting in ORNs and PNs through presynaptic and postsynaptic inhibition, as shown mainly in the adult fly (12, 13, 26–30). LNs project to several uni- and multiglomerular PNs, and PNs project to higher brain areas such as the mushroom body and the lateral horn (4).

We study the circuit and computation presynaptic to PNs, i.e., occurring from ORN somas to ORN axons and LNs. Specifically, we examine the subcircuit formed by all $D = 21$ ORNs and those 4 LN types (on each side of the brain) that reciprocally connect with ORNs (4) (Fig. 1*A*, *SI Appendix*, Fig. 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 (*SI Appendix*, Figs. S1 and S2*A*). This amounts to 8 ORNs–LN connections per side (3 BTs, 2 BDs, 2 KSs, and 1 P0s) and 16 on both sides. See *SI Appendix*, Tables S1 and S2 for a list of all acronyms and mathematical variables used in the paper.

We use the number of synaptic contacts in parallel between two neurons as a proxy for the synaptic weight (2, 7–9) (but see refs. 10 and 11). In the linear approximation, the change in the postsynaptic neuron activity due to a change in the presynaptic neuron activity is proportional to the synaptic weight connecting them.

We focus our analysis on the synaptic counts of the feedforward ORNs → LN connections. We call $\mathbf{w}_{LN}$ the $D = 21$ dimensional vector containing the synaptic counts of the connections from the 21 ORNs to one LN. Because all the entries of this synaptic count vector $\mathbf{w}_{LN}$ share the same postsynaptic neuron, this

vector is likely proportional to the corresponding synaptic weight vector. Conversely, the synaptic count vector from one LN to all 21 ORNs may not be proportional to the corresponding synaptic weight vector, because each connection affects a different postsynaptic ORN, which potentially has different electrical properties. This makes the entries of a feedback synaptic count vector not directly comparable. Yet, the feedforward and feedback synaptic count vectors are somewhat correlated (*SI Appendix*, Fig. S2).

While the study (4) divided LNs into the above types based on their neuronal lineage, morphology, and qualitative connectivity, we also find that these types are innervated differently by ORNs (Fig. 1*B*). Indeed, the average correlation of $\mathbf{w}_{\mathrm{LN}}$s within each LN type is higher than between LN types (Fig. 1*C*). Thus, for a part of our study (Figs. 2 and 3 *A* and *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.

ORNs → LN Synaptic Count Vectors Are Adapted to Odor Representations in ORNs. Several studies proposed that LNs could facilitate the decorrelation of the neural representation of odors (14, 15, 32–35). To perform such decorrelation, the circuit must be adapted to or "know about" the correlations in the activity patterns (36). We investigate whether this is the case in this olfactory circuit by testing whether the $\mathbf{w}_{\mathrm{LNtype}}$s contain signatures of ORN activity patterns.

An ensemble of ORN activity patterns $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ ($t = 1, \ldots, 170$) was obtained using Ca^{2+} fluorescence imaging of ORN somas in response to a set of 34 odorants at 5 dilutions (5) (Fig. 2*A* and *SI Appendix*). 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).

We examine the Pearson correlation coefficients between the activity patterns $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ and the ORNs → $\mathrm{LN}_{\mathrm{type}}$ synaptic count vectors $\{\mathbf{w}_{\mathrm{LNtype}}\}$ (Fig. 2 *C* and *D* for $\mathbf{w}_{\mathrm{BT}}$ and two odors; Fig. 2*B* for all four $\mathbf{w}_{\mathrm{LNtype}}$s and all activity patterns $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$). After controlling for multiple comparisons (31), we find that the $\mathbf{w}_{\mathrm{LNtype}}$s for the Broad Trio and Picky 0 maintain significant correlations ($P < 0.05$) with a selection of ORN activity patterns, BT being highly correlated with the largest set of $\mathbf{x}^{(t)}$s. This suggests that the synaptic count vectors of at least these two LN types are more adapted to these activity patterns than would be expected by chance (see *SI Appendix*, Fig. S4 and *SI Appendix* for additional evidence). This supports the hypothesis that the circuit is at least partially adapted to ORN activity patterns and that it could perform a computation like decorrelation of input stimuli.

Each $\mathbf{w}_{\mathrm{LNtype}}$ exhibits a different "connectivity tuning curve" shape (Fig. 2*G*), $\mathbf{w}_{\mathrm{BT}}$ being correlated with the largest set of $\mathbf{x}^{(t)}$s, and $\mathbf{w}_{\mathrm{P0}}$ the most highly correlated to a few $\mathbf{x}^{(t)}$s, and the $\mathbf{w}_{\mathrm{BD}}$ and $\mathbf{w}_{\mathrm{KS}}$ the most weakly correlated. Biologically, this could signify that the BT type is activated by the largest set of odors and P0 only by a few odors. One possibility is that a different set of odors activates each LN class.

A Normative and Mechanistic Model of the ORN-LN Circuit. We aim to understand the circuit's computation and organization using a top-down, normative (also called principle-driven) approach, which involves formulating an optimization problem. Such an approach provides us with a theoretical understanding of the computation and organizational principles of the circuit. Although a bottom-up modeling approach requires unavailable

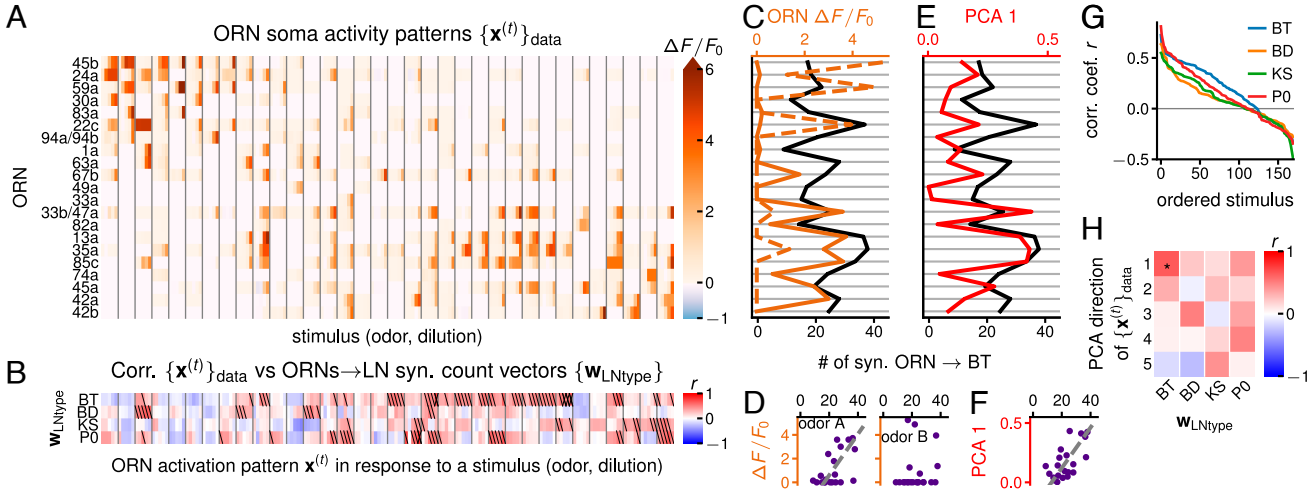

Fig. 2. Alignment of ORNs → LN synaptic count vectors with odor representations in ORN activity. (*A*) Ca^{2+} $\Delta F/F_0$ activity patterns $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ in ORN somas in response to 34 odors (separated by vertical gray lines) at 5 dilutions ($10^{-8}, \ldots, 10^{-4}$) from ref. 5. See *SI Appendix*, Fig. S3 for odor labels and scaled $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$. (*B*) Correlation between the four ORNs → $\mathrm{LN}_{\mathrm{type}}$ synaptic count vectors ($\mathbf{w}_{\mathrm{LNtype}}$ for BT, BD, KS, and P0) with the odor representations $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ from (*A*). Slash: significant at 0.05 level; cross: significant at 0.05 FDR (false discovery rate) (31). *P*-values calculated by shuffling the entries of each $\mathbf{w}_{\mathrm{LNtype}}$ (50,000 permutations). (*SI Appendix*, Figs. S4*A* and S5). (*C*) ORNs → Broad Trio synaptic count vector $\mathbf{w}_{\mathrm{BT}}$ superimposed with ORN activity patterns $\mathbf{x}^{(\mathrm{A})}$ and $\mathbf{x}^{(\mathrm{B})}$ in response to the ligands 2-heptanone (odor A) and 2-acetylpyridine (odor B) at dilution 10^{-4}. y-axis: ORN, follows order of (*A*). (*D*) Scatter plot representation of (*C*). $\mathbf{w}_{\mathrm{BT}}$ is more strongly tuned to $\mathbf{x}^{(\mathrm{A})}$ ($r = 0.6$, $P = 0.004$) than to $\mathbf{x}^{(\mathrm{B})}$ ($r = 0.14$, $P = 0.3$). *P*-values not adjusted for multiple comparisons. (*E*) $\mathbf{w}_{\mathrm{BT}}$ superimposed on the 1st PCA direction of $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$. *y*-axis: ORN, follows order of (*A*). (*F*) Scatter plot representation of (*E*) ($r = 0.65$, $P = 0.001$). *P*-values are not adjusted for multiple comparisons. (*G*) LN "connectivity tuning curves": correlation coefficients sorted in decreasing order from (*B*) for each $\mathbf{w}_{\mathrm{LNtype}}$. (*H*) Correlation coefficient *r* between the top 5 PCA directions of $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ and the four $\mathbf{w}_{\mathrm{LNtype}}$s (*SI Appendix*, Fig. S6 *A*, *B*, and *E*). Two-sided *P*-values calculated by shuffling the entries of each $\mathbf{w}_{\mathrm{LNtype}}$ (50,000 permutations). *: significance at 0.05 FDR.

physiological circuit parameters, we verified our predictions with a connectome-constrained model (Fig. 6).

Previous studies suggest that analogous circuits perform stimulus whitening or decorrelation (14, 15, 32–35), and our analysis above supports the possibility of such a computation. A class of optimization problems based on the similarity-matching principle and solvable by circuits similar to the ORN-LN one has been shown to be capable of implementing whitening, principal subspace extraction, and clustering (20, 21, 37). Note that the circuit's synaptic weights are adapted (optimized) to the ensemble of inputs to perform such computation.

