Manifold learning for olfactory habituation to strongly fluctuating backgrounds

François X. P. Bourassa[a, b, c, ✉], **Paul François**[c, d], **Gautam Reddy**[a, ✉], **and Massimo Vergassola**[e, f]

[a]Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544, USA
[b]Department of Physics, McGill University, 3600 rue University, Montréal, QC, H3A 2T8, Canada
[c]Département de biochimie et médecine moléculaire, Université de Montréal, 5155 Chemin de la Rampe, Montréal, QC, H3T 1J4, Canada
[d]MILA Québec, 6666, rue Saint-Urbain, Montréal, QC, H2S 3H1, Canada
[e]Laboratoire de Physique de l'Ecole normale supérieure (ENS), Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France
[f]Department of Physics, University of California, San Diego, 9500 Gillman Drive, La Jolla, CA 92093, USA

Abstract

Animals rely on their sense of smell to survive, but important olfactory cues are mixed with confounding background odors that fluctuate due to atmospheric turbulence. It is unclear how the olfactory system habituates to such stochastic backgrounds to detect behaviorally important odors. Here, we explicitly consider the high-dimensional nature of odor coding, the natural statistics of odor fluctuations and the architecture of the early olfactory pathway. We show that their combination favors a manifold learning mechanism for olfactory habituation over alternatives based on predictive filtering. Manifold learning is implemented in our model by a biologically plausible network of inhibitory interneurons in the early olfactory pathway. We demonstrate that plasticity rules based on IBCM or online PCA are effective at implementing this mechanism in turbulent conditions and outperform previous models relying on mean background subtraction. Interneurons with an IBCM plasticity rule acquire selectivity to independently varying odors. This manifold learning mechanism offers a path towards distinguishing plasticity rules in experiments and could be leveraged by other biological circuits facing fluctuating environments.

olfaction | **fluctuating environments** | **habituation** | **manifold learning** | **theoretical neuroscience**

Correspondence: ***fb3@princeton.edu, greddy@princeton.edu***

Introduction

Most of us have experienced an odor fading to imperceptibility after prolonged exposure. Habituation is a basic building block of sensory cognition, allowing us to pay attention to weak but important cues relevant for survival [1, 2]. Across sensory modalities, numerous mechanisms for sensory adaptation and habituation filter out irrelevant information; these mechanisms must be considered in light of the statistical features of natural scenes [3–7]. Though olfactory habituation to regular stimuli is behaviorally well-characterized [8], less is known about its computational basis in neural circuits facing naturalistic environments.

The physics of odor transport poses a difficult habituation problem, challenging simple models such as mean background subtraction. Unlike in vision and audition, olfactory signals are transported by a physical medium that is turbulent at the spatial scales relevant for behavior. Wind velocities in such environments have complex spatial and temporal fluctuations, which segregate air into patches of odor and clean air (Fig. 1A). An olfactory sensory apparatus thus receives a highly intermittent sensory signal, where clumps of intense odor detections ('whiffs') are separated by seconds to minutes of relatively clean air ('blanks') (Fig. 1B) [9–11]. These strongly non-Gaussian statistics make it non-trivial for olfactory circuits to identify new odors mixed with a dominant, fluctuating background.

The neurobiology of early olfaction outlines the underlying circuit structure solving this habituation problem across animal species (Fig. 1C). In insects, odors are first detected by olfactory receptors (ORs) located on olfactory sensory neurons (OSNs) in the antennae. Each OSN often expresses one olfactory receptor type (among $\sim$ 50 different receptor types in the fruit fly). Axons from OSNs expressing the same receptor project to distinct locations called glomeruli in the antennal lobe. Projection neurons (PNs) integrate signals from a few glomeruli and project to higher order processing centers, including the mushroom body (where associations are formed) and the lateral horn (which drives innate behaviors) [12, 13]. A strikingly similar architecture is present in mammals: glomeruli are located in the olfactory bulb where OSN input is processed and broadcast to diverse subcortical and cortical regions [14].

A characteristic feature of ORs is their broad tuning to many odor molecules, such that each OSN is activated by multiple odors [15]. Thus, OSN activity induced by a behaviorally relevant odor appearing in a naturalistic environment is masked by possibly many background odors [16]. Biophysical mechanisms for adaptation in a single OSN allow for adapting the dynamic range of spiking output to the statistics of receptor activity [17–20]. However, these single-neuron mechanisms do not, on their own, disentangle contributions from a new and relevant odor from those of irrelevant backgrounds, suggesting that background subtraction occurs at the neural population coding level, in a later stage of the olfactory pathway [21–24].

The antennal lobe (AL) in insects and the olfactory bulb (OB) in mammals are likely candidate regions for olfactory habituation. Since most olfactory receptors are promiscuous, the population activity of glomeruli in these regions represents an efficient combinatorial code for odors [25, 26]. The AL and OB further contain extensive local networks of inhibitory interneurons that mediate inter-glomerular crosstalk before signals are broadcast to downstream processing centers [27–30]. Consistent with this picture, plasticity mechanisms in *Drosophila* antennal lobe inhibitory interneurons (called lateral neurons (LNs)) have been linked to the formation of olfactory memories during habituation [6, 8, 22, 31–35]. Building on these results, Shen *et al.* proposed a neural model for olfactory habituation where LNs learn and subtract a time-averaged background signal by integrating glomerular activity over timescales of minutes [36]. While a mean filtering mechanism is plausible when backgrounds are stable, it cannot filter out the strong fluctuations of naturalistic olfactory scenes (*Supp. Materials*, sec. 3), hinting that other mechanisms are at play.

Here, we propose a conceptually distinct model for olfactory habituation to broadly activating backgrounds that fluctuate on physically relevant timescales. As in previous proposals, we assume that background subtraction is mediated by plasticity mechanisms in inhibitory interneurons within the AL and OB. Our model is motivated by the fact that the representation of a fluctuating odor traverses a one-dimensional manifold (*i.e.*, a curve) in a much higher-dimensional glomerular activity space [23]. Consequently, a mixture of backgrounds spans a manifold of dimensionality equal to the number of independently varying odors in the mixture. A network that learns this low-dimensional manifold can thereby subtract out background activity by projecting instantaneous activity to the low-dimensional background manifold, highlighting components that are orthogonal to it (Fig. 1D). Our proposed 'manifold projection' mechanism thus relies on the high-dimensional nature of olfactory coding, and is still applicable when odors fluctuate on timescales comparable to timescales of neural signal propagation. Hence, the characteristics of olfactory stimuli and circuits outline a distinct habituation mechanism at the level of neural population codes, which could also be leveraged by other biological circuits facing fluctuating backgrounds in high-dimensional input spaces.

The structure of the paper is as follows. We first use a minimal mathematical model to delineate the physical and sensory coding regimes where a manifold projection strategy outcompetes a predictive filtering mechanism for new signal recognition among strong background fluctuations. Next, we propose a biologically plausible model for manifold learning implemented by inhibitory interneurons in the early olfactory system. We consider two local plasticity rules (IBCM and BioPCA), which find the linear subspace spanned by the background odors. Interneurons equipped with either rule perform considerably better against fluctuating backgrounds than prior models and perform comparably against each other. A detailed mathematical analysis of the IBCM rule shows that interneurons acquire selectivity to independent odors in the mixture. Finally, we show that both plasticity rules are robust across a range of physiologically relevant physical and computational regimes.

Results

Regimes of predictive filtering and manifold learning. We begin by delineating the physical and computational regimes in which a *manifold learning* strategy for habituation outperforms a *predictive filtering* strategy. We consider an olfactory system habituating for time T to a fluctuating background, $\mathbf{b}(t)$. The components of these vectors represent the glomerular activations corresponding to each OR type, and thus reflect the coordinates of an odor in an N_S-dimensional olfactory coding space. As in the rest of the paper, the backgrounds $\mathbf{b}(t)$ are mixtures of N_B odor vectors,

$$\mathbf{b}(t) = \sum_{\gamma=1}^{N_\mathrm{B}} c_\gamma(t)\hat{\mathbf{s}}_\gamma, \tag{1}$$

where the concentration of the γth odor at time t is $c_\gamma(t)$. Here, we have assumed additive odor mixtures to keep our analysis tractable; our conceptual argument should also extend to non-additive odor mixture coding [17, 37–39] when combined with algorithms for curved manifolds (*Discussion*). The system subsequently responds at time T to a mixture of the fluctuating background $\mathbf{b}(T)$ and a new, behaviorally relevant target odor $\mathbf{s}_\mathrm{new}$ which the target aims to recognize; the total input is $\mathbf{b}(T) + \mathbf{s}_\mathrm{new}$. An idealized circuit subtracts a vector $\mathbf{u}_T$ from the target-background mixture, where

$$\mathbf{u}_T = \sum_{j=1}^{T-1} v_j \mathbf{b}(T-j) + P(\mathbf{b}(T) + \mathbf{s}_\mathrm{new})\,. \tag{2}$$

The first term represents a weighted average over the background's history, corresponding to the predictive filtering strategy, where scalar coefficients v_j are set by the second-order statistics of background fluctuations. The second term corresponds to a simplified manifold learning strategy, where the matrix P projects the current stimulus to the subspace spanned by the background. The parameters v_j and P are learned during habituation such that the target odor $\mathbf{s}_\mathrm{new}$ is recovered (in the mean squared error sense) by subtracting $\mathbf{u}_T$ from the current input.

We analytically optimized v_j and P to delineate how the two strategies contribute to recovering the target odor $\mathbf{s}_\mathrm{new}$ from the mixture with $\mathbf{b}(T)$, which depends on background statistics (*Supp. Materials*, sec. 1). Fig. 1E illustrates the optimal reconstruction error for different autocorrelation timescales of background fluctuations (τ), and dimensionalities of the olfactory coding space (N_S, Fig. S1A-B). While the combined strategy is by construction always better than each individual strategy, manifold learning alone explains all the performance when fluctuations are fast and the dimensionality is large (small τ, large N_S).