To understand the circuit, we first postulate an optimization problem (Eq. **4**) based on the similarity-matching principle and solvable by a circuit with the ORN-LN architecture (see *Methods* and *SI Appendix*). To match this architecture, similarity-matching takes place between ORN axon and LN activities, which seeks to maintain that distances (similarities) between neural representations at the level of ORN axons and LNs. Specifically, if the representations of two odors are similar (dissimilar) in ORN axons, their representations will also tend to be similar (dissimilar) in LNs. Second, we derive the circuit models (Eqs. **5**–**7**) that solve this optimization problem with the recorded ORN soma activity described above (5) as input. Third, we compare the synaptic weight organization of the circuit model with the connectome (4) (Figs. 2, 3, and 4) and find that the circuit model accounts for multiple experimental observations. We thus conclude that the similarity-matching principle and the optimization problem widely explain the biological circuit's organization. Lastly, we describe in detail the computation performed by the circuit model (Figs. 5 and 6).

Mathematically, given a set of T activity patterns in D ORN somas as input, $\{\mathbf{x}^{(t)}\}_{t=1\ldots T}$, the optimization provides us as output the activity patterns in the D ORN axons $\{\mathbf{y}^{(t)}\}_{t=1\ldots T}$ and K LNs $\{\mathbf{z}^{(t)}\}_{t=1\ldots T}$. The circuit model performing the computation of the optimization has the following parameters: $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] := \mathbf{E}[\mathbf{y}^{(t)}\mathbf{z}^{(t)\mathrm{T}}]$ and $\mathbf{M} = \{m_{i,j}\}_{i,j=1\ldots K} := \mathbf{E}[\mathbf{z}^{(t)}\mathbf{z}^{(t)\mathrm{T}}]$, which are proportional to the connection weights between ORNs and LNs, and between LNs, respectively. In addition to K, the number of LNs, the model contains only one effective parameter ρ^2, corresponding to the ratio between feedback inhibition and feedforward excitation strengths.

We consider two optimization problems leading to two circuit models, differing in their domain of optimization: 1) a linear circuit, LC-K with K LNs, Eq. **6**, with no constraint on the optimization domain; 2) a nonnegative circuit, NNC-K, Eq. **7**, with nonnegative constrains on ORN axon and LN activity ($\mathbf{y}^{(t)} \geq 0, \mathbf{z}^{(t)} \geq 0$). This constraint renders the NNC more biologically plausible than the LC, and the NNC indeed predicts the structural data better than the LC (below). However, only for the LC we can derive the analytical expressions for $\mathbf{W}$, $\mathbf{M}$, $\{\mathbf{y}^{(t)}\}$, and $\{\mathbf{z}^{(t)}\}$, whereas for the NNC we must rely on numerical simulations (*SI Appendix*). Because both models are closely related, we examine the analytical solution of the LC to quantitatively understand the relationship between input and output variables, describe the circuit's function in a mathematically tractable manner, and substantiate the numerical results for the NNC.

Given an input $\{\mathbf{x}^{(t)}\}$, the optimal synaptic weights can be found by solving the optimization problem offline (Eqs. **4** and **5**), or online with Hebbian plasticity (Eq. **8**). The latter implies that the circuit model's synaptic weights can adapt to solve the optimization problem on any ORN activity patterns ensemble, in an unsupervised manner. This would correspond to activity-dependent synaptic plasticity in the biological circuit, which was, so far, only observed in the adult *Drosophila* (16–19). Given the specific wiring of some LNs such as Keystone and Picky 0 in the biological circuit (4), it is very likely that the synaptic weights of these (and potentially other) LNs are largely genetically predetermined and were set over evolutionary time scales (similar to an offline setting). It is unknown which mechanisms determine the synaptic weights in the biological circuit, and it is beyond the scope of this study to elucidate them.

Next, we characterize the computation performed by the LC and the NNC as well as the connectivity (in terms of $\mathbf{W}$ and $\mathbf{M}$) that supports the computation. In short, in the LC, LNs extract and encode the top K PCA subspace of the input in ORN somas and the ORNs → LN synaptic weight vectors $\{\mathbf{w}_k\}$ span that subspace. In the NNC, LNs encode soft cluster/feature memberships of the odor representations in ORN somas and $\{\mathbf{w}_k\}$ are related to cluster locations. In both models, the ORN axons encode a partially whitened and normalized version of the ORN soma activity due to LN feedback inhibition.

Predictions of the ORN–LN Connection Weight Vectors. We start by analyzing our models' predictions in terms of circuit connectivity. In the LC-K, the $\{\mathbf{w}_k\}_{k=1\ldots K}$ (proportional to the ORNs ↔ LN connection weight vectors) are linearly independent and span the same K dimensional subspace as the top K PCA directions $\{\mathbf{u}_{X,i}\}_{i=1\ldots K}$ of the uncentered input $\{\mathbf{x}^{(t)}\}$ (*SI Appendix*):

$$\mathbf{w}_k = \sum_{i=1}^{K} a_{k,i}\mathbf{u}_{X,i}. \qquad [1]$$

This ensures that LNs extract the top K PCA subspace of the input (below). The $\{a_{i,j}\}_{i,j=1\ldots K}$ are coefficients with a degree of freedom, arising from the nonuniqueness of the optimization solution. Thus, the $\mathbf{w}_k$s do not necessarily correspond to specific PCA directions of the input and are not orthogonal. Because the model predictions rely on "uncentered PCA," i.e., PCA without prior centering of the data, we use such PCA throughout the paper.

We probe this structural prediction by testing the alignment between the four ORNs → LN synaptic count vectors, $\{\mathbf{w}_{\mathrm{LNtype}}\}$ and the first 5 PCA directions of the ORN activity data, $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ (Fig. 2 *E*, *F*, and *H*). We find that only $\mathbf{w}_{\mathrm{BT}}$ is significantly correlated with the first PCA direction. Because this is uncentered PCA, this direction closely resembles the mean activity direction. We compare with the top 5 (instead of 4, as the number of $\mathbf{w}_{\mathrm{LNtype}}$s) PCA directions to account for the potential discrepancy between this ORN activity dataset and the true ORN activity.

Next, to test Eq. **1** directly, we examine the alignment of the subspaces spanned by the four $\mathbf{w}_{\mathrm{LNtype}}$s and the top five PCA directions of $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ (*SI Appendix*, Fig. S7). While ≈ 1 more dimension is significantly aligned than is randomly expected, supporting the results of Fig. 2*H*, there is no complete alignment. In summary, although $\mathbf{w}_{\mathrm{BT}}$ aligns with the top PCA direction of $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$, and the connectivity and activity subspaces are more aligned than expected by chance, the LC does not account for the connectivity of most LN types.

Next, we study the $\{\mathbf{w}_k\}_{k=1\ldots 4}$ predicted by the NNC-4 ($K = 4$ as the number of LN types) optimized on $\{\mathbf{x}^{(t)}\}_{\mathrm{data}}$ (Fig. 2*A*), for $0.1 \leq \rho \leq 10$. For $\rho \lessapprox 3.1$, three of the four $\mathbf{w}_k$s align

significantly with a $\mathbf{w}_{\text{LNtype}}$ (BT, BD, and P0, Fig. 3 *A* and *B*). In a perfect fit between model and data, each $\mathbf{w}_{\text{LNtype}}$ is aligned one $\mathbf{w}_k$. $\mathbf{w}_{\text{KS}}$ is not significantly correlated with any of the $\mathbf{w}_k$s, but NNC-5 has one $\mathbf{w}_k$ significantly aligned with $\mathbf{w}_{\text{KS}}$ (*SI Appendix*, Fig. S6*H*). The significant alignment of $\mathbf{w}_4$ with both $\mathbf{w}_{\text{BT}}$ and $\mathbf{w}_{\text{P0}}$ could arise due to partial correlation between $\mathbf{w}_{\text{LNtype}}$s (Fig. 1*C*). Furthermore, we find a similarity between the model and the data in terms of alignment of the ORNs → LN connection weight vectors with the ORN activity vectors $\{\mathbf{x}^{(t)}\}_{\text{data}}$ (*SI Appendix*, Fig. S8).

In summary, the ORN → LN connection weights predicted by the NNC model strongly resemble the synaptic counts in $\{\mathbf{w}_{\text{LNtype}}\}$, but do not provide an exact one-to-one correspondence. This analysis confirms that all the $\mathbf{w}_{\text{LNtype}}$s are adapted to ORN activity patterns. It also corroborates the hypothesis that the similarity-matching principle and the optimization problem have explanatory power for the organization of the biological circuit. Later we discuss the potential reasons for the nonexact alignment between the model and the data.

Emergence of LN Groups in the NNC. In the connectome, LNs are grouped by type and several $\mathbf{w}_{\text{LN}}$s are similar (Figs. 1 *B* and *C* and 3*C*). Do LN groups naturally emerge in our models? In the LC, $\{\mathbf{w}_k\}_{k=1\ldots K}$ spans the top K-dimensional principal subspace of the input $\{\mathbf{x}^{(t)}\}$, resulting in distinct $\mathbf{w}_k$s and thus no LN group emerges.

In the NNC, however, we observe the formation of LN groups. For example, in NNC-8 (8 LNs as on each side of the larva) trained on $\{\mathbf{x}^{(t)}\}_{\text{data}}$, several $\mathbf{w}_k$s are similar, especially for smaller

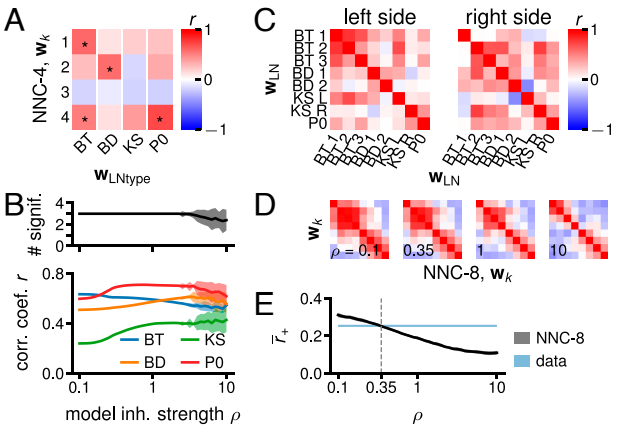

Fig. 3. Prediction of the connectivity with the NNC and emergence of LN groups. (*A*) Correlation between the four ORNs → LN connection weight vectors $\{\mathbf{w}_k\}$ from NNC-4 ($\rho = 1$) and the four ORNs → LN_{type} synaptic count vectors $\{\mathbf{w}_{\text{LNtype}}\}$ (*SI Appendix*, Fig. S6 *C*, *D*, *F*, *G*, and *H*). One-sided *P*-values calculated by shuffling the entries of each $\mathbf{w}_{\text{LNtype}}$ (50,000 permutations). *: significant at 0.05 FDR. (*B*) *Bottom*: maximum correlation coefficient (mean ± SD) of the four $\mathbf{w}_k$s from NNC-4 with the four $\mathbf{w}_{\text{LNtype}}$s for different values of ρ (50 simulations per ρ), encoding the feedback inhibition strength. *Top*: number of $\mathbf{w}_{\text{LNtype}}$s significantly correlated with at last one $\mathbf{w}_k$ from NNC-4 (FDR at 5%). For $\rho \gtrapprox 3.1$, not all simulations converge to the same $\{\mathbf{y}^{(t)}\}$, $\{\mathbf{z}^{(t)}\}$, and $\{\mathbf{w}_k\}$, potentially due to existence of multiple global optima or simulations only finding local optima. (*C*) Correlation between the $\mathbf{w}_{\text{LN}}$s on the left and right sides of the larva, portraying that several $\mathbf{w}_{\text{LN}}$s are similar. (*D*) Same as (*C*) for the eight $\mathbf{w}_k$s arising from NNC-8 and with $\rho = 0.1, 0.35, 1, 10$. Matrices ordered using hierarchical clustering and $\mathbf{w}_k$s ordered accordingly (*SI Appendix*). (*E*) Mean rectified correlation coefficient $\bar{r}_+$ ($r_+ := \max[0, r]$) from (*C*) (blue band delimited by the value for left and right circuit) and from NNC-8 (black line, mean ± SD, 50 simulations per ρ). $\bar{r}_+$ obtained by averaging all the r_+ from a correlation matrix, i.e., (*C*) or (*D*), excluding the diagonal.

ρ (Fig. 3*D*). Given that the $\mathbf{w}_k$s point toward the cluster locations in the ORN axon activity space, the grouping of $\mathbf{w}_k$s is influenced by 1) ORN activity pattern statistics (closer clusters elicit more aligned $\mathbf{w}_k$s), 2) the number of LNs (having more LNs than clusters lead to several similar $\mathbf{w}_k$s), and 3) the value of ρ (higher ρ leads to more separated clusters in ORN axons and thus dissimilar $\mathbf{w}_k$s) (*SI Appendix*, Figs. S9 and S10).

For the biological circuit, we lack exact measures of the factors (e.g., $\{\mathbf{x}^{(t)}\}$ and ρ) that influence $\{\mathbf{w}_k\}$ grouping. Nevertheless, we inquire whether NNC-8 can, in principle, generate a $\mathbf{w}_k$ grouping similar to the biological circuit for different values of ρ. At $\rho = 0.35$, the mean rectified correlation coefficient $\bar{r}_+$ ($r_+ := \max[0, r]$) between all $\mathbf{w}_k$s of the NNC-8 matched that of the connectome (Fig. 3*E*). While this value of ρ, which corresponds to a relatively low feedback inhibition in the model, should not be interpreted as the "true" value in the actual biological circuit, it falls within the range found above ($\rho \lessapprox 3.1$).

In summary, within a reasonable parameter range, the NNC reproduces another property of the biological circuit: the emergence of LN groups.

Relation between LN–LN and Feedforward ORNs → LN Connection Weights. The ORN-LN circuit contains reciprocal inhibitory LN–LN connections (Fig. 4*A*) whose connectivity patterns and roles are not fully understood. In our models, these connections are symmetric: the synaptic weights $\text{LN}_i \to \text{LN}_j$ and $\text{LN}_j \to \text{LN}_i$ are equal. This is largely verified in the connectome, except for the P0, which inhibits the KSs, but is not strongly inhibited by them. Theoretical predictions of the LC-K model (with K LNs) state that the strength of LN–LN connections ($\mathbf{M} = \{m_{\text{LNi, LNj}}\}_{i,j=1\ldots K}$) and ORN–LN connections ($\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$) are related (*SI Appendix*):

$$\mathbf{M}^2 = \mathbf{M}^\top\mathbf{M} \propto \mathbf{W}^\top\mathbf{W} \Leftrightarrow \mathbf{M} \propto (\mathbf{W}^\top\mathbf{W})^{1/2}, \qquad [2]$$

where $\top$ is the matrix transpose. This relationship is exact for the LC and approximate for the NNC. The ith column of $\mathbf{M}$, $\mathbf{m}_i$, is the LNs → LN_i (and LN_i → LNs) synaptic weight vector. The ith column of $\mathbf{W}$, $\mathbf{w}_i$, is proportional to the ORNs → LN_i (and LN_i → ORNs) synaptic weight vector. From Eq. **2** follows that: 1) $\|\mathbf{w}_i\|/\|\mathbf{m}_i\| = \text{const}$, i.e., the ratio between the magnitude of the ORNs → LN and LNs → LN synaptic weight vectors is the same at each LN. The magnitude is a proxy for the total synaptic strength of a synaptic weight vector. 2) $\measuredangle(\mathbf{w}_i, \mathbf{w}_j) = \measuredangle(\mathbf{m}_i, \mathbf{m}_j)$, where $\measuredangle(\mathbf{a}, \mathbf{b})$ is the angle between two vectors $\mathbf{a}$ and $\mathbf{b}$. Thus 2 LNs with a similar (different) connectivity pattern with the ORNs have a similar (different) connectivity pattern with LNs.

We test whether Eq. **2** holds in the connectome (Fig. 4), and find a significant correlation ($r = 0.73$, $P = 0.006$) between the off-diagonal entries of matrices $\mathbf{M}$ and $(\mathbf{W}^\top\mathbf{W})^{1/2}$, suggesting a meticulous co-organization of the ORN–LN and LN–LN connections. We lack the values of the LN neural leaks, which correspond to the diagonal entries of $\mathbf{M}$ (Eqs. **6** and **7**).

In summary, the synaptic weight organization in the NNC model resembles that the connectome in several key ways: the synaptic counts $\mathbf{w}_{\text{LNtype}}$, the emergence of LN groups, and the relationship between ORNs → LN and LN–LN. The LC model, on the other hand, fails at explaining several of these structural features.

Circuit Model Computation and Coding Efficiency. We next explore the computations of the LC and NNC. In both models,

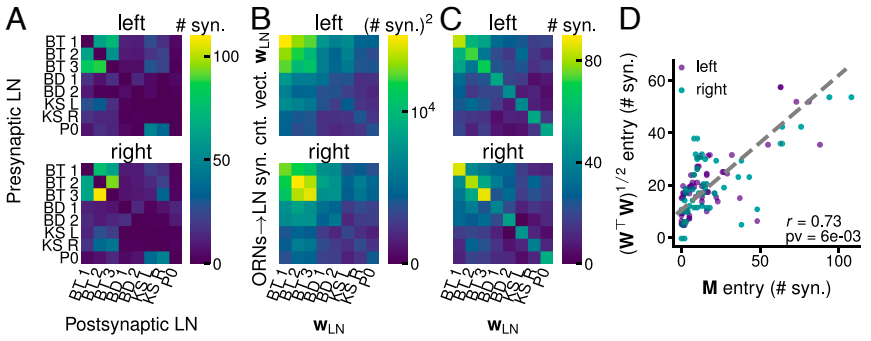

Fig. 4. Relationship between LN–LN (**M**) and ORNs → LN (**W**) synaptic counts in the connectome reconstruction. (*A*) LN–LN connections synaptic counts **M** on the left and right sides of the larva. (*B*) $\mathbf{W}^\mathsf{T}\mathbf{W}$ with $\mathbf{W} = [\mathbf{w}_{\mathrm{LN}1}, ..., \mathbf{w}_{\mathrm{LN}8}]$ on the left and right sides. Thus each entry is $\mathbf{w}_{\mathrm{LN}i}^\mathsf{T}\mathbf{w}_{\mathrm{LN}j}$, the scalar product between 2 ORNs → LN synaptic count vectors $\mathbf{w}_{\mathrm{LN}}$. (*C*) $(\mathbf{W}^\mathsf{T}\mathbf{W})^{1/2}$, i.e., the square root of the matrices in (*B*). (*D*) Entries of **M** vs entries of $(\mathbf{W}^\mathsf{T}\mathbf{W})^{1/2}$, excluding the diagonal, for both sides. *r*: Pearson correlation coefficient. pv: one-sided *P*-value calculated by shuffling the entries of each $\mathbf{w}_{\mathrm{LN}}$ independently, which assures that each LN keeps the same total number of synapses. Shuffling the entries of **M** in addition to shuffling each $\mathbf{w}_{\mathrm{LN}}$ leads to *P*-value $< 10^{-4}$.

upon ORN soma activation, the computation is implemented dynamically through the ORN–LN loop and converges exponentially to a steady state (Eqs. **6** and **7**). Given inputs $\{\mathbf{x}^{(t)}\}$, the circuit's outputs are the converged representations in ORN axons, $\{\mathbf{y}^{(t)}\}$, and LNs, $\{\mathbf{z}^{(t)}\}$.

Efficient encoding of odor representations in ORN is crucial for downstream processing. Odor representations can be visualized as points in a neural space, where each axis is the activity of an ORN. We consider a circuit with just $D = 2$ ORNs and $K = 2$ LNs, and an artificial input dataset of two odors A and B (Fig. 5 *A* and *D*). Given $\mathbf{x}^A$ and $\mathbf{x}^B$ the representations of the two odors: the larger the angle $\measuredangle(\mathbf{x}^A, \mathbf{x}^B)$, the easier the two odors can be discriminated, and the more efficiently the space is utilized. We quantify the efficiency of the encoding by the coefficient of variation of the PCA variances, $\{\sigma_i^2\}$, of the representation: $\mathrm{CV}_\sigma = \mathrm{SD}[\{\sigma_i^2\}]/\mathrm{mean}[\{\sigma_i^2\}]$. If all the variances are equal ($\mathrm{CV}_\sigma = 0$), the representation is white, and the encoding space is efficiently used (38). A larger CV_σ indicates a less optimal space utilization. We study the PCA variances and "whiteness" of uncentered data because we assume downstream neurons experience uncentered activity. We further describe the computation in terms of the modification of the stimulus representations.

LC: Extraction of the Principal Subspace by LNs and Partial Equalization of PCA Variances in ORN Axons. We first describe the computation in the LC. Given activity patterns $\{\mathbf{x}^{(t)}\}$ in the D ORN somas, we call $\{\mathbf{u}_{X,i}\}$ and $\{\sigma_{X,i}^2\}$ $(i = 1, ..., D)$ the PCA directions and variances of the uncentered $\{\mathbf{x}^{(t)}\}$ (Fig. 5*D*). The activity of the K LNs, $\{\mathbf{z}^{(t)}\}$, encodes the top K PCA subspace of $\{\mathbf{x}^{(t)}\}$, i.e., spanned by $\{\mathbf{u}_{X,i}\}_{i\le K}$ (Fig. 5*B*). How exactly LNs encode the subspace is a degree of freedom of the optimization, and thus the activity of individual LNs does not necessarily align with the PCA directions of the input. When $K < D$, LNs perform a dimensionality reduction of the ORN soma activity.