The spectrum of turbulent fluctuations is dominated by brief whiffs and blanks that can reach down to ∼10 ms (Fig. 1B); predictive filtering, since it acts as a change detector, would constantly respond to these whiffs. Taking the filtering time step to be the smallest olfactory delay functionally perceptible to mice and human (30-60 ms) [40, 41], we estimate that $\tau < 100$ for typical turbulent backgrounds (Fig. S1C). The number of OR types spans from a few tens to thousands across different animals. The olfactory space is high-dimensional compared to the typical number of in-

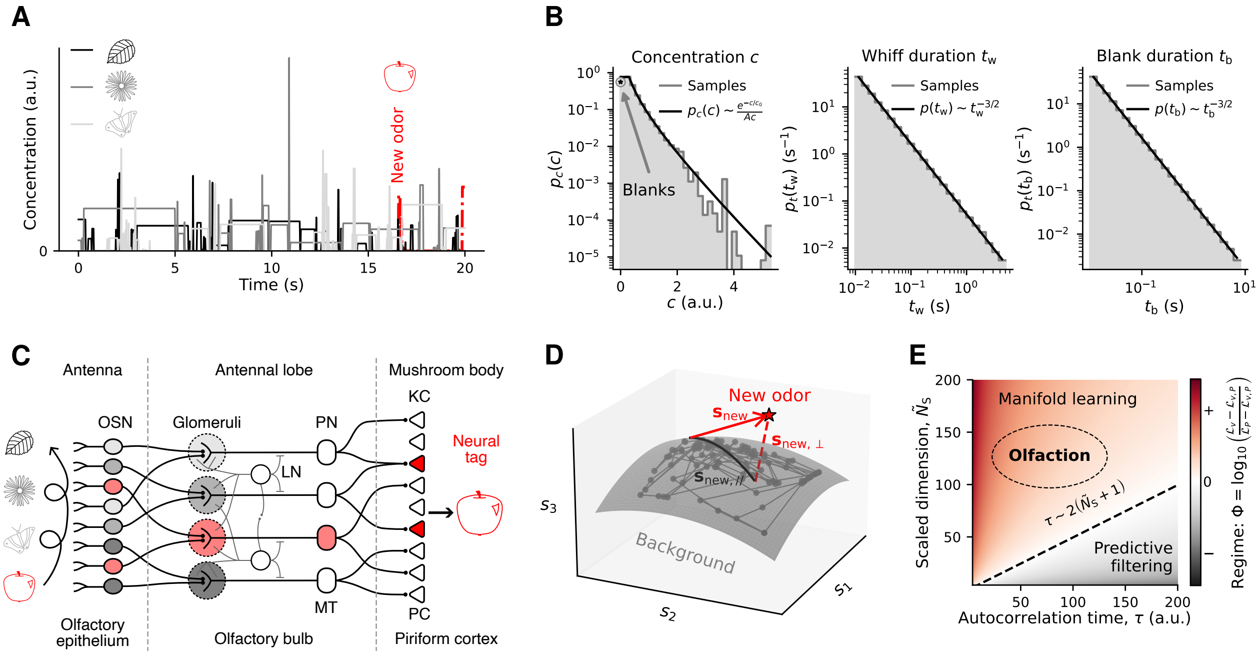


Fig. 1. Olfactory systems face stochastic, turbulent odor mixture inputs, for which manifold learning may be an optimal habituation strategy. (**A**) Illustration of background and new odor concentrations time series, strongly fluctuating in a series of whiffs and blanks, according to the turbulent atmosphere statistics derived in [9]. (**B**) Stationary probability distributions of whiff concentrations and whiff and blank durations. (**C**) Structure of early layers in the olfactory network, annotated with fly (top) and mouse (bottom, when different from fly) anatomical regions and cell types. (**D**) Illustration of a hypothetical low-dimensional subspace spanned by background odors in the space of OSN activities s_i, with sample mixtures generated by log-normal odor concentrations. A new odor, $\mathbf{s}_{\mathrm{new}}$, generally has a component, $\mathbf{s}_{\mathrm{new},\perp}$, lying outside of the background manifold. (**E**) Log-ratio of the difference in loss functions for new odor recognition between manifold learning ($\mathcal{L}_P$), predictive filtering ($\mathcal{L}_v$), or the combination of both strategies ($\mathcal{L}_{v,P}$). For a sensory system tasked to detect new odors within fast background fluctuations in a high-dimensional space, as it is the case with olfaction, manifold learning is the dominant strategy (loss $\mathcal{L}_P \approx \mathcal{L}_{v,P}$). The olfactory space dimensionality is rescaled by the relative variance of background and new odors: $\tilde{N}_{\mathrm{S}} = N_{\mathrm{S}}\sigma^2/\sigma^2_{\mathrm{new}}$.

dependent odor sources that might prevail in a natural landscape. Thus, physics and neurobiology together indicate that olfaction lies within the regime where a manifold learning strategy is most effective.

Models of manifold learning in the early olfactory circuit to improve new odor recognition. We now develop biologically plausible models of olfactory habituation that rely on manifold learning. Following [36, 42], we formulate a mathematical description (Fig. 2A) of the early olfactory circuit (Fig. 1C). We use *Drosophila* cell types for conciseness, but the model generalizes to other organisms. The key component for habituation is a layer of N_{I} lateral interneurons (LN) which receive inputs $\mathbf{s}$ from olfactory sensory neurons (OSN) via synaptic weights M and are coupled with lateral connections L, thus having activities $\bar{\mathbf{h}} = LM\mathbf{s}$. OSNs excite projection neurons (PNs) with unit synaptic weights and LNs inhibit PNs with synaptic weights W. The net activity of the PNs is thus $\mathbf{y} = \mathbf{s} - W\bar{\mathbf{h}}$.

Intuitively, in our model, interneurons learn to project inputs onto the low-dimensional subspace of background odors, and subtract these projections from PNs. Interneurons can therefore perform (linear) manifold learning with projection matrix WLM. Lastly, PN activities are projected on a large layer of N_{K} Kenyon cells (KC) by sparse random connections (fixed, not learned). The condition $N_{\mathrm{K}} \gg N_{\mathrm{S}}$ ensures that the 5 % most active KCs represent a distinct neural tag z for each possible odor. This dimensional expansion from PNs to KCs implements locality-sensitive hashing of input identity [42].

Next, we consider biologically realistic synaptic plasticity rules for weights M, L, and W to achieve adequate manifold learning within this network. The optimal manifold projection matrix P derived for Fig. 1E involves non-local terms and moments of the new odor distribution inaccessible to the network (eq. 9). Instead, we postulate a simple, unsupervised, local learning rule for the inhibitory weights W: they evolve during habituation to minimize the norm of the PN activity $\mathbf{y}$ by using interneuron activities $\bar{\mathbf{h}}$. This optimization principle results in simple Hebbian dynamics (see *Methods*),

$$\frac{\mathrm{d}W_{ij}}{\mathrm{d}t} = \alpha y_i \bar{h}_j - \beta W_{ij} \tag{3}$$

with learning rate α and a regularization rate β. This simple update rule allows us to compare different models for the projection weights M and L. We consider two models: (a) the Intrator, Bienenstock, Cooper, and Munro (IBCM) model of synaptic plasticity [43–46], and (b) a biologically plausible online implementation of principal components analysis (BioPCA) [47] (for full model definitions, see *Methods*). The IBCM model was proposed to explain neuronal selectivity to specific stimulus components [43]. Its connections with independent component analysis (ICA) [48, 49] suggest that it could provide a biologically meaningful basis for learning the