LNs inhibit ORN axons, altering their odor representation $\{\mathbf{y}^{(t)}\}$ (Fig. 5*D*). However, the PCA directions $\{\mathbf{u}_{Y,i}\}$ of ORN axon activity remain the same as in ORN somas, i.e., $\{\mathbf{u}_{Y,i}\} = \{\mathbf{u}_{X,i}\}$. Thus, this transformation from soma to axons only

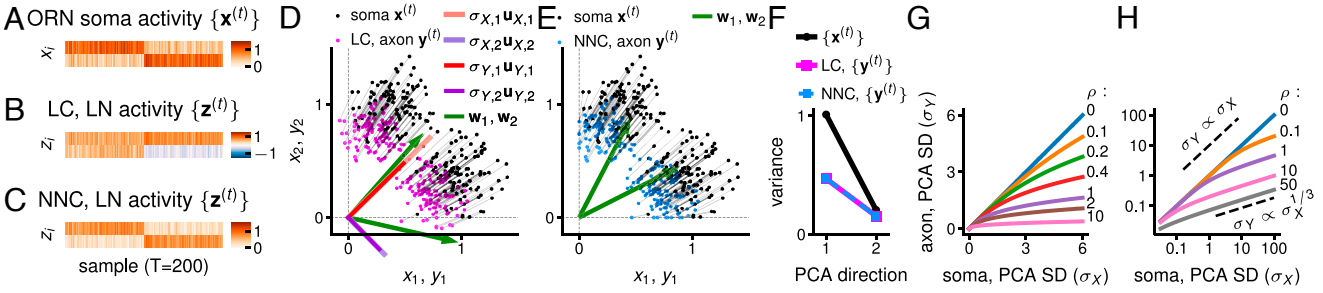

Fig. 5. Computation in the LC and NNC. (*A*) Artificial ORN soma activity patterns ($\{\mathbf{x}^{(t)}\}$, $D = 2$ ORN somas), generated with two Gaussian clusters of 100 points each centered at (1, 0.3) and (0.3, 1), SD = 0.17. This input is fed to the LC-2 (i.e., $K = 2$ LNs) (*B*, *D*, and *F*) and the NNC-2 (*C*, *E*, and *F*), $\rho = 1$. (*B*) LN activity, $\{\mathbf{z}^{(t)}\}$, in the LC-2. Because of a degree of freedom in LC, LN activity can be any rotation of the activity depicted here, i.e., $\mathbf{Q} \cdot \mathbf{z}$, where **Q** is a rotation (orthogonal) matrix. (*C*) LN activity, $\{\mathbf{z}^{(t)}\}$, in the NNC-2. LNs encode cluster memberships. (*D*) Scatter plot of the activity patterns in ORN somas ($\{\mathbf{x}^{(t)}\}$, black, from (*A*) and in ORN axons in the LC-2 ($\{\mathbf{y}^{(t)}\}$, magenta). $\sigma_{X,i}\mathbf{u}_{X,i}$, $\sigma_{Y,i}\mathbf{u}_{Y,i}$: vectors of the PCA directions of uncentered $\{\mathbf{x}^{(t)}\}$ and $\{\mathbf{y}^{(t)}\}$ scaled by the SD of that direction. $\mathbf{w}_k$ (green): direction of an ORNs → LN synaptic weight vector in the LC-2 from (*B*). Rotating the LN output $\{\mathbf{z}^{(t)}\}$ would alter the $\mathbf{w}_k$s, but not the $\{\mathbf{y}^{(t)}\}$. (*E*) Scatter plot of the activity patterns in ORN somas ($\{\mathbf{x}^{(t)}\}$, black, from (*A*) and in ORN axons in the NNC-2 ($\{\mathbf{y}^{(t)}\}$, blue). All activities are nonnegative and the $\mathbf{w}_k$s point toward the cluster locations, enabling the clustering observed in (*C*). (*F*) The PCA variances of the activity are less dispersed in ORN axons (output, $\{\mathbf{y}^{(t)}\}$) than in ORN somas (input, $\{\mathbf{x}^{(t)}\}$) for the LC and NNC. The output representation is thus partially whitened. The LC and NNC are similar in terms of their PCA variances. (*G* and *H*) Transformation of the SD (σ_X, σ_Y) of PCA directions from ORN somas ($\{\mathbf{x}^{(t)}\}$) to ORN axons ($\{\mathbf{y}^{(t)}\}$) in the LC model on linear and logarithmic scales, for different values of ρ (different line colors), encoding inhibition strength. When $\rho = 0$, the output equals the input. The higher the ρ, the smaller the PCA variances in the ORN axon.

stretches and does not rotate the cloud of representations in the neural space. This absence of rotation (called "zero-phase") makes the axonal and somatic activity maximally similar (23). This is advantageous for downstream processing because the evolving representation in ORN axons, computed dynamically via LN activation, is thus maximally close to the converged representation, allowing meaningful downstream processing before the complete representation convergence.

The PCA variances $\{\sigma^2_{Y,i}\}$ and $\{\sigma^2_{Z,i}\}$ of $\{\mathbf{y}^{(t)}\}$ and $\{\mathbf{z}^{(t)}\}$ are (Fig. 5 *D* and *F*):

$$\begin{cases} \sigma_{Y,i}\left(1+\rho^2\sigma^2_{Y,i}\right)=\sigma_{X,i} & 1\le i\le K \quad \textbf{[3a]} \\ \sigma_{Y,i}=\sigma_{X,i} & K+1\le i\le D \quad \textbf{[3b]} \\ \sigma_{Z,i}=\rho\sigma_{Y,i} & 1\le i\le K. \quad \textbf{[3c]} \end{cases}$$

Hence, the variances of the last D–K PCA directions in ORN somas ($\{\mathbf{x}^{(t)}\}$) remain unaltered in ORN axons ($\{\mathbf{y}^{(t)}\}$). The variances of top K PCA directions in ORN somas are diminished according to Eq. **3a** (Fig. 5 *G* and *H*): relatively large PCA variances in ORN somas ($\sigma^2_{X,i}\gg\rho^2$) are shrunken with a cubic root in ORN axons ($\sigma_{Y,i}\approx\sqrt[3]{\sigma_{X,i}/\rho^2}$), relatively small PCA variances ($\sigma^2_{X,i}\ll\rho^2$) remain virtually unchanged ($\sigma_{Y,i}\approx\sigma_{X,i}$). The PCA variances in LN activity ($\{\mathbf{z}^{(t)}\}$) are proportional to those in ORN axon activity ($\{\mathbf{y}^{(t)}\}$) (Eq. **3c**). (Note the indices i of the PCA directions and variances in ORN axons have been set to match those in ORN somas, and do not follow the usual decreasing order).

This transformation generally results in a smaller coefficient of variation of PCA variances, CV_σ, in the output $\{\mathbf{y}^{(t)}\}$ than in the input $\{\mathbf{x}^{(t)}\}$ (*SI Appendix*, see below, Fig. 6*D*). The PCA variances are then less spread and the odor representations are encoded more efficiently. Because the PCA variances are partially equated and no rotation occurs, this transformation is a partial (Zero-phase) ZCA-whitening.

NNC: Clustering by LNs and Partial Equalization of PCA Variances in ORN Axons. We next explore the computation of the NNC, where LN ($\{\mathbf{z}^{(t)}\}$) and ORN axon ($\{\mathbf{y}^{(t)}\}$) activities are nonnegative. LNs implement symmetric nonnegative matrix factorization (SNMF) on ORN axon activity, which consists of clustering and feature discovery (*SI Appendix*) (37). SNMF