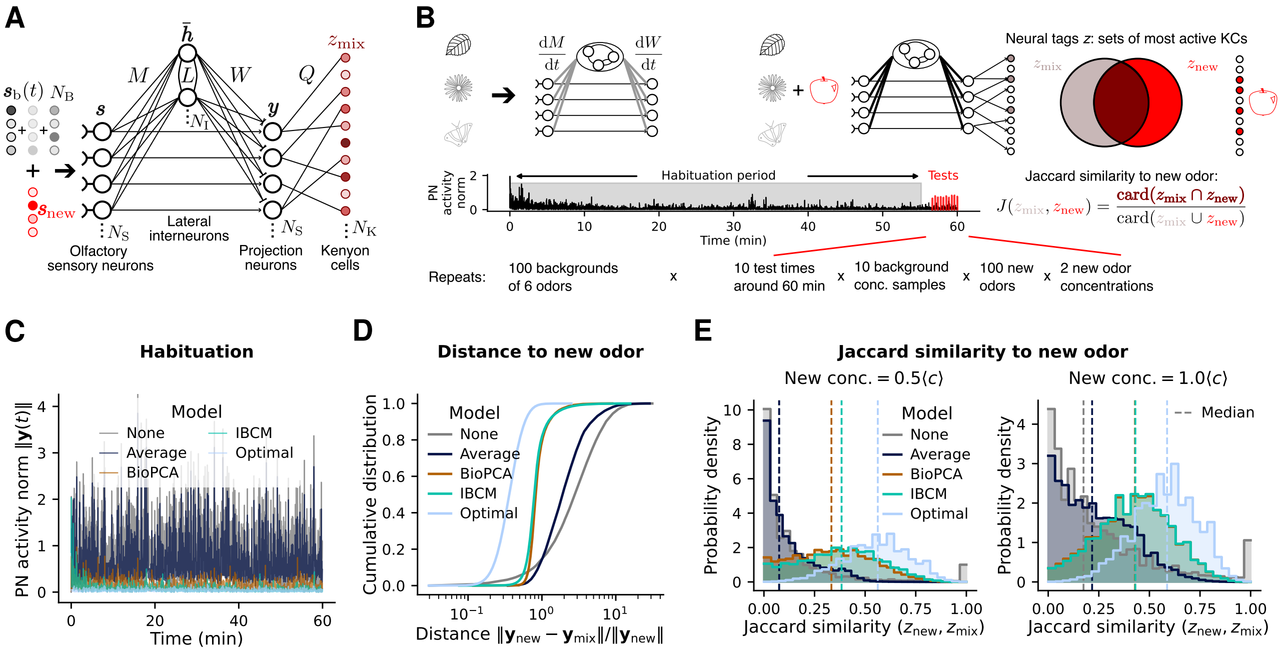


Fig. 2. Recognition of new odors after habituation to a background with different learning models. (**A**) Mathematical description of the olfactory network. (**B**) Schematic of the numerical experiments performed to assess habituation performance. We generate a set of background odors randomly, then integrate the network's synaptic plasticity equations for 60 minutes of habituation to a simulated background time series, where odor concentrations fluctuate according to the turbulent stochastic process illustrated in Fig. 1A-B (see *Supp. Materials* sec. 2 for simulation methods). We then compute the network's response to a new odor $\mathbf{s}_{\text{new}}$ mixed with the background. The recognition performance is quantified by the Jaccard similarity between the neural tag of the mixture, z_{mix}, and the neural tag (pre-habituation) of the new odor alone, z_{new}. This procedure is repeated for several test times, random backgrounds, samples of each background, random new odors, and different new odor concentrations, for each habituation model (none, average subtraction, BioPCA, IBCM). Parameter values are listed in the *Methods*. (**C**) Sample time series of the norm of PN activity, to illustrate the extent of habituation (decrease in PN activity when exposed to the background) in each model. "Optimal": response with the optimal manifold learning matrix P (no predictive filtering) derived for Fig. 1E. (**D**) Cumulative distribution of the Euclidean distance between new odors $\mathbf{s}_{\text{new}}$ and each model's PN response to new odors mixed with the background, $\mathbf{y}_{\text{mix}}$, after habituation, across all background, odors, and new odor concentrations tested. (**E**) Distribution of Jaccard similarities $J(z_{\text{n}}, z_{\text{mix}})$ of the various models, across all backgrounds and odor samples tested.

manifold of non-Gaussian backgrounds.

We perform initial numerical simulations, outlined in Fig. 2B, to assess the performance of different habituation schemes. We compare these rules to the average subtraction mechanism proposed in [36] ("Average"), as well as with the absence of habituation ("None") and the optimal manifold learning matrix P derived in the previous section (see *Methods*). The dynamical equations for each plasticity rule are integrated in the presence of a background as in Eq. (1), with concentrations fluctuating according to the turbulent stochastic process of Fig. 1A-B. After this habituation period, we present the network with mixtures of the background and new odors, $\mathbf{s}_{\text{mix}} = \mathbf{s}_{\text{b}}(t) + \mathbf{s}_{\text{new}}$. To assess how well that odor is decoded from the mixture, we compare its output with the neural tag z_{new} of the new odor alone.

We find that while average subtraction cannot inhibit the strong fluctuations of turbulent backgrounds (*Supp. Materials*, sec. 3), both the IBCM and BioPCA networks significantly reduce PN activity in response to the background (Fig. 2C) comparably to the optimal manifold learning matrix P. This confirms that both models provide adequate projections on the background subspace, allowing the Hebbian rule for W (Eq. (3)) to achieve its function of minimizing PN activity.

Moreover, we compare the models' performance for new odor recognition after habituation at the level of PN activities (Fig. 2D) and neural tags (Fig. 2E). For both metrics, average subtraction provides a very limited improvement compared to recognition without habituation, due to the strong background fluctuations caused by turbulence. In contrast, manifold learning implementations significantly improve odor recognition, with both IBCM and BioPCA networks performing similarly well. With respect to the distance in PN activity between the new odor alone and mixed with the background (Fig. 2D), these models result in a $\sim$ 3-fold improvement, but fall short of the optimum by a similar factor. This is not surprising, since P is fine-tuned for the distribution of new odors, which is unknown to the IBCM and BioPCA networks. Nonetheless, in terms of the Jaccard similarity between neural tags of the mixture and the new odor (Fig. 2B, right), the IBCM and BioPCA networks perform within 15 % similarity of the optimum (Fig. 2E), producing responses much more similar to the new odor than to background odors (Fig. S2). These models thus recover, from a mixture with the background, roughly 50 % of the KCs which are most activated by the new odor alone and define its identity, even when the new odor is present at just half the average whiff concentration (Fig. 2E, left). These results prompt us to investigate in more detail how the IBCM and BioPCA models learn the background manifold.

Analysis of background habituation by the IBCM model. We first focus on the IBCM model, since its mechanisms are less intuitive than PCA and have classically been characterized for visual input processes alternating between a fixed set of vectors [44, 50]. In its simplest form (Fig. 3A), the IBCM model describes the slow variation of a neuron's synaptic input weights $\mathbf{m}$ (a row in matrix M) as

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \mu h(h-\Theta)\mathbf{s}(t) \quad \text{where } h = \mathbf{m}\cdot\mathbf{s}(t) \\ \frac{\mathrm{d}\Theta}{\mathrm{d}t} &= \frac{1}{\tau_\Theta}(h^2-\Theta) \end{aligned} \tag{4}$$

where h is the activity in response to input $\mathbf{s}(t)$, μ is the learning rate, and Θ is an internal threshold converging to a temporal average $\Theta = \langle h^2 \rangle$ ($\langle\cdot\rangle$ denotes averages over the input fluctuations). In practice, we include lateral mean-field inhibition between interneurons with coupling parameter η, a mild nonlinearity to h preventing excessively large activations, and a small ($\varepsilon \ll 1$) decay term $-\varepsilon\mu\mathbf{m}$. For simulations with turbulent backgrounds, we use a variant of the model from [45] where the learning rate is divided by Θ to speed up convergence (see *Methods*).

The $\mathbf{m}$ equation introduces a competition between input patterns: inputs that cause $h > \Theta$ are further reinforced, while sub-threshold ones are further depressed; consequently, a neuron responds specifically to some inputs and does not respond to others [43]. This mechanism works when the threshold time scale τ_Θ is slow enough to average over fast input fluctuations $\mathbf{s}(t)$, yet still fast compared to the learning rate μ. This separation of time scales is to ensure $\mathbf{m}$ does not vary much while Θ averages over fluctuations; oscillations arise in the synaptic weights [51] without the separation $\tau_\Theta \ll 1/\mu$. This criterion can nonetheless be achieved during olfactory habituation over the course of 30-60 minutes.

To gain insight into how IBCM neurons learn the background subspace, we examine the fixed point equations of the model averaged over fast input fluctuations (*Supp. Materials*, sec. 4), finding exact expressions for these solutions in terms of the background concentration moments. From a linear stability analysis (*Supp. Materials*, sec. 4F and Fig. S3), we find that the only stable fixed points are those where the alignment of the synaptic weights with background odor vectors, $h_\gamma = \mathbf{m}\cdot\hat{\mathbf{s}}_\gamma$, take a large positive value h_{sp} ("specific") for *one* odor γ and a small, possibly negative value h_{ns} ("non-specific") for all other background components. Hence, an IBCM neuron learns to selectively respond to one background odor: the classical specificity property of this model [44] thus extends to quite general input stochastic processes of the form given in eq. 1. From the solutions for $\mathbf{m}$ and $\mathbf{h}$, we also derive analytical expressions for the fluctuation-averaged inhibitory W weights at steady-state (*Supp. Materials*, sec. 4G). Overall, our results show that a network of IBCM neurons performs manifold learning by having each neuron selectively suppress one background odor. Lateral inhibitory coupling between IBCM neurons (matrix L) help to push each neuron towards a different odor component [52].

To confirm our analysis, we perform numerical simulations with a simpler, weakly non-Gaussian background (Fig. 3B, see *Methods*). As expected, each IBCM neuron evolves over time to align with one background component (Fig. 3C,D), with steady-state average values of the dot products $\bar{h}_\gamma$ closely matching our analytical predictions $h_{\mathrm{sp}}, h_{\mathrm{ns}}$ (dashed lines). Different IBCM neurons (labeled by colors) become specific to different odors (line transparencies). The W weights also converge to steady-state average values matching our analytical results (Fig. 3E). The network of IBCM neurons performs habituation effectively, reducing both the mean and standard deviation (fluctuations) of the PN activity norm below $\sim 10\%$ of the input levels. Characterizing further the learning dynamics, we observe that the selectivity of IBCM neurons is acquired in two phases, first approaching a saddle point before converging to a selective fixed point (*Supp. Materials* sec. 6, Fig. S4, and Fig. 3C). This selectivity is driven by skewness (non-zero third moment) in the background statistics [44, 53] (Fig. S5).

Importantly, these properties of IBCM neurons are robust across different background statistics. We observe similar specificity and learning dynamics in IBCM neurons exposed to log-normal background fluctuations (Fig. S6) and in the simulations of Fig. 2 with the full turbulent statistics (Fig. 4A-B). In the latter case, averaging over long whiffs and blanks requires slower learning rates $1/\tau_\Theta$ and μ (see *Methods*and Table S2); despite stronger fluctuations, IBCM neurons align with a single background odor each.

Comparison of IBCM and BioPCA learning. As a point of comparison, we also analyze how the BioPCA network learns the background manifold in fluctuating environments. Despite strongly non-Gaussian statistics, the network converges to the expected PCA decomposition fixed point (see *Methods*, *Supp. Materials*, sec. 5 and sec. 6C). The matrix L becomes nearly diagonal, containing the principal values (Fig. 4C, Fig. S6B), while the rows of M converge to (scaled versions of) the principal component vectors, as evidenced by the error on their alignment decreasing to $\sim 1\,\%$ (Fig. 4D, Fig. S6C, Fig. S7).

On the whole, both models achieve similar habituation levels within $\sim$ 30 minutes. However, they rely on distinct mechanisms and converge to different vector bases for the background manifold. BioPCA neurons learn principal components: linear combinations of the true odors, distinguished by their variance. IBCM neurons, in contrast, rely on higher statistical moments of the inputs to select individual odor sources. Both models require in principle one neuron per background dimension, $N_{\mathrm{I}} = N_{\mathrm{B}}$, to span the background subspace. Superfluous neurons have little effect for the BioPCA model, where they reach principal values $L_{ii} \approx 0$. For the IBCM network, extra neurons are helpful, increasing the probability that each background odor will be selected by at least one of them. Notwithstanding these differences, both models produce very similar habituation and odor recognition performances in Fig. 2. They are also similar in their robustness to OSN noise (Fig. S8), and in their performance when combined with alternate Hebbian rules for the W weights based on different L^P norms (*Supp. Materials* sec. 7 and Fig. S9).

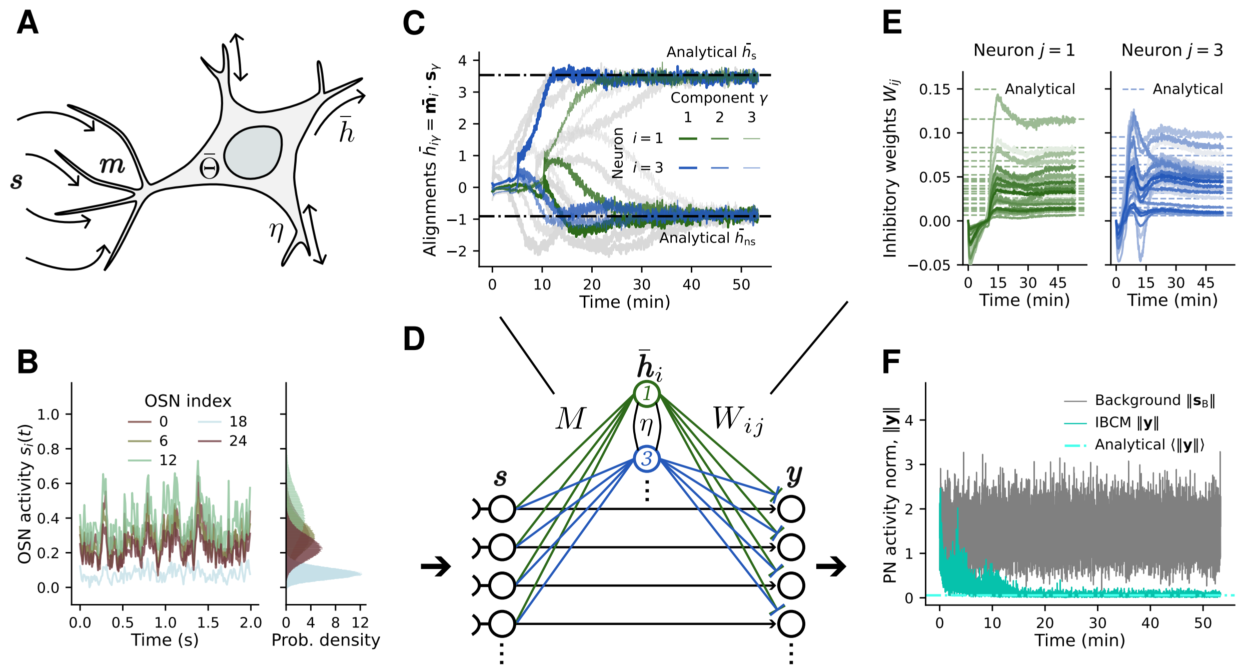


Fig. 3. The IBCM model of synaptic plasticity learns olfactory background components. (**A**) Illustration of an IBCM neuron's inputs from OSNs, $\mathbf{s}$, synaptic weights $\mathbf{m}$, activity $\bar{h}$, and internal threshold $\bar{\Theta}$. (**B**) OSN activities for a simple background process with 3 odors and weakly non-Gaussian concentration fluctuations. (**C**) Time series of the IBCM neurons' alignment with the different background odors (two neurons are highlighted in green and blue, with line widths indicating different odors γ). Each neuron predominantly aligns with one odor (large dot product value $\bar{h}_{\mathrm{s}}$ for one odor, small value $\bar{h}_{\mathrm{ns}}$ for the others). This steady-state alignment matches our analytical predictions (dashed lines, *Supp. Materials* sec. 4). (**D**) Location, in the olfactory network model, of the different weights and neurons illustrated in other panels. (**E**) Time series of the IBCM neurons' inhibitory synaptic weights, W, for the neurons highlighted in (C). The steady-state values of these weights can also be derived analytically (dashed lines) and align well with the simulations. (**F**) Norm of the response of projection neurons (PN) to the background process during habituation. The analytical prediction (dashed line) over-estimates the actual inhibition, since it neglects the contribution of fluctuations in M and W.

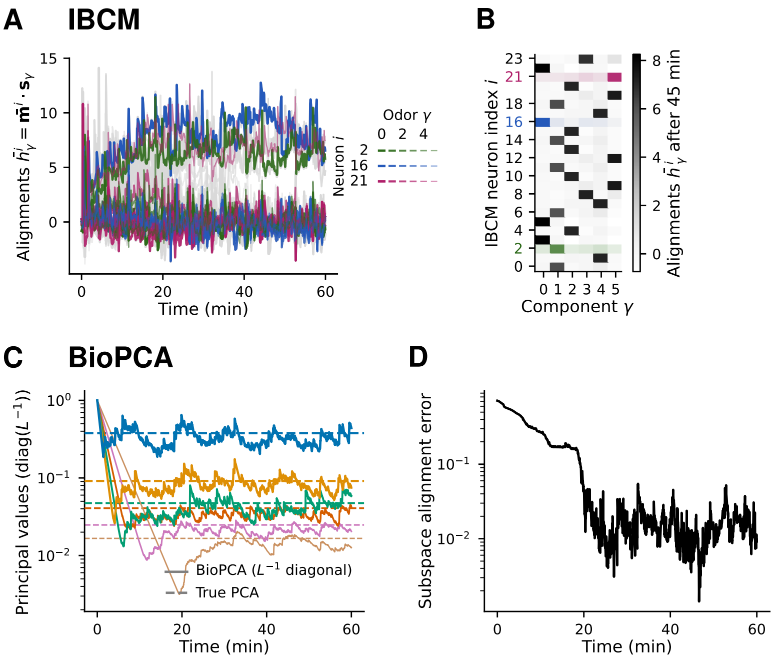


Fig. 4. Habituation of IBCM and BioPCA neurons to turbulent olfactory backgrounds. (**A**) Time series of the IBCM neurons' synaptic weight alignment with background odor during habituation to a six-odor background with the turbulent concentration stochastic process illustrated in Fig. 1A-B. Three neurons are highlighted with colors. (**B**) Table of each neuron's alignment after habituation, showing that IBCM neurons becomes selective for one odor even in this strongly fluctuating background. (**C**) Time series of the principal values learned by lateral interneurons obeying the BioPCA model during habituation to the same turbulent background. These principal values are stored in the inverse of the self-coupling weights L_{ii}^{-1} (inverse of the diagonal entries in the LN coupling matrix L). The principal values learned by the first N_B neurons converge to averages equal to the N_B non-zero eigenvalues of a PCA decomposition of the background (dashed horizontal lines). (**D**) Alignment error between the background subspace and the principal components learned by the BioPCA LNs (the rows of the LM matrix should be the principal components), confirming the model does learn the PCA decomposition of the background.

Performance in various olfactory space conditions. To understand the similar performance of the IBCM and BioPCA versions of manifold learning, we investigate the effect of various olfactory space parameters on them. We perform numerical simulations analogous to those of Fig. 2B for increasingly large olfactory space dimensions (N_{S}) and for a wider range of new odor concentrations (Fig. 5A). We consider dimensionalities ranging from half (25) that of the fruit fly (50) up to human (300) and mouse (1000) levels. While the performance of the optimal manifold learning algorithm increases with N_{S} up to a nearly perfect score, the IBCM and BioPCA networks reach a very similar plateau at $N_{\mathrm{S}} \sim 100$ (Fig. 5B). Remarkably, this plateau corresponds to the similarity between the new odor tag, z_{new}, and the tag $z_{\mathrm{new},\perp}$ of the new odor component orthogonal to the background, $\mathbf{y}_{\mathrm{new},\perp}$ ("orthogonal" pink line, Fig. 5B). This observation clarifies why both models perform similarly well: the local rules for W (Eq. (3)), based on minimizing PN activity, cause the parallel component of the new odor to be subtracted at the same time as the background. As long as the M weights provide complete projections of the inputs, the network produces a response $\mathbf{y}_{\mathrm{mix}} \approx \mathbf{s}_{\mathrm{new},\perp}$. In comparison, the optimal matrix P preserves some of the new odor's parallel component, thus maximizing the recognition of $\mathbf{s}_{\mathrm{new}}$.

Still, the levels of odor recognition reached by the IBCM and BioPCA models are significant, several standard deviations above chance similarity (black line, Fig. 5B). They also perform well above the average subtraction model, which was similar to no habituation: in that model, new odor signals are masked by strong fluctuations away from the average. For higher new odor concentrations, manifold learning provides a more modest improvement (Fig. 5C and Fig. S10), because habituation is not as crucial for very strong new odor whiffs which dominate the background. Overall, both IBCM and BioPCA interneurons subtract the background manifold from olfactory inputs while preserving new odor signals in regimes that are physically and computationally relevant for biological systems.

Discussion

Sensory adaptation to olfactory backgrounds is particularly challenging due to strong fluctuations generated by turbulent mixing in naturalistic conditions. We showed that predictive filtering strategies, which act on individual stimulus features, cannot adequately distinguish between changes in activity due to new odors and changes in activity due to fluctuations in the background. An alternative class of habituation strategies, manifold learning, could better identify new odors by learning to subtract projections of the instantaneous inputs onto the low-dimensional background manifold. We propose that inhibitory interneurons, which modulate the activity of principal neurons in early olfactory pathways, implement a manifold learning strategy for habituation. We explore two classes of synaptic plasticity rules, each of which combines a Hebbian-like rule and a linear projection learning rule (IBCM or BioPCA). Our analysis shows that these simplified linear manifold learning strategies are near-optimal for a range of physiologically relevant parameters, including when background odors display strong fluctuations such as those encountered in turbulent environments. Both plasticity rules show comparable performance on a habituation task, but learn distinct stimulus features. Notably, IBCM neurons select biologically relevant projections corresponding to independently varying components in the background mixture.