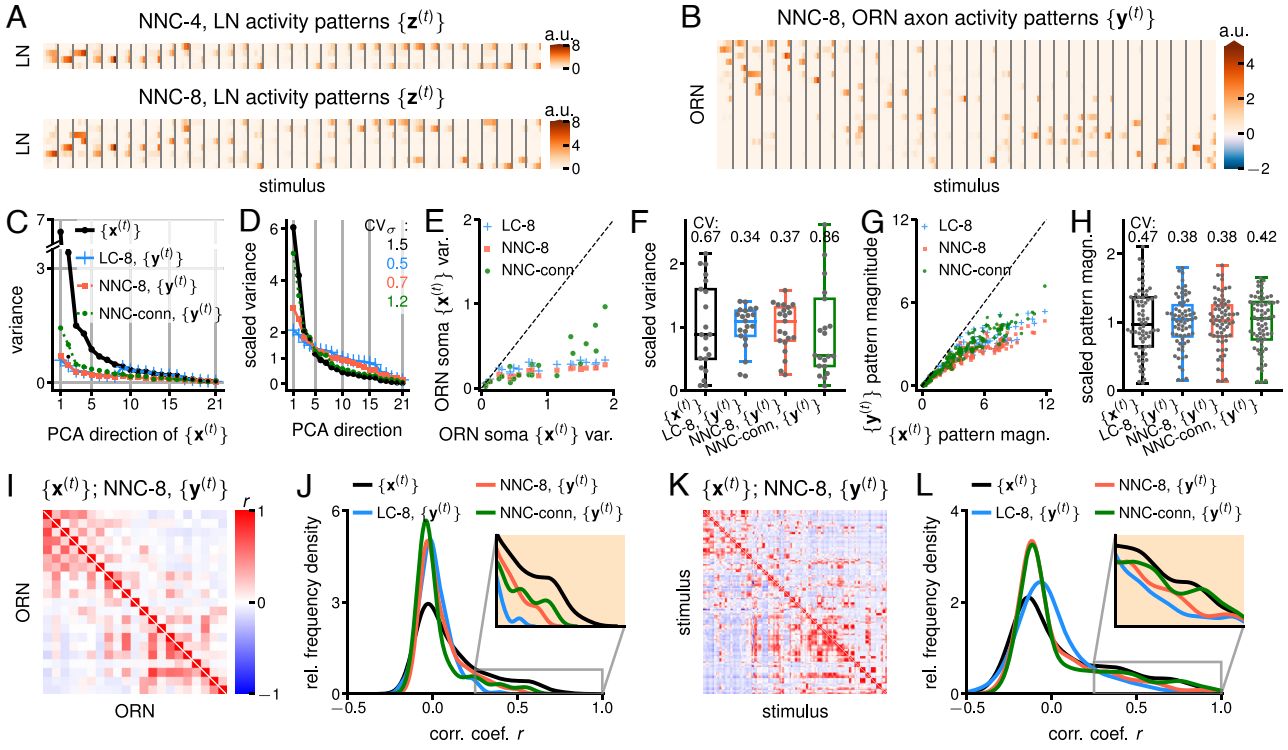

Fig. 6. Computation in the LC, NNC, and NNC-conn models in response to $\{\mathbf{x}^{(t)}\}_{\text{data}}$ (Fig. 2*A*): clustering, partial whitening, normalization, and decorrelation. (*A*) LN activity, $\{\mathbf{z}^{(t)}\}$, for the NNC-4 and NNC-8 models (*SI Appendix*, Fig. S11). LNs are mostly active in response to the odors to which their connectivity is the most aligned (*SI Appendix*, Fig. S8*A*). (*B*) ORN axon activity, $\{\mathbf{y}^{(t)}\}$, in the NNC-8. (*C*) Variances of odor representations in ORN somas ($\{\mathbf{x}^{(t)}\}_{\text{data}}$) and axons ($\{\mathbf{y}^{(t)}\}$) in the PCA directions of uncentered ($\{\mathbf{x}^{(t)}\}_{\text{data}}$). The variances decrease the strongest in the directions of the highest initial variance. (*D*) Uncentered PCA variances $\{\mathbf{x}^{(t)}\}_{\text{data}}$ and $\{\mathbf{y}^{(t)}\}$ scaled by their mean to portray the spread of variances. (*E*) Uncentered variances of activity at ORN axons ($\{\mathbf{y}^{(t)}\}$, output) vs. in ORN somas ($\{\mathbf{x}^{(t)}\}_{\text{data}}$, input). (*F*) Box plot of the ORN activity variances from (*E*) scaled by their mean to show the spread of variances. (*G*) Magnitude of the 170 activity patterns in ORN axons $\{\mathbf{y}^{(t)}\}$ vs in somas $\{\mathbf{x}^{(t)}\}_{\text{data}}$. (*H*) Box plot of the activity pattern magnitudes from (*G*) (only for top two dilutions 10^{-5} and 10^{-4}) scaled by their mean to show the spread of magnitudes. (*I*) Correlations between the activity of ORN somas ($\{\mathbf{x}^{(t)}\}_{\text{data}}$, *Lower Left* triangle) and of ORN axon activity in NNC-8 ($\{\mathbf{y}^{(t)}\}$, *Upper Right* triangle). (*J*) Smoothed histogram of the channel correlation coefficients from (*I*), excluding the diagonal (based on n=210 values). In all models, at the axonal level, there are more correlation coefficients around zero and fewer at higher values. (*K*) Correlations between activity patterns (i.e., odor representations) in ORN somas ($\{\mathbf{x}^{(t)}\}_{\text{data}}$, *Lower Left* triangle) and in ORN axons for NNC-8 ($\{\mathbf{y}^{(t)}\}$, *Upper Right* triangle). (*L*) Smoothed histogram of the activity pattern correlation coefficients from (*K*) (only for top two dilutions 10^{-5} and 10^{-4}, n = 2,278). Similar effect as for channels in (*J*). The decorrelation in the LC is more effective than in the NNC. The decorrelation in NNC-conn is not as pronounced as for the other two models. $\rho=2$ in this figure. a.u.: arbitrary units, stands for appropriate unit of neural activity. See *SI Appendix*, Figs. S12–S16 for the alignment of PCA direction, the LC, the NNC, the NNC-conn, and $\rho=10$.

is essentially "soft" K-means clustering, allowing inputs to belong to multiple clusters. Clustering satisfies the optimization's objective of nonnegative LN activity and maximally conserved distances between stimulus representations in ORN axons and LNs. Thus LN activity, $\{\mathbf{z}^{(t)}\}$, encodes the cluster membership of odor representations in ORN axons ($\{\mathbf{y}^{(t)}\}$), and the ORN → LN synaptic weight vectors, $\{\mathbf{w}_k\}$, point toward clusters (Fig. 5 *C* and *E*). Unlike the LC, there is no degree of freedom in LN activity.

The activity in ORN axons in NNC resembles that in LC, only without negative values, and the PCA variances are also similar (Figs. 5 *D*–*F*).

Circuit Model Computation on the ORN Activity Dataset. Next, to better comprehend the potential computation of the ORN-LN circuit, we study the computation of the NNC on the dataset of odor representation in ORNs, $\{\mathbf{x}^{(t)}\}_{\text{data}}$ (Fig. 2*A*). We also show the LC. We set the parameter that regulates the inhibition strength $\rho = 2$ to clearly represent the effect of the odor representation transformation in ORNs. K, the number of LNs, is set to 1, 4 (as the number of LN types) or 8 (as the number of LNs on one side of the larva). We also examine the computation of a nonnegative circuit model (NNC-conn) with connectivity weights proportional to the synaptic counts of the connectome (*SI Appendix*). Because for NNC-conn multiple unknown model parameters need to be guessed, and this circuit might not be adapted to the specific statistics of $\{\mathbf{x}^{(t)}\}_{\text{data}}$, its computation might not accurately reflect that of the true circuit, and the discrepancies with the normative models might be a consequence of this. Nevertheless, we find many similarities between NNC-conn and NNC-8, further supporting our predictions regarding circuit computation. Fig. 6 exhibits the main results, *SI Appendix*, Figs. S13, S14, and S15 display additional analysis of the LC, NNC, and NNC-conn, respectively.

As above, LNs in the LC encode the top K-dimensional PCA subspace of ORN soma activity (*SI Appendix*, Fig. S11*B*). LNs in the NNC softly cluster odors, as observed by their sparser activity and their correspondence with ORN activity patterns (Fig. 6*A*). LN activity in NNC-conn is also rather sparse.

In all models, ORN axon activity ($\{\mathbf{y}^{(t)}\}$) is weaker than in somas (Fig. 6*B*). While it is also sparser and nonnegative in the NNC models, in the LC, it contains negative values, which may not be biologically plausible.

Next, we compare the PCA variances of the odor representations in ORN somas ($\{\sigma^2_{X,i}\}$) and axons ($\{\sigma^2_{Y,i}\}$) (Fig. 6*C*). In the NNC models, variances decrease for all PCA directions. In the LC, however, only the variances of the top K PCA directions decrease. This difference results from the nonnegativity constraint in the NNC models, which affects all stimulus directions. The spread of PCA variances $\{\sigma^2_{Y,i}\}$ decreases in all models (smaller CV_σ, Fig. 6*D*) indicating a whiter representation in the ORN axons. This effect is the weakest in the NNC-conn. Changing the number of LNs impacts the NNC less than the LC. In the LC, only the order of the PCA directions of $\{\mathbf{x}^{(t)}\}$ and $\{\mathbf{y}^{(t)}\}$ changes, because K of them are shrunken (*SI Appendix*, Fig. S12 *A* and *B*). For the NNC, the PCA directions are slightly altered, but their order mostly remains (*SI Appendix*, Fig. S12 *C* and *D*). In the NNC-conn, the PCA directions are modified more strongly (*SI Appendix*, Fig. S12*E*).

Considering the decreased spread of PCA variances, we inquire whether activity becomes more evenly distributed among ORNs, an important property of efficient coding. Both the LC and NNC decrease the (uncentered) activity variance of "high-variance ORNs" and leave "low-variance ORNs" virtually unaffected, reducing the CV of ORN variance (Fig. 6 *E* and *F*). The NNC-conn, however, exhibits an increase in CV due to several "high-variance ORNs" being not strongly dampened.

Subsequently, we investigate changes in the magnitude of ORN soma and axon activity patterns. The magnitude is the length of an activity pattern vector in the $D = 21$ dimensional ORN space and is a proxy for the total activity of all ORNs in response to an odor. Similarly to ORN variances, the magnitude of large-magnitude patterns decreases, whereas small-magnitude patterns remain unchanged, decreasing the spread of pattern magnitudes (Fig. 6 *G* and *H*). These effects resemble a divisive normalization-type computation, also reported in *Drosophila* (13, 25).

In line with the less dispersed PCA variances in ORN axons, in all models ORNs and odor representations are more decorrelated in the axons than in the somas (Fig. 6 *I*–*L*), consistent with partial whitening.

Additionally, we investigate the effect of adjusting the model parameter ρ, which regulates feedback inhibition strength. A higher ρ ($\rho = 10$, *SI Appendix*, Fig. S16) leads to decreased activity in ORN axons and smaller PCA variances, reduced spread of PCA variances, channels and patterns norms, stronger decorrelation of ORNs and patterns. When inhibition is eliminated ($\rho \to 0$), the axonal and somatic ORN activity become identical. Although it is unknown if inhibition is modulated in the real circuit, altering this parameter allows us to understand this circuit's potential.

In summary, NNC analysis predicts that the ORN-LN circuit clusters odors with LNs and performs partial ZCA-whitening and normalization of odor representations in ORN axons. This results in a more efficiently encoded output with more decorrelated and equalized ORNs and odor representations, ultimately enhancing odor discrimination downstream.

Computation without LN–LN Connections. Lastly, we investigate the role of LN–LN connections by considering two alternative circuit models. First, we consider an LC or NNC circuit adapted to an input ensemble (i.e., Fig. 6) and remove the LN–LN connections, which corresponds to setting the off-diagonal elements in $\mathbf{M}$ to 0 (*SI Appendix*, Fig. S17). This manipulation leads to less sparse LN activity in the NNC, altered PCA directions in the axonal activity relatively to the soma, increased inhibition, and more dissimilar odor representations in ORN axons compared to somas. Thus, in an already "adapted" circuit, LN–LN connections improve clustering in LNs for the NNC, regulate inhibition, and maintain similar representation in ORN axons and somas.

Second, we consider the slightly different optimization problem that leads to an ORN-LN circuit without LN–LN connections (*SI Appendix*) (39). In the linear case, the whitening is complete (i.e., the first K PCA variances that are larger than $1/\rho^2$ become equal) and the K LNs still encode the top K dimensional subspace of the input. However, with nonnegativity constraints on ORN axon and LN activity, all LNs display the same activity, lacking differentiation (*SI Appendix*, Fig. S18). Thus, in this case, LN–LN connections are imperative for clustering.

Discussion

Combining the *Drosophila* larva olfactory circuit connectome, ORN activity data, and a normative model, we advance the

understanding of sensory computation and adaptation, quantitatively link ORN activity statistics, functional data, and connectome, and make testable predictions. We reveal a canonical circuit model capable of autonomously adapting to different environments, while maintaining the critical computations of partial whitening, normalization, and feature extraction. Such a circuit architecture may arise in other brain areas and may be applicable in machine learning and signal processing. Using ORN activity patterns as input, our normative framework accounts for the biological circuit structural organization and identifies in the connectome signatures of circuit function and adaptation to ORN activity. Such an approach offers a general framework to understand circuit computation (40, 41) and could provide valuable insights into more neural circuits, whose structural and activity data become available (1, 2).