The biological underpinnings of our proposed model for olfactory background manifold projection are supported by previous experimental and theoretical studies. All connections in our network structure (OSN to PN, OSN to LN, LN to PN) are abundant in the connectome [12]. Habituation on the time scale of minutes has been shown to occur predominantly at the level of PNs in flies [22, 54] or M/T cells in mice [29]. Several studies found lateral inhibitory signals (GABA, glutamate) and their receptors for such signals (GABA-A, NDMA) to be essential to habituation [6, 27, 34, 35]. Our model relies on odor-specific PN inhibition by PN-to-LN plasticity for habituation, as observed in *Drosophila* [8] and honeybees [55]. Of note, we neglected feedback of PNs on LNs [31] and instead considered a feedforward network for mathematical simplicity, to illustrate our concept of manifold learning. Moreover, recent theoretical work has argued that the PN-LN connectivity pattern reflects correlations in PN activity, suggesting that the PN-LN circuit whitens odor representations in the antennal lobe [56]. However, the authors focused on hardwired computations preadapted to a given set of odors (*i.e.*, offline), whereas we addressed a different problem altogether, showing that online PCA is one plausible set of synaptic plasticity rules to achieve background subtraction in fluctuating environments.

Our theory provides salient predictions that could be tested experimentally. The simplest observable feature is the decrease in both the mean and variance of PN (or M/T cells in mice) activity after 20-60 minutes of exposure to turbulent odor mixtures (Fig. 2C and 3F). This phenomenon has already been observed for simpler backgrounds in *Drosophila* [8], which motivated our study. It could be directly tested in mice by calcium fluorescence imaging of glomeruli [37]. PN or glomerular activity should however be restored in response to new odors orthogonal to the learned background. In comparison, temporal average filtering would fail to reduce PN activity (Fig. 2C), while filtering based on recent samples (as in eq. 2) would rapidly suppress the response to new odors as well.

A more subtle feature in our proposed model is that lateral interneuron activity (LN) should, conversely, closely track background stimuli in real time to keep inhibiting PN responses (Fig. S8D-E). It may be experimentally challenging, however, to single out interneurons and record their fast fluctuations. A corollary of this model feature (which might be easier to measure) is that, since LN activity reflects projections on the learned background manifold, these neurons should become silent if the stimulus is suddenly switched to new odors with null projections on the previous subspace. Their activity should slowly recover on the time scale of habituation as they learn the new background.

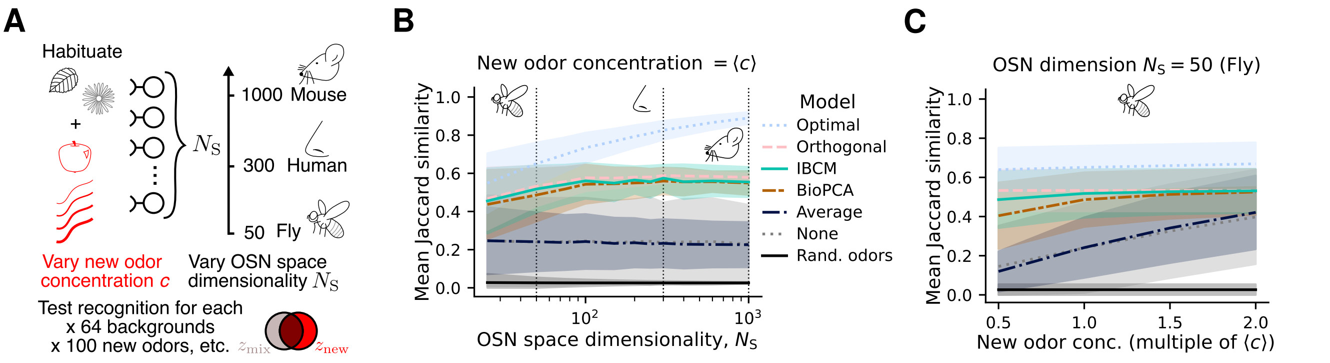


Fig. 5. Recognition performance as a function of olfactory space dimensionality and new odor concentration. (**A**) For various olfactory space dimensionalities $N_{\rm S}$ (*i.e.*, number of OSN types) and new odor concentrations $c_{\rm n}$ at test times, we perform simulations like those in Fig. 2: the model habituates to a six-odor turbulent background for ∼60 minutes before a new odor is introduced in the mixture. Each $(N_{\rm S}, c_{\rm new})$ condition is tested across 64 backgrounds, 100 new odors, 5 test times post-habituation, and 4 background samples at each test time. (**B**) New odor recognition performance, quantified by the Jaccard similarity between the new odor and the response to the mixture after habituation, as a function of $N_{\rm S}$, for different manifold learning models. Results shown here are for the new odor $c_{\rm new}$ equal to the average concentration of background odors, $\langle c \rangle$ (see Fig. S10 for all concentrations). "Optimal": manifold learning matrix W derived in Fig. 1E. "Orthogonal": similarity between the entire new odor and its component orthogonal to the background. "Rand. odors": similarity between two randomly selected odors (*i.e.*, similarity by chance). The shaded area represents one standard deviation across replicates. (**C**) Recognition performance as a function of the concentration at which the new odor is presented (multiples of $\langle c \rangle$), for the fly case ($N_{\rm S} = 50$; see *Supp. Materials*, Fig. S10A for all dimensions). Same legend as (B).

A third feature of habituation by manifold learning is that new odor recognition performance decreases with the distance to the background subspace (*i.e.*, as the norm of the orthogonal component $\mathbf{s}_{\rm n,\perp}$ decreases; Fig. S2C-D). This dependence would not be as strong in predictive filtering strategies. This correlation between odor recognition and distance from new odors could be tested in behavioral experiments.

Further theoretical work will also be necessary to refine our proposed implementation of manifold learning in olfactory circuits, and to assess how this strategy may be coupled with other odor recognition mechanisms. Schemes more sophisticated than a Hebbian rule would be necessary to reach the optimal performance promised by manifold learning (Figure 5) or to fully exploit the biologically relevant projections learned by IBCM neurons. Also, in our study, we focused on linear background manifolds (eq. 1) that could be decomposed into linear projections ($\mathbf{h} = LM\mathbf{s}$); while manifold projection also applies conceptually to non-additive odor mixtures, this extension will require olfactory circuit implementations of algorithms for curved manifold learning, such as manifold tiling [57]. Moreover, by choosing to focus on early olfactory processing, we neglected neuromodulatory inputs [58–60] and feedbacks from higher-level cognitive functions [61, 62]. For instance, while we assumed background and new odors are merely defined by their order of presentation, long-term memory of odors and other computations in the piriform cortex [63] likely help mammals focus their attention on relevant cues rather than on uninformative odors for,*e.g.*, odor trail tracking [64, 65]. Future investigation on this aspect could draw upon recent advances on attention mechanisms in artificial learning models [66]. Conversely, the concept of background manifold projection could prove useful for algorithms performing figure-ground segregation in time-varying signals, such as in video object detection [67].

Beyond olfaction, the interplay between habituation and attention also arises in other biological systems performing chemodetection in fluctuating environments [68]. For instance, in T cell antigen recognition [69], both (immune) memory and (T cell receptor) signal processing networks play important roles for pathogen detection amid a sea of irrelevant (self) antigens. Overall, we hope that our proposed model of habituation via manifold learning will motivate further theoretical and experimental efforts to clarify how living systems meet the challenge of adaptation to fluctuating backgrounds.

Materials and Methods

Odor vectors and concentrations. In our models, odors have a fixed, unit-normed direction, and an amplitude along that axis set by their (fluctuating) concentration: $\mathbf{s}(t) = c\hat{\mathbf{s}}$. Except for the idealized setup of Fig. 1E, vectors for background ($\hat{\mathbf{s}}_\gamma$) and new ($\hat{\mathbf{s}}_{\rm new}$) odors are drawn from the same distribution $\mathcal{P}_{\hat{\mathbf{s}}}$, by sampling i.i.d. exponential elements, then normalizing each vector. New odors are tested at fixed concentrations $c_{\rm new}$. Background concentrations $c_\gamma(t)$ follow a stochastic process, usually (Figs. 2, 5) the turbulent process illustrated in Fig. 1A-B. We simulate each odor concentration as a telegraph-like process, alternating blanks and whiffs with stochastic durations and whiff concentrations. The power-law distribution of whiffs and blanks durations (t_w, t_b) have a lower cutoff at 10 ms and upper cutoffs at 5 s (whiffs) or 8 s (blanks), respectively. The whiff concentration distribution has a scale $c_0 = 0.6$. We also considered weakly non-Gaussian (Fig. 3) and log-normal (Fig. S6) background concentrations, by simulating a multivariate Ornstein-Uhlenbeck $\{g_\gamma(t)\}$, then transforming these variables as $c_\gamma(t) = g_\gamma(t) + \nu g_\gamma(t)^2$ or $c_\gamma(t) = 10^{g_\gamma(t)}$, respectively. We used a short autocorrelation time $\tau = 20$ ms, an average $\langle g \rangle = 1/\sqrt{N_{\rm B}}$, and standard deviation $\sigma_g = 0.3$. For the non-Gaussian case, we set $\nu = 0.2$. Details of the stochastic simulation methods are provided in *Supp. Materials* sec. 2.

Optimizing predictive filtering and manifold learning regimes. In Eq. (2), we introduced an idealized inhibitory network response combining manifold learning and predictive filtering.

The objective of this network is to minimize the squared distance between its response, $\mathbf{b}(T) + \mathbf{s}_{\mathrm{n}} - \mathbf{u}(T)$, and the target odor alone, $\mathbf{s}_{\mathrm{n}}$. The corresponding loss is

$$\mathcal{L}_{v,P} = \left\langle \left\| \mathbf{b}(T) - \sum_{j=1}^{T-1} v_j \mathbf{b}(T-j) - P(\mathbf{b}(T) + \mathbf{s}_{\mathrm{new}}) \right\|^2 \right\rangle , \tag{5}$$

where the average is taken across samples (concentrations) from a given background and across new odors. Our goal was to determine the ideal performance of such a network, and the contribution of each habituation strategy depending on olfactory space parameters. We therefore solved for the optimal scalar coefficients v_j and the optimal $N_{\mathrm{S}} \times N_{\mathrm{S}}$ matrix P, as an upper bound on the mechanisms that real networks could learn during habituation. We minimized the loss by solving $\frac{\partial \mathcal{L}}{\partial v_j} = \frac{\partial \mathcal{L}}{\partial P_{ij}} = 0$. *Supp. Materials* sec. 1 details the calculation and the result; we obtained a general solution for any background mixture, as defined in Eq. (1), considering zero-mean and statistically independent concentrations.

For Fig. 1E, we considered a simpler particular case where background odors are orthogonal, new odors are drawn uniformly from the unit hypersphere, and background concentrations have an exponential autocorrelation function with time constant τ, $\langle c_\gamma(t) c_\rho(t+s) \rangle = \sigma^2 \delta_{\gamma\rho} e^{-|s|/\tau}$. The minimized loss is then

$$\mathcal{L}_{v,P} = N_{\mathrm{B}} \sigma^2 (1 - e^{-2/\tau}) / (1 + \tilde{N}_{\mathrm{S}} (1 - e^{-2/\tau})) \tag{6}$$

where $\tilde{N}_{\mathrm{S}} = (\sigma^2 / \sigma_n^2) N_{\mathrm{S}}$ and σ_n^2 is the new odor concentration variance. In the figure, we compared $\mathcal{L}_{v,P}$ with the limiting cases of pure predictive filtering ($\mathcal{L}_v$, setting $P = 0$) and pure manifold learning ($\mathcal{L}_P$, setting $v_j = 0$), which respectively give losses

$$\mathcal{L}_v = N_{\mathrm{B}} \sigma^2 (1 - e^{-2/\tau}) \ , \quad \mathcal{L}_P = N_{\mathrm{B}} \sigma^2 / (1 + \tilde{N}_{\mathrm{S}}) \ . \tag{7}$$

For the "optimal" strategy in Figures 2 and 5, we needed to derive the general solution (for non-orthogonal background and new odors drawn from $\mathcal{P}_{\hat{\mathbf{s}}}$) for pure manifold learning ($v = 0$) when the background has a non-zero average (as was the case in our simulations). In that case, the optimal projection matrix is

$$P = (\langle \mathbf{b}\mathbf{b}^{\intercal} \rangle + \langle \mathbf{b} \rangle \langle \mathbf{s}_{\mathrm{new}} \rangle^{\intercal}) M^{+} \tag{8}$$

where M^{+} is the Moore-Penrose pseudo-inverse (or the usual matrix inverse, when it exists) of

$$M = \langle \mathbf{b}\mathbf{b}^{\intercal} \rangle + \langle \mathbf{s}_{\mathrm{new}} \mathbf{s}_{\mathrm{new}}^{\intercal} \rangle + \langle \mathbf{b} \rangle \langle \mathbf{s}_{\mathrm{new}} \rangle^{\intercal} + \langle \mathbf{s}_{\mathrm{new}} \rangle \langle \mathbf{b} \rangle^{\intercal} \ . \tag{9}$$

We evaluated numerically the moments $\langle \mathbf{b}\mathbf{b}^{\intercal} \rangle$, $\langle \mathbf{s}_{\mathrm{new}} \mathbf{s}_{\mathrm{new}}^{\intercal} \rangle$, etc., by sampling unit vectors $\mathbf{s}_\gamma$, $\mathbf{s}_{\mathrm{new}}$ from $\mathcal{P}_{\hat{\mathbf{s}}}$ and background concentrations c_γ from the stationary distribution of the background process at hand – for Figs. 2 and 5, the turbulent statistics shown in Fig. 1A-B.

Mathematical model of the olfactory network. We model the instantaneous response of the olfactory network (Fig. 2A) to a stimulus $s(t)$ received at time t as the following set of neural activities in its different layers:

$$\bar{\mathbf{h}}(t) = \phi(LM\mathbf{s}(t)) \quad \text{(interneurons)} \tag{10}$$

$$\mathbf{y}(t) = \mathbf{s}(t) - W\bar{\mathbf{h}}(t) \quad \text{(PNs)} \tag{11}$$

$$\mathbf{z}(t) = 5\% \text{ most active in } R_\theta(Q\mathbf{y}(t)) \quad \text{(KCs)} \tag{12}$$

where ϕ is an element-wise nonlinearity and R_θ clips elements below threshold θ. We used $\phi(x) = A_{\mathrm{sat}} \tanh(x/A_{\mathrm{sat}})$ for IBCM neurons to saturate their activity at a large $A_{\mathrm{sat}} = 50$ for numerical stability; most of the time, $x \ll A_{\mathrm{sat}}$ is in the linear part of this function, so $\mathbf{y} \approx \mathbf{s} - WLM\mathbf{s}$. We did not apply a nonlinearity for the BioPCA network, nor for the IBCM model on simpler backgrounds (Fig. 3).

Then, the neural tag z is computed as in [36]. First, PN activities are projected to the KC layer by the sparse $N_{\mathrm{K}} \times N_{\mathrm{S}}$ binary matrix Q. Then, R_θ clips Kenyon cells (KCs) with activity below threshold; we set $\theta = \frac{1}{60} \times fN_{\mathrm{S}} \times \langle s_i \rangle$, where $\langle s_i \rangle$ is the average OSN activity in the current input $\mathbf{s}(t)$ and $f = 6/50$, the fraction of PNs forming a synapse with one KC in *Drosophila*. Finally, the neural tag z is the set of all non-zero KCs with activity above the 95th percentile of all KC activities.

The matrix Q is generated by randomly picking fN_{S} PNs to project to each KC (*i.e.*, picking fN_{S} non-zero elements in Q's row for that KC). We generated a new Q for each background tested in the numerical experiments in Figs. 2 and 5. When varying the olfactory space dimensionality N_{S} in Fig. 5, we preserved the relative size of PN and KC layers $N_{\mathrm{S}}/N_{\mathrm{K}} = 50/2000$ found in *Drosophila*; hence, for mice with $N_{\mathrm{S}} = 1000$ OR types, we used $N_{\mathrm{K}} = 40,000$ cortical cells (KC equivalent), and the Q matrix had $fN_{\mathrm{S}} = 120$ M/T cells (PN equivalent) projecting to each cortical cell. This ratio aligned with experimental estimates in mice giving ∼ 200 glomeruli connected to a cortical cell, or 10 % sparsity in Q [70–72]. The other matrices (M, W, L in BioPCA) are slowly updated according to synaptic plasticity rules during a habituation run.

Hebbian learning rule for W. The W weights are learned according to the Hebbian rule in Eq. (3). This rule derives from minimizing the average squared PN activity with L^2 regularization on the W_{ij}:

$$\mathcal{L}_W = \frac{1}{2} \langle \mathbf{y}^2 \rangle + \frac{\beta}{2\alpha} \sum_{i,j} W_{ij}^2 \tag{13}$$

where we recall that $\mathbf{y} = \mathbf{s} - W\bar{\mathbf{h}}$. Taking the W dynamics to be a gradient descent on $\mathcal{L}_W$ with rate α, $\frac{\mathrm{d}W_{ij}}{\mathrm{d}t} = -\alpha \frac{\partial \mathcal{L}_W}{\partial W_{ij}}$, yields the aforementioned Hebbian rule. The average $\langle\rangle$ is replaced by time averaging over a time window $1/\alpha$ using a slow rate α to implement online averaging of fast background fluctuations.

Average subtraction model. The "negative image" subtraction model proposed in [36] is effectively a mean filtering or average subtraction model. It can be recast in the form of our network structure by having $N_{\mathrm{I}} = 1$ interneuron with fixed activity $h = 1$, without M or L weights. The W weights are then a vector $\mathbf{w}_{\mathrm{avg}}$, which is subtracted from the PN response input since $h = 1$: $\mathbf{y}(t) = \mathbf{s}(t) - \mathbf{w}_{\mathrm{avg}}$. The Hebbian rule above is then

$$\frac{\mathrm{d}w_{\mathrm{avg}}}{\mathrm{d}t} = \alpha(\mathbf{s}(t) - \mathbf{w}_{\mathrm{avg}}) - \beta \mathbf{w}_{\mathrm{avg}} \ , \tag{14}$$

which makes $\mathbf{w}_{\mathrm{avg}}$ align, at steady-state, with the average of the background over a time window, $\mathbf{w}_{\mathrm{avg}} = \frac{\alpha}{\alpha+\beta} \langle s \rangle$ (*Supp. Materials*, sec. 3 for details).

Network of IBCM neurons. Equation 4 presents the simplest form of the IBCM model for a single neuron. In our olfactory network, we consider N_{I} IBCM neurons with constant mean-field lateral inhibitory coupling, as proposed in [44], corresponding here to a matrix L with 1 on the diagonal and $-\eta$ off-diagonal. Consequently, the reduced activity $\bar{h}_i$ of neuron i (*i.e.*, element i of $\bar{\mathbf{h}} = \phi(LM\mathbf{s})$) is

$$\bar{h}_i = \phi\left(\mathbf{m}_i \cdot \mathbf{s}(t) - \eta \sum_{j \neq i} \mathbf{m}_j \cdot \mathbf{s}(t) \right) = \phi(\bar{\mathbf{m}}_i \cdot \mathbf{s}(t)), \tag{15}$$

where we defined the inhibited weights $\bar{\mathbf{m}}_i = \mathbf{m}_i - \sum_{j \neq i} \mathbf{m}_j$, and where ϕ is the $\tanh$ nonlinearity introduced in Eq. (10). The complete dynamical equation for the synaptic weights $\mathbf{m}_i$ incoming into IBCM neuron i has additional terms due to this coupling,

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}_i}{\mathrm{d}t} =& \mu_{\bar{\Theta}_i} \bar{h}_i \left(\bar{h}_i - \bar{\Theta}_i \right) \phi'(\bar{\mathbf{m}}_i \cdot \mathbf{s}(t)) \mathbf{s}(t) \\ & - \eta \sum_{j \neq i} \mu_{\bar{\Theta}_j} \left(\bar{h}_j - \bar{\Theta}_j \right) \phi'(\bar{\mathbf{m}}_j \cdot \mathbf{s}(t)) \mathbf{s}(t) - \varepsilon \mu \mathbf{m}_i \ , \end{aligned} \tag{16}$$

where ϕ' is the derivative of the nonlinearity. We have also added a small decay term $-\varepsilon \mu \mathbf{m}_i$ to eliminate any component orthogonal to the background manifold in the random initial weights. For simulations with turbulent background statistics, we scaled the learning rate in the first two terms as

$$\mu_{\bar{\Theta}_i} = \frac{\mu}{(\bar{\Theta}_i)^2 + k_\theta^2} \ . \tag{17}$$

This form is similar to the variant introduced in [45], but we added a constant k_Θ in the denominator to prevent blowups at $t = 0$, where we initialize $\bar{\Theta} = 0$. For simpler backgrounds (Figs. 3, Fig. S6), we did not include this variant and simply used $\mu_\Theta = \mu$. Moreover, the internal threshold of each neuron, $\bar{\Theta}_i$, evolves as

$$\frac{\mathrm{d}\bar{\Theta}_i}{\mathrm{d}t} = \frac{1}{\tau_\Theta} ((\bar{h}_i)^2 - \bar{\Theta}_i) \tag{18}$$

such that it tracks the reduced neuron activity $\bar{h}_i$ averaged on an intermediate time scale τ_Θ.