Model and Biological Circuit: Similarities and Differences. In this paper, we compare the structural predictions of our normative approach to the connectome. The NNC model, when adapted to the ORN activity dataset (5), accounts for key structural characteristics (Figs. 3 and 4), for example, the ORNs → LN connection weight vectors. We ask two questions: 1) Why does the strong resemblance between model and data arise, when the available odor dataset most probably imperfectly matches the true larva odor environment? 2) Why isn't the resemblance even greater, and could the imperfect fit suggest that the model inadequately explains the biological circuit?

For 1), a possibility is that generic correlations between ORNs arise in large enough ORN activity datasets, causing robust features in the model connectivity. These correlations could result from the intrinsic chemical properties of ORN receptors. Odor statistics would also influence the connection weights, but to a lesser degree. Thus, a more naturalist activity dataset could further improve model predictions.

For 2), first, due to intrinsic noise and variability, no model could be 100% accurate in predicting connectivity. In fact, variability in synaptic count and innervation arises for *Drosophilas* raised in similar environments (27, 42), indicating potential "imprecision" of development and/or learning. We also observe variability in the left vs. right side connectivity (Fig. 1*B*). Second, incomplete ORN activity statistics may decrease prediction accuracy. Third, synaptic count might not exactly reflect synaptic strength (11). Finally, our model being a simplification of reality misses additional factors shaping circuit connectivity.

Our analysis indicates that the matches between model and data likely do not result from chance only, suggesting that the similarity-matching principle influences circuit organization. However, our unsupervised approach assumes that no odor is "special" for the animal, and thus LNs in the circuit model cluster odors solely based on their representations in the ORN activity space. This contrasts with the biological ORN-LN circuit, where LNs such as Keystone and Picky 0 have specific downstream connections likely related to survival needs and different hardwired animal behaviors (4, 43), requiring them to detect particular odors. Consequently, the connectivity of such LNs might contribute to the imperfect one-to-one correspondence between the model and the connectome (e.g., KS in NNC-4, Fig. 3*A*).

The circuit model can learn the optimal connection weights autonomously via Hebbian learning, offering the capacity to adapt to different environments. Studies in adult *Drosophila* reveal that glomeruli sizes (and thus ORN–LN or ORN–PN synaptic weights) or activity depend on the environment in which the *Drosophila* grew up (16–19). It is, however, unknown if activity-dependent plasticity also occurs in the larval ORN-LN circuit and whether the observed synaptic counts are a result of such plasticity. If present, it is unclear whether the short 6-h life of the larva from which the connectome was reconstructed allows substantial learning to occur and whether changes in synaptic weights would translate to different synaptic counts (11).

Resolving connectomes of larvae raised in different odor environments and at different times of their life, probing synaptic plasticity, and recording ORN responses to the full odor ensemble present in its environment would help clarify the influence of noise, plasticity, and genetics in circuit shaping.

Roles of LNs. LNs form a significant part of the neural populations in the brain, perform diverse computational functions, and exhibit extremely varied morphologies and excitabilities (27, 44). We propose a dual role for LNs in this olfactory circuit: altering the odor representation in ORNs and extracting ORN activity features, available for downstream use (4). In the olfactory system of *Drosophila* and zebrafish, LNs perform multiple computations, such as gain control, normalization of odor representations, and pattern and channel decorrelation (12–15, 32, 45), which is consistent with our results. Also, in *Drosophila* the LN population expands the temporal bandwidth of synaptic transmission and temporally tunes PN responses (28, 29, 46), which was not addressed here.

In topographically organized circuits, such as in the visual periphery or in the auditory cortex, distinct LN types uniformly tile the topographic space, and each LN type extracts a specific feature of the input, e.g., in the retina (47). In nontopographically organized networks, however, the organization and role of LNs remains a matter of research and controversy (27, 48). We study a subcircuit with four LN types, and most types contain several similarly connected LNs (Fig. 1). What is the function of multiple similar LNs in the ORN-LN circuit, as also observed in the NNC (Fig. 3 *C–E*)? First, LNs might differentiate further as the larva grows. Second, several LNs might help expand the dynamic range of a single LN. What are the features extracted by LNs in the *Drosophila* larva? Our NNC model and the distinct connectivity patterns of LN types in the connectome (4), suggest that different LN types are activated in response to different sets of odors. The extracted features might relate to clusters in ORN activity and to prewired, animal-relevant odors. Since several ORNs → LN connection weight vectors $\{\mathbf{w}_k\}$ in the NNC model resemble those in the biological circuit, the odor clusters identified by the model likely correspond to the set of odors that activate LNs in the biological circuit. The feedforward synaptic count vector from ORNs to the Broad Trio $\mathbf{w}_{\mathrm{BT}}$, which aligns with the first PCA direction of ORN activity and with an ORNs → LN connection weight vector $\mathbf{w}_k$ in the NNC model (Figs. 2*H*, 3 *A* and *B*) could potentially encode the mean ORN activity and thus be related to the global odor concentration (26). Other LNs might encode features of odors, such as aromatic vs. long-chain alcohols (5), or specific information influencing larva behavior (4, 43), but more experiments are required to definitely resolve the features. While our conclusions differ from a study that found that LN activation is invariant to odor identity (48), that study imaged several LNs simultaneously and might thus have missed the selectivity of individual LNs.

The connectome reveals LN–LN connections, which we propose play a key role in clustering and shaping the odor representation, and are co-organized with the ORN–LN connections (Fig. 4). To the best of our knowledge, the role of LN–LN

connections and their relationship to ORN–LN connections is relatively unexplored.

In summary, our study emphasizes the importance 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.

Circuit Computation, Partial ZCA-Whitening, and Divisive Normalization. We propose that the circuit's effect on the neural representation of odors in ORNs corresponds to partial ZCA-whitening and divisive normalization (Figs. 5 and 6). Such computations, which reduce correlations originating from the sensory system and the environment, have appeared in efficient coding and redundancy reduction theories (22, 25, 36, 38, 49, 50). Partial whitening is in fact a solution to mutual information maximization in the presence of input noise (38). In this circuit too, complete whitening might also not be desirable due to potential noise amplification. Thus, keeping low-variance signal directions of the input unchanged and dampening larger ones is consistent with mutual information maximization. Our conclusions are in line with reports of pattern decorrelation and/or whitening in the olfactory system in zebrafish (14, 15, 32, 33) and mice (34, 35).

The computation in our model also resembles divisive normalization, an ubiquitous computation in the brain (25), proposed for the analogous circuit in the adult *Drosophila* (12, 13). In its simplest form, divisive normalization is defined as $Y_j = \alpha X_j^n/(\sigma^n + \sum_k X_k^n)$, where Y_j is the response of neuron j, X_i is the driving input of neuron i, α is the maximum response of the output neuron and σ and n determine the offset and slope of the neuronal sigmoidal response curve, respectively (25). Divisive normalization captures two effects of neuronal and circuit computation: 1) neural response saturation with increasing input up to a maximum spiking rate α, arising from the neuron's biophysical properties; 2) dampening of the response of a given neuron when other neurons also receive input, often due to lateral inhibition (but see ref. 51). Aspect (1) is absent in our model but could be implemented with a saturating nonlinearity. Depending on the biological value of the maximum output, our model might not accurately capture responses for high-magnitude inputs. However, signatures of (2) are evident in the saturation of the activity pattern magnitudes in ORN axons for increasing ORN soma activity pattern magnitudes (Fig. 6*G*). Activity patterns of large magnitude correspond to activity at higher odor concentrations and with a high number of active ORNs. Because such input directions are more statistically significant in our dataset, these stimuli are more strongly dampened by LNs (which encode such directions) than those with few ORNs active. Thus, our model presents a possible linear implementation of a crucial aspect of divisive normalization, which in itself is a nonlinear operation.

Although the basic form of divisive normalization performs channel decorrelation, and not activity pattern decorrelation (13, 14, 32), our models perform both channel and pattern decorrelation. Nevertheless, a modified version of divisive normalization, which includes different coefficients for the driving inputs in the denominator (52), performs pattern decorrelation too, as our circuit model. The proposed neural implementations of divisive normalization usually require multiplication by the feedback (52, 53), 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 normalization, or learn sparse representations in an unsupervised manner (21, 37, 52, 54–59). However, these studies either lack a normative/optimization approach or have a different circuit architecture or synaptic learning rules. Using a normative approach has the advantage of directly investigating the underlying principles of neural functioning and also potentially providing a mathematically tractable understanding of the circuit structure and function.

Our study complements machine learning approaches to understand neural circuit organization (60, 61). These approaches use supervised learning and backpropagation to train an artificial neural network to perform tasks such as odor or visual classification. In the olfactory system, circuit configurations arising from this optimization, which could mimic the evolutionary process, display many connectivity features found in biology (61). Unlike these approaches, we propose a general principle governing the transformation of neural representations, similarity-matching, and also a mechanism to learn autonomously during the animal's lifetime.

Materials and Methods

Optimization Problems Describing the ORN-LN Circuit. We use a normative approach to study the ORN-LN circuit. We formulate two optimization problems that can be solved by a circuit model with the ORN-LN architecture. Studying the circuit model computation is then equivalent to studying the solution of an optimization problem. We derive analytical expressions describing different aspects of the computation and the circuit synaptic organization (*SI Appendix*).