Network of BioPCA neurons. We could use the "inverse-free PSP" version of the biologically plausible online PCA (BioPCA) proposed in [47] directly for the M and L weights of the interneuron layer. The model converges to a fixed point where the L matrix is diagonal with the principal values in it, and where the matrix LM contains the principal vectors in its rows, with norms specified by the pre-defined diagonal matrix $\underline{\Lambda}$ [47, Lemma 3]. The model specifies dynamical update rules for M and $L' = L^{-1}$, the inverse of L, rather than L directly. To avoid non-biological matrix inverse computations, the vector of interneuron activities $\mathbf{h}$ is computed with a Taylor series for L,

$$\bar{\mathbf{h}}(t) = \left({L'}_d^{-1} - {L'}_d^{-1} L'_o {L'}_d^{-1} \right) M \mathbf{s}(t) \approx LM\mathbf{s} \ , \tag{19}$$

where ${L'}_d$ contains the diagonal of L', and ${L'}_d$ contains the off-diagonal terms. This approximation is accurate at the fixed point where L' is diagonal ($L'_o \to 0$). The BioPCA dynamical update rules converging to this PCA decomposition are

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mu_M \left(\bar{\mathbf{h}} \, \mathbf{s}^\intercal - M \right) \tag{20}$$

$$\frac{dL'}{dt} = \mu_L \left(\bar{\mathbf{h}} \, \bar{\mathbf{h}}^\intercal - \underline{\Lambda} L' \underline{\Lambda} \right) \tag{21}$$

In practice, we set $\underline{\Lambda}_{kk} = \Lambda \left(1 - \frac{\lambda_r (k-1)}{(N_\mathrm{I} - 1)} \right)$, where Λ is the scaling factor for M weights described below to make IBCM and BioPCA perform similarly, and λ_r is the range of Λ values (between 0 and 1). We followed the original paper's recommendation for the linear decrease of Λ_{kk} with k and for setting $\mu_L = 2\mu_M$ For further comparison with the original paper [47], note the following equivalence between our notation $\leftrightarrow$ theirs: $M \leftrightarrow W$, $L \leftrightarrow M^{-1}$, $\mathbf{s} \leftrightarrow \mathbf{x}$, and $\bar{\mathbf{h}} \leftrightarrow \mathbf{y}$.

In our simulations, we added an extra interneuron applying the average subtraction mechanism described in Eq. (14), with $\beta = 0$, upstream of the BioPCA model. This way, the BioPCA network learned the decomposition of the covariance matrix rather than of $\langle \mathbf{s}\mathbf{s}^\intercal \rangle$, which still includes the average $\langle s \rangle$. This choice did not change the model performance, but made it more interpretable.

To measure the convergence of M's columns to the PCA vectors, in Fig. 4D, we computed, at each time point, the subspace alignment error proposed by [47],

$$E_\mathrm{Pro}(M) = \min_{Q \in \mathbb{O}_{N_\mathrm{S}}} \frac{\|FQ - U_\mathrm{PCA}\|_\mathrm{Frob}^2}{\|U_\mathrm{PCA}\|_\mathrm{Frob}^2} \tag{22}$$

where columns of U_PCA contains the N_B PCA vectors with non-zero eigenvalues, $F = (\underline{\Lambda}^{-1} LM)^\intercal$ contains the eigenvectors learned in the network's projection weights, and $\|U\|_\mathrm{Frob}^2 = \mathrm{Tr}(U^\intercal U)$ is the Frobenius matrix norm. The Q matrix minimizing the distance to give the alignment error solves the so-called orthogonal Procrustes problem and is $Q = U_F V^\intercal$ where U_F, V come from the SVD of $U_\mathrm{PCA} F^\intercal = U_F \Sigma V^\intercal$ [73].

Scaling parameter Λ for M weights. In the BioPCA model, the scale parameter Λ in the $\underline{\Lambda}$ matrix (see below eq. 21) controls the magnitude of weights M. This scale influences the strength of habituation: larger $M \sim \Lambda$ weights allow smaller $W \sim 1/\Lambda$ weights that are less constrained by regularization (β term in eq. 13) and thus further reduce PN activity. We set Λ to the value necessary to achieve the same background reduction level as the IBCM network, as predicted by our analytical calculations for the IBCM fixed points and post-habituation PN activity; see *Supp. Materials*, sec. 8 for detailed expressions. Of note, we rescaled the μ_L rate to μ_L/Λ^2 in the BioPCA model (eq. 21) to preserve exactly the same learning dynamics for any Λ, just with M weights scaled up or down.

For comparison, we introduced a similar scale parameter $\Lambda_\mathbf{IBCM}$ in the IBCM model, but we generally kept it equal to 1 (its implicit value by default), since that was sufficient to achieve complete background manifold projection (Fig. S11). Similar to BioPCA, scaling of the learning rate μ was required for $\Lambda_\mathbf{IBCM} \neq 1$ (*Supp. Materials*, sec. 8 for details).

Numerical simulations and model parameter values. We integrated the stochastic differential equations of the network, with the background processes simulated as described above, using an Euler scheme with time step $\Delta t = 10$ ms. Below, we give rates in scaled units where this time step = 1.

By default, we performed simulations lasting $360,000$ time steps (1 hour) with $N_\mathrm{S} = 25$ dimensions, $N_\mathrm{K} = 40 N_\mathrm{S}$ Kenyon cells, $N_\mathrm{B} = 6$ background odors, $N_\mathrm{I} = 24$ IBCM neurons or $N_\mathrm{I} = N_\mathrm{B}$ BioPCA neurons. For W Hebbian learning, we used $\alpha = 10^{-4}$ and $\beta = 2 \times 10^{-5}$. However, for Fig. 3 and Fig. S6, we used $N_\mathbf{B} = 3$, $N_\mathrm{I} = 6$ (IBCM), $\alpha = 2.5 \times 10^{-4}$, and $\beta = 5 \times 10^{-5}$.

For the IBCM weights, we used by default $\mu = 1.25 \times 10^{-3}$, $\tau_\Theta = 1600$, $\eta = 0.6/N_\mathrm{I}$, $k_\Theta = 0.1$, $\epsilon = 0.005$, and $A = 50$ as the maximum amplitude of the nonlinearity ϕ. For the simple background in Fig. 3, we used $\mu = 1.5 \times 10^{-3}$, $\tau_\Theta = 200$, $\eta = 0.5/N_\mathrm{I}$, we did not apply the ϕ nonlinearity or divide the learning rate by $k_\Theta + \Theta$. Also, for the simulations in higher dimensions in Fig. 5, we used a slower learning $\mu = 7.5 \times 10^{-4}$ and $\tau_\Theta = 2000$. For the BioPCA model, we used by default $\mu = 10^{-4}$, $\mu_L = 2\mu$, and $\lambda_r = 0.5$ (the range of $\underline{\Lambda}_{kk}$ entries). For Fig. 4, we fixed $\Lambda_\mathrm{PCA} = 8$ instead of the exact value making BioPCA and IBCM inhibit the background equivalently (described above). A full list of parameter definitions and values is provided in Tables S1 and S2.

The background process was initialized to a random sample from its stationary distribution. We initialized the W weights to zero, and the M weights to random i.i.d. normal samples with standard deviation 0.2 (or 0.3 for Fig. 4) for IBCM, or standard deviation $\Lambda_{\mathrm{PCA}}/\sqrt{N_{\mathrm{S}}}$ for BioPCA. For the latter model, we initialized L to the identity matrix (as recommended in the original paper); for IBCM, we initialized $\bar{\Theta}$ to the value of $\bar{h}$ with the initial weights and background.

Code Availability. All simulation code is available on Github: https://github.com/frbourassa/olfactory_habituation

Author contributions. All authors designed research, FXPB performed research, and FXPB and GR wrote the paper with input from PF and MV.

Competing interests. The authors declare no competing interest.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation, through the Center for the Physics of Biological Function (PHY-1734030). F.X.P.B. was supported in part by an award from the Fonds de recherche du Québec – Nature et technologies. P.F. was supported by grants from the NSERC Discovery and CIHR Programs, and by a Fonds Courtois award. M.V. was supported by NIH via the Grant No. RF1NS128865-01. G.R acknowledges support from a joint research agreement between NTT Research Inc. and Princeton University.

Bibliography

1. R. F. Thompson and W. A. Spencer, Habituation: a model phenomenon for the study of neuronal substrates of behavior, Psychological review **73**, 16 (1966).
2. C. H. Rankin, T. Abrams, R. J. Barry, S. Bhatnagar, D. F. Clayton, J. Colombo, G. Coppola, M. A. Geyer, D. L. Glanzman, S. Marsland et al., Habituation revisited: an updated and revised description of the behavioral characteristics of habituation, Neurobiology of learning and memory **92**, 135 (2009).
3. B. Wark, A. Fairhall, and F. Rieke, Timescales of inference in visual adaptation, Neuron **61**, 750 (2009).
4. A. I. Weber and A. L. Fairhall, The role of adaptation in neural coding, Current opinion in neurobiology **58**, 135 (2019).
5. M. A. Webster, Visual adaptation, Annual review of vision science **1**, 547 (2015).
6. D. Chaudhury, L. Manella, A. Arellanos, O. Escanilla, T. A. Cleland, and C. Linster, Olfactory bulb habituation to odor stimuli, Behavioral neuroscience **124**, 490 (2010).
7. D. Pérez-González and M. S. Malmierca, Adaptation in the auditory system: an overview, Frontiers in integrative neuroscience **8**, 19 (2014).
8. S. Das, M. K. Sadanandappa, A. Dervan, A. Larkin, J. A. Lee, I. P. Sudhakaran, R. Priya, R. Heidari, E. E. Holohan, A. Pimentel et al., Plasticity of local GABAergic interneurons drives olfactory habituation, Proc Natl Acad Sci USA **108**, E646 (2011).
9. A. Celani, E. Villermaux, and M. Vergassola, Odor Landscapes in Turbulent Environments, Phys. Rev. X **4**, 041015 (2014).
10. E. Yee, P. R. Kosteniuk, G. M. Chandler, C. A. Biltoft, and J. F. Bowers, Statistical characteristics of concentration fluctuations in dispersing plumes in the atmospheric surface layer, Boundary-Layer Meteorology **65**, 69 (1993).
11. G. Reddy, V. N. Murthy, and M. Vergassola, Olfactory Sensing and Navigation in Turbulent Environments, Annu. Rev. Condens. Matter Phys. **13**, 191 (2022).
12. P. Schlegel, A. S. Bates, T. Stürner, S. R. Jagannathan, N. Drummond, J. Hsu, L. Serratosa Capdevila, A. Javier, E. C. Marin, A. Barth-Maron et al., Information flow, cell types and stereotypy in a full olfactory connectome, Elife **10**, e66018 (2021).
13. R. Benton, Drosophila olfaction: past, present and future, Proceedings of the Royal Society B **289**, 20222054 (2022).
14. L. B. Vosshall and R. F. Stocker, Molecular Architecture of Smell and Taste in Drosophila, Annu. Rev. Neurosci. **30**, 505 (2007).
15. E. A. Hallem and J. R. Carlson, Coding of Odors by a Receptor Repertoire, Cell **125**, 143 (2006).
16. C. Martelli and D. A. Storace, Stimulus driven functional transformations in the early olfactory system, Frontiers in Cellular Neuroscience **15**, 684742 (2021).
17. G. Reddy, J. D. Zak, M. Vergassola, and V. N. Murthy, Antagonism in olfactory receptor neurons and its implications for the perception of odor mixtures, eLife **7**, e34958 (2018).
18. T. Tsukahara, D. H. Brann, S. L. Pashkovski, G. Guitchounts, T. Bozza, and S. R. Datta, A transcriptional rheostat couples past activity to future sensory responses, Cell **184**, 6326 (2021).
19. S. R. Olsen, V. Bhandawat, and R. I. Wilson, Divisive Normalization in Olfactory Population Codes, Neuron **66**, 287 (2010).
20. N. Kadakia and T. Emonet, Front-end Weber-Fechner gain control enhances the fidelity of combinatorial odor coding, eLife **8**, e45293 (2019).
21. J.-M. Devaud, A. Acebes, and A. Ferrus, Odor exposure causes central adaptation and morphological changes in selected olfactory glomeruli in Drosophila, Journal of Neuroscience **21**, 6274 (2001).
22. S. Sachse, E. Rueckert, A. Keller, R. Okada, N. K. Tanaka, K. Ito, and L. B. Vosshall, Activity-dependent plasticity in an olfactory circuit, Neuron **56**, 838 (2007).
23. C. Martelli and A. Fiala, Slow presynaptic mechanisms that mediate adaptation in the olfactory pathway of Drosophila, Elife **8**, e43735 (2019).
24. S. G. Solomon and A. Kohn, Moving Sensory Adaptation beyond Suppressive Effects in Single Neurons, Current Biology **24**, R1012 (2014).
25. K. Krishnamurthy, A. M. Hermundstad, T. Mora, A. M. Walczak, and V. Balasubramanian, Disorder and the Neural Representation of Complex Odors, Frontiers in Computational Neuroscience **16** (2022).
26. K. A. Fulton, D. Zimmerman, A. Samuel, K. Vogt, and S. R. Datta, Common principles for odour coding across vertebrates and invertebrates, Nature Reviews Neuroscience **25**, 1 (2024).
27. T. W. Margrie, B. Sakmann, and N. N. Urban, Action potential propagation in mitral cell lateral dendrites is decremental and controls recurrent and lateral inhibition in the mammalian olfactory bulb, Proceedings of the National Academy of Sciences **98**, 319 (2001).
28. Y.-H. Chou, M. L. Spletter, E. Yaksi, J. C. S. Leong, R. I. Wilson, and L. Luo, Diversity and wiring variability of olfactory local interneurons in the Drosophila antennal lobe, Nature Neuroscience **13**, 439 (2010).
29. A. Vinograd, Y. Livneh, and A. Mizrahi, History-Dependent Odor Processing in the Mouse Olfactory Bulb, Journal of Neuroscience **37**, 12018 (2017).
30. M. N. Economo, K. R. Hansen, and M. Wachowiak, Control of Mitral/Tufted Cell Output by Selective Inhibition among Olfactory Bulb Glomeruli, Neuron **91**, 397 (2016).
31. I. P. Sudhakaran, E. E. Holohan, S. Osman, V. Rodrigues, K. Vijayraghavan, and M. Ramaswami, Plasticity of recurrent inhibition in the Drosophila antennal lobe, Journal of Neuroscience **32**, 7225 (2012).
32. I. Twick, J. A. Lee, and M. Ramaswami, Olfactory habituation in Drosophila—odor encoding and its plasticity in the antennal lobe, Progress in Brain Research **208**, 3 (2014).
33. Y. Suzuki, J. E. Schenk, H. Tan, and Q. Gaudry, A population of interneurons signals changes in the basal concentration of serotonin and mediates gain control in the Drosophila antennal lobe, Current Biology **30**, 1110 (2020).
34. A. M. Dacks, D. S. Green, C. M. Root, A. J. Nighorn, and J. W. Wang, Serotonin modulates olfactory processing in the antennal lobe of Drosophila, Journal of neurogenetics **23**, 366 (2009).
35. A. Larkin, S. Karak, R. Priya, A. Das, C. Ayyub, K. Ito, V. Rodrigues, and M. Ramaswami, Central synaptic mechanisms underlie short-term olfactory habituation in Drosophila larvae, Learning & memory **17**, 645 (2010).
36. Y. Shen, S. Dasgupta, and S. Navlakha, Habituation as a neural algorithm for online odor discrimination, Proc Natl Acad Sci USA **117**, 12402 (2020).
37. J. D. Zak, G. Reddy, M. Vergassola, and V. N. Murthy, Antagonistic odor interactions in olfactory sensory neurons are widespread in freely breathing mice, Nature Communications **11**, 3350 (2020).
38. P. Pfister, B. C. Smith, B. J. Evans, J. H. Brann, C. Trimmer, M. Sheikh, R. Arroyave, G. Reddy, H.-Y. Jeong, D. A. Raps et al., Odorant receptor inhibition is fundamental to odor encoding, Current Biology **30**, 2574 (2020).
39. V. Singh, N. R. Murphy, V. Balasubramanian, and J. D. Mainland, Competitive binding predicts nonlinear responses of olfactory receptors to complex mixtures, Proceedings of the National Academy of Sciences **116**, 9598 (2019).
40. T. Ackels, A. Erskine, D. Dasgupta, A. C. Marin, T. P. A. Warner, S. Tootoonian, I. Fukunaga, J. J. Harris, and A. T. Schaefer, Fast odour dynamics are encoded in the olfactory system and guide behaviour, Nature **593**, 558 (2021).
41. Y. Wu, K. Chen, C. Xing, M. Huang, K. Zhao, and W. Zhou, Human olfactory perception embeds fine temporal resolution within a single sniff, Nature Human Behavior **8**, 2168 (2024).
42. S. Dasgupta, C. F. Stevens, and S. Navlakha, A neural algorithm for a fundamental computing problem, Science **358**, 793 (2017).
43. E. L. Bienenstock, L. N. Cooper, and P. W. Munro, Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex, Journal of Neuroscience **2**, 32 (1982).
44. N. Intrator and L. N. Cooper, Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions, Neural Networks **5**, 3 (1992).
45. C. C. Law and L. N. Cooper, Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory, Proceedings of the National Academy of Sciences **91**, 7797 (1994).
46. L. N. Cooper and M. F. Bear, The BCM theory of synapse modification at 30: interaction of theory with experiment, Nature Reviews Neuroscience **13**, 798 (2012).
47. V. Minden, C. Pehlevan, and D. B. Chklovskii, Biologically Plausible Online Principal Component Analysis Without Recurrent Neural Dynamics, in *2018 52nd Asilomar Conference on Signals, Systems, and Computers* (IEEE, Pacific Grove, CA, USA, 2018) pp. 104–111.
48. E. Oja, Simplified neuron model as a principal component analyzer, Journal of Mathematical Biology **15**, 267 (1982).
49. A. Hyvärinen and E. Oja, Independent component analysis: algorithms and applications, Neural Networks **13**, 411 (2000).
50. N. Intrator and J. I. Gold, Three-Dimensional Object Recognition Using an Unsupervised BCM Network: The Usefulness of Distinguishing Features, Neural Computation **5**, 61 (1993).
51. L. Udeigwe, P. Munro, and G. Ermentrout, Emergent Dynamical Properties of the BCM Learning Rule, Journal of mathematical neuroscience **7**, 2 (2017).
52. G. C. Castellani, N. Intrator, H. Shouval, and L. N. Cooper, Solutions of the BCM learning rule in a network of lateral interacting nonlinear neurons, Network: Computation in Neural Systems **10**, 111 (1999).
53. M. Froc and M. C. W. van Rossum, Slowdown of BCM plasticity with many synapses, Journal of Computational Neuroscience **46**, 141 (2019).
54. J. Cafaro