We define the following variables: an input matrix $\mathbf{X} = [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(T)}]$ of T samples, and outputs $\mathbf{Y} = [\mathbf{y}^{(1)}, ..., \mathbf{y}^{(T)}]$, $\mathbf{Z} = [\mathbf{z}^{(1)}, ..., \mathbf{z}^{(T)}]$. $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ are D-dimensional vectors, while $\mathbf{z}^{(t)}$ are K-dimensional. $\mathbf{x}^{(t)}$, $\mathbf{y}^{(t)}$, and $\mathbf{z}^{(t)}$ represent the activity patterns of D ORN somas (i.e., the inputs), D ORN axons and K LNs, respectively. We call b^* an optimal value (solution) of a variable b. In the results section, we drop the *. We postulate the following similarity-matching-inspired optimization problem (e.g., ref. 20), which seeks the optimal output activities $\mathbf{Y}^*$ and $\mathbf{Z}^*$ given an input $\mathbf{X}$:

$$\min_{\mathbf{Y}} \max_{\mathbf{Z}} \frac{T}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 - \frac{\rho^2}{4}\left\|\mathbf{Y}^\mathsf{T}\mathbf{Y} - \frac{1}{\rho^2}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right\|_F^2 + \frac{\rho^2}{4}\left\|\mathbf{Y}^\mathsf{T}\mathbf{Y}\right\|_F^2, \quad [4]$$

where $\|\cdot\|_F^2$ is the square of the matrix Frobenius (Euclidean) norm. The term $\|\mathbf{X}-\mathbf{Y}\|_F^2$ drives the activity of the ORN axons $\mathbf{Y}$ toward the activity of ORN somas $\mathbf{X}$ and ensures that $\mathbf{Y}^* = \mathbf{X}$ when there is no activity in the LNs. The terms $\|\mathbf{Y}^\mathsf{T}\mathbf{Y} - 1/\rho^2\mathbf{Z}^\mathsf{T}\mathbf{Z}\|_F^2$ and $\|\mathbf{Y}^\mathsf{T}\mathbf{Y}\|_F^2$ align the similarities between the activities of ORN axons and LNs and puts a 4th order penalty on the norm of $\mathbf{Y}$; they correspond to the bidirectional all-to-all connectivity between ORN axons and LNs, as well as between LNs, but no direct connectivity between ORN axons; such similarity-matching terms permit a significant change of neural representation and a change of dimensionality, which takes place between ORN axons and LNs. ρ is a parameter related to the strength of the dampening in $\mathbf{Y}$ and affects both the optima $\mathbf{Y}^*$ and $\mathbf{Z}^*$.

We consider this optimization in two search domains for $\mathbf{Y}$ and $\mathbf{Z}$. One without any constraints on $\mathbf{Y}$ and $\mathbf{Z}$, representing the linear circuit (LC) model, and one with nonnegativity constraints ($\mathbf{Y} \geq 0$, $\mathbf{Z} \geq 0$), representing the nonnegative circuit (NNC) model. Nonnegativity constraints account for the fact that neural activity is usually nonnegative, or at least not symmetric in the negative and positive directions. The optimal $\mathbf{Y}^*$ and $\mathbf{Z}^*$ can be found analytically for the LC, and through numerical simulations for the NNC. Note that one cannot always guarantee converging to a global optimum for the NNC (62).

We prove that a neural circuit with ORN-LN architecture can solve this optimization problem (*SI Appendix*, Online algorithm). In brief, we introduce into the optimization problem two auxiliary matrices $\mathbf{W} := \mathbf{Y}\mathbf{Z}^\mathsf{T}/T$ and $\mathbf{M} := \mathbf{Z}\mathbf{Z}^\mathsf{T}/T$, which naturally map onto ORNs–LNs and LNs–LNs synaptic

weights, respectively. By construction, $\mathbf{M}$ is symmetric, i.e., $\mathbf{M} = \mathbf{M}^\mathsf{T}$. The new objective function is then optimized over the variables $\{\mathbf{y}^{(t)}\}$, $\{\mathbf{z}^{(t)}\}$, $\mathbf{W}$, and $\mathbf{M}$. Writing the gradient descent/ascent over $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ provides the neural dynamics equations, with $\mathbf{W}$ and $\mathbf{M}$ related to the synaptic weights (Eqs. **6** and **7**). The optimal $\mathbf{W}^*$ and $\mathbf{M}^*$:

$$\mathbf{W}^* = \mathbf{Y}^*\mathbf{Z}^{*\mathsf{T}}/T, \quad \mathbf{M}^* = \mathbf{Z}^*\mathbf{Z}^{*\mathsf{T}}/T, \qquad [5]$$

can be found "offline" by obtaining the optimal $\mathbf{Y}^*$ and $\mathbf{Z}^*$ in Eq. **4**, or in the "online setting," through unsupervised, Hebbian learning, where $\mathbf{W}$ and $\mathbf{M}$ are updated after each stimulus presentation (Eq. **8**, see below).

Circuit Neural Dynamics. A solution to the optimization problem Eq. **4** without the nonnegativity constraints can be implemented by the following differential equations describing the LC, whose steady-state solutions correspond to the optima for $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ for given $\mathbf{M}$ and $\mathbf{W}$ (*SI Appendix*, Online algorithm). These equations naturally map onto the ORN-LN neural circuit dynamics (dropping the sample index (t) for simplicity of notation):

$$\begin{cases} \tau_y d\mathbf{y}(\tau)/d\tau = -\mathbf{y}(\tau) - \mathbf{W}\mathbf{z}(\tau) + \mathbf{x} \\ \tau_z d\mathbf{z}(\tau)/d\tau = -\mathbf{M}\mathbf{z}(\tau) + \rho^2\mathbf{W}^\mathsf{T}\mathbf{y}(\tau), \end{cases} \qquad [6]$$

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 $\rho^2\mathbf{W}$ and $\mathbf{W}$ contain the synaptic weights of the feedforward ORNs $\rightarrow$ LN and feedback LN $\rightarrow$ ORNs connections, respectively. Thus, the feedforward connection vectors are proportional to the feedback vectors, and the parameter ρ sets the ratio. The assumption of proportionality is reasonable considering the connectivity data (*SI Appendix*, Fig. S2 *A*, *B*, and *D*). Off-diagonal elements of the $K \times K$ matrix $\mathbf{M}$ contain the weights of the LN - LN inhibitory connections, whereas the diagonal entries encode the LN leaks. In the absence of LN activity and at steady state, the equations satisfy $\mathbf{y} = \mathbf{x}$, i.e., somatic and axonal activities of ORNs are identical. In the absence of input ($\mathbf{x} = 0$) 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 through the term $-\mathbf{W}\mathbf{z}(\tau)$ and 2) LN activity is driven by the activity in ORN axonal terminals by $\rho^2\mathbf{W}^\mathsf{T}\mathbf{y}(\tau)$ and inhibited by LNs through the term $-\mathbf{M}\mathbf{z}(\tau)$. Note that changing ρ in the objective function leads to different optimal $\mathbf{W}^*$ and $\mathbf{M}^*$.

When optimized online, the optimization problem Eq. **4** with the nonnegativity constraints gives rise to the following equations describing the NNC:

$$\begin{cases} \mathbf{y}(\tau+1) = [\mathbf{y}(\tau) + \epsilon(\tau)\,(-\mathbf{y}(\tau) - \mathbf{W}\mathbf{z}(\tau) + \mathbf{x})]_+ \\ \mathbf{z}(\tau+1) = \left[\mathbf{z}(\tau) + \epsilon(\tau)\left(-\mathbf{M}\mathbf{z}(\tau) + \rho^2/\mathbf{W}^\mathsf{T}\mathbf{y}(\tau)\right)\right]_+, \end{cases} \qquad [7]$$

where $\epsilon(\tau)$ is the step size parameter and $[\mathbf{x}]_+ := \max[\mathbf{0}, \mathbf{x}]$ is a component-wise rectification. Here, τ is a discrete-time variable. These equations are analog to Eq. **6**, but also satisfying constraints on the activity: $y_i(\tau) \geq 0, z_i(\tau) \geq 0, \forall\tau, i$. Such constraints are implemented by formulating circuit dynamics in discrete time and using a projected gradient descent.

We call LC-K the linear circuit model implemented by Eq. **6** and NNC-K the nonnegative circuit model implemented by Eq. **7**, with K LNs.

Note that there is a manifold of implementations of the same computation by a circuit model. First, one can introduce a parameter γ (*SI Appendix*), that scales the feedforward and feedback connections as well as the magnitude of LN activity, in such a way that the ORN axon activity remains the same. Second, multiplying the whole equation in Eq. **6** or Eq. **7** would not alter the converged output, but would scale the circuit time constants and synaptic weights.

Synaptic Plasticity. The circuit model is capable of reaching the optimal synaptic weights $\mathbf{W}^*$ and $\mathbf{M}^*$, which solve the optimization problem Eq. **4**, in an unsupervised manner, with Hebbian plasticity. In practice, as the circuit receives a stimulus $\mathbf{x}^{(t)}$ (ORN soma activation), it performs a computation that yields a steady state output activity in ORN axons $\mathbf{y}^{(t)}$ and LNs $\mathbf{z}^{(t)}$ (with Eq. **6** or Eq. **7**); the synaptic weights are then updated using Hebbian rules:

$$\begin{aligned} \mathbf{W}^{(t+1)} &= \mathbf{W}^{(t)} + \epsilon_1(t)\left(\mathbf{y}^{(t)}\mathbf{z}^{(t)\mathsf{T}} - \mathbf{W}^{(t)}\right) \\ \mathbf{M}^{(t+1)} &= \mathbf{M}^{(t)} + \epsilon_2(t)\left(\mathbf{z}^{(t)}\mathbf{z}^{(t)\mathsf{T}} - \mathbf{M}^{(t)}\right), \end{aligned} \qquad [8]$$

where $\epsilon_i(t)$ are learning rates. These equations arise when optimizing Eq. **4** online. We assume that the ORN soma activation $\mathbf{x}^{(t)}$ is present long enough so that $\mathbf{y}^{(t)}(\tau)$ and $\mathbf{z}^{(t)}(\tau)$ reach steady state values. During this iterative process of synaptic updating, where the circuit model "learns"/"adapts" to the stimulus ensemble $\{\mathbf{x}^{(t)}\}$, the synaptic weights converge toward "optimum" steady state Eq. **5** (which might require multiple learning epochs over the $\{\mathbf{x}^{(t)}\}$). Note that the neural leaks of LNs (diagonal values of $\mathbf{M}$) are set (Eq. **5**) and updated (Eq. **8**) similarly to the synaptic weights ($\mathbf{W}$ and off-diagonal of $\mathbf{M}$).

Data, Materials, and Software Availability. The connectome and activity datasets are available in refs. (4) and (5). Code for generating the analysis and all the figures is available in GitHub (https://github.com/chapochn/ORN-LN_circuit) (63).

ACKNOWLEDGMENTS. We thank Aravinthan D.T. Samuel, Jacob Baron, Guangwei Si, Thomas Frank, Victor Minden, Anirvan Sengupta, Eftychios A. Pnevmatikakis, Siavash Golkar, David Lipshutz, and Shiva GhaaniFarashahi for discussions and/or comments on the manuscript. C.P. was supported by an NSF Award DMS-2134157 and the Intel Corporation through the Intel Neuromorphic Research Community.

Author affiliations: [a]Center for Computation Neuroscience, Flatiron Institute, New York, NY 10010; [b]Department of Neurology, New York University School of Medicine, New York, NY 10016; [c]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; [d]Center for Brain Science, Harvard University, Cambridge, MA 02138; [e]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA 02138; and [f]Neuroscience Institute, New York University School of Medicine, New York, NY 10016

1. K. Eichler *et al.*, The complete connectome of a learning and memory centre in an insect brain. *Nature* **548**, 175–182 (2017).
2. L. K. Scheffer *et al.*, A connectome and analysis of the adult *Drosophila* central brain. *eLife* **9**, e57443 (2020).
3. S. Aimon *et al.*, Fast near-whole-brain imaging in adult *Drosophila* during responses to stimuli and behavior. *PLOS Biol.* **17**, e2006732 (2019).
4. M. E. Berck *et al.*, The wiring diagram of a glomerular olfactory system. *eLife* **5**, e14859 (2016).
5. G. Si *et al.*, Structured odorant response patterns across a complete olfactory receptor neuron population. *Neuron* **101**, 950–962.e7 (2019).
6. R. I. Wilson, Early olfactory processing in drosophila: Mechanisms and principles. *Annu. Rev. Neurosci.* **36**, 217–241 (2013).
7. C. L. Barnes, D. Bonnéry, A. Cardona, "Synaptic counts approximate synaptic contact area in *Drosophila*" (Tech. Rep., bioRxiv, December 2020).
8. S. Takemura *et al.*, A visual motion detection circuit suggested by *Drosophila* connectomics. *Nature* **500**, 175–181 (2013).
9. N. Holderith *et al.*, Release probability of hippocampal glutamatergic terminals scales with the size of the active zone. *Nat. Neurosci.* **15**, 988–997 (2012).
10. Y. Akbergenova, K. L. Cunningham, Y. V. Zhang, S. Weiss, J. T. Littleton, Characterization of developmental and molecular factors underlying release heterogeneity at *Drosophila* synapses. *eLife* **7**, e38268 (2018).
11. C. H. Bailey, E. R. Kandel, K. M. Harris, Structural components of synaptic plasticity and memory consolidation. *Cold Spring Harbor Perspect. Biol.* **7**, a021758 (2015).
12. S. R. Olsen, R. I. Wilson, Lateral presynaptic inhibition mediates gain control in an olfactory circuit. *Nature* **452**, 956–960 (2008).
13. S. R. Olsen, V. Bhandawat, R. I. Wilson, Divisive normalization in olfactory population codes. *Neuron* **66**, 287–299 (2010).
14. A. A. Wanner, R. W. Friedrich, Whitening of odor representations by the wiring diagram of the olfactory bulb. *Nat. Neurosci.* **23**, 433–442 (2020).
15. R. W. Friedrich, Neuronal computations in the olfactory system of zebrafish. *Annu. Rev. Neurosci.* **36**, 383–402 (2013).

16. J.-M. Devaud, A. Acebes, A. Ferrús, Odor exposure causes central adaptation and morphological changes in selected olfactory glomeruli in *Drosophila*. *J. Neurosci.* **21**, 6274–6282 (2001).
17. I. P. Sudhakaran *et al*., Plasticity of recurrent inhibition in the *Drosophila* antennal lobe. *J. Neurosci.* **32**, 7225–7231 (2012).
18. S. Sachse *et al*., Activity-dependent plasticity in an olfactory circuit. *Neuron* **56**, 838–850 (2007).
19. S. Das *et al*., Plasticity of local GABAergic interneurons drives olfactory habituation. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E646–E654 (2011).
20. C. Pehlevan, A. Sengupta, D. B. Chklovskii, Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks? *Neural Comput.* **30**, 84–124 (2018).
21. C. Pehlevan, D. B. Chklovskii, "Optimization theory of Hebbian/anti-Hebbian networks for PCA and whitening" in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015* (2016), pp. 1458–1465.
22. H. B. Barlow, "Possible principles underlying the transformations of sensory messages" in *Sensory Communication*, W. A. Rosenblith, Ed. (The MIT Press, 1961), pp. 217–234.
23. A. Kessy, A. Lewin, K. Strimmer, Optimal whitening and decorrelation. *Am. Stat.* **72**, 309–314 (2018).
24. A. J. Bell, T. J. Sejnowski, The "independent components" of natural scenes are edge filters. *Vis. Res.* **37**, 3327–3338 (1997).
25. M. Carandini, D. J. Heeger, Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
26. K. Asahina, M. Louis, S. Piccinotti, L. B. Vosshall, A circuit supporting concentration-invariant odor perception in *Drosophila*. *J. Biol.* **8**, 9 (2009).
27. Y.-H. Chou *et al*., Diversity and wiring variability of olfactory local interneurons in the *Drosophila* antennal lobe. *Nat. Neurosci.* **13**, 439–449 (2010).
28. A. J. Kim, A. A. Lazar, Y. B. Slutskiy, Projection neurons in *Drosophila* antennal lobes signal the acceleration of odor concentrations. *eLife* **4**, e06651 (2015).
29. K. I. Nagel, E. J. Hong, R. I. Wilson, Synaptic and circuit mechanisms promoting broadband transmission of olfactory stimulus dynamics. *Nat. Neurosci.* **18**, 56–65 (2015).
30. G. Laurent, Olfactory network dynamics and the coding of multidimensional signals. *Nat. Rev. Neurosci.* **3**, 884–895 (2002).
31. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., B: Stat. Methodol.* **57**, 289–300 (1995).
32. R. W. Friedrich, M. T. Wiechert, Neuronal circuits and computations: Pattern decorrelation in the olfactory bulb. *FEBS Lett.* **588**, 2504–2513 (2014).
33. R. W. Friedrich, G. Laurent, Dynamic optimization of odor representations by slow temporal patterning of mitral cell activity. *Science* **291**, 889–894 (2001).
34. O. Gschwend *et al*., Neuronal pattern separation in the olfactory bulb improves odor discrimination learning. *Nat. Neurosci.* **18**, 1474–1482 (2015).
35. S. Giridhar, B. Doiron, N. N. Urban, Timescale-dependent shaping of correlation by olfactory bulb lateral inhibition. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5843–5848 (2011).
36. E. P. Simoncelli, B. A. Olshausen, Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
37. C. Pehlevan, D. B. Chklovskii, "A Hebbian/Anti-Hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features" in *Conference Record - Asilomar Conference on Signals, Systems and Computers* (2015), pp. 769–775.
38. J. J. Atick, A. N. Redlich, What does the retina know about natural scenes? *Neural Comput.* **4**, 196–210 (1992).
39. D. Lipshutz, C. Pehlevan, D.B Chklovskii, Interneurons accelerate learning dynamics in recurrent neural networks for statistical adaptation. arXiv [Preprint] (2022). http://arxiv.org/abs/2209.10634 (Accessed 3 October 2022).
40. Y. Bahroun, D. Chklovskii, A. Sengupta, A similarity-preserving network trained on transformed images recapitulates salient features of the fly motion detection circuit. *Adv. Neural Inf. Process. Syst.* **32** (2019).
41. S. Golkar, D. Lipshutz, Y. Bahroun, A. Sengupta, D. Chklovskii, A simple normative network approximates local non-Hebbian learning in the cortex. *Adv. Neural Inf. Process. Syst.* **33**, 7283–7295 (2020).
42. W. F. Tobin, R. I. Wilson, W.-C. A. Lee, Wiring variations that enable and constrain neural computation in a sensory microcircuit. *eLife* **6**, e24838 (2017).
43. K. Vogt *et al*., Internal state configures olfactory behavior and early sensory processing in *Drosophila* larvae. *Sci. Adv.* **7**, eabd6900 (2021).
44. R. Hattori, K. V. Kuchibhotla, R. C. Froemke, T. Komiyama, Functions and dysfunctions of neocortical inhibitory neuron subtypes. *Nat. Neurosci.* **20**, 1199–1208 (2017).
45. P. Zhu, T. Frank, R. W. Friedrich, Equalization of odor representations by a network of electrically coupled inhibitory interneurons. *Nat. Neurosci.* **16**, 1678–1686 (2013).
46. K. I. Nagel, R. I. Wilson, Mechanisms underlying population response dynamics in inhibitory interneurons of the *Drosophila* antennal lobe. *J. Neurosci.* **36**, 4325–4338 (2016).
47. R. H. Masland, The neuronal organization of the retina. *Neuron* **76**, 266–280 (2012).
48. E. J. Hong, R. I. Wilson, Simultaneous encoding of odors by channels with diverse sensitivity to inhibition. *Neuron* **85**, 573–589 (2015).
49. M. D. Plumbley, "A Hebbian/anti-Hebbian network which optimizes information capacity by orthonormalizing the principal subspace" in *Proceedings of IEE Conference on Artificial Neural Networks* (1993), pp. 86–90.
50. R. Linsker, Self-organization in a perceptual network. *Computer* **21**, 105–117 (1988).
51. T. K. Sato, B. Haider, M. Häusser, M. Carandini, An excitatory basis for divisive normalization in visual cortex. *Nat. Neurosci.* **19**, 568–570 (2016).
52. Z. M. Westrick, D. J. Heeger, M. S. Landy, Pattern adaptation and normalization reweighting. *J. Neurosci.* **36**, 9805–9816 (2016).
53. D. J. Heeger, Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
54. P. D. King, J. Zylberberg, M. R. DeWeese, Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *J. Neurosci.* **33**, 5475–5485 (2013).
55. M. Zhu, C. J. Rozell, Modeling inhibitory interneurons in efficient sensory coding models. *PLoS Comput. Biol.* **11**, e1004353 (2015).
56. B. A. Olshausen, D. J. Field, Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vis. Res.* **37**, 3311–3325 (1997).
57. A. A. Koulakov, D. Rinberg, Sparse incomplete representations: A potential role of olfactory granule cells. *Neuron* **72**, 124–136 (2011).
58. S. D. Wick, M. T. Wiechert, R. W. Friedrich, H. Riecke, Pattern orthogonalization via channel decorrelation by adaptive networks. *J. Comput. Neurosci.* **28**, 29–45 (2010).
59. J. J. Atick, A. N. Redlich, Convergent algorithm for sensory receptive field development. *Neural Comput.* **5**, 45–60 (1993).
60. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
61. P. Y. Wang, Y. Sun, R. Axel, L. F. Abbott, G. R. Yang, Evolving the olfactory system with machine learning. *Neuron* **109**, 3879–3892.e5 (2021).
62. A. M. Sengupta, M. Tepper, C. Pehlevan, A. Genkin, D. B. Chklovskii, "Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks" in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS 2018* (Curran Associates Inc., Red Hook, NY, 2018), pp. 7080–7090.
63. N. M. Chapochnikov, C. Pehlevan, D. B. Chklovskii, Python code for paper "Normative and mechanistic model of an adaptive circuit for efficient encoding and feature extraction". GitHub. https://github.com/chapochn/ORN-LN_circuit/. Deposited 28 June 2023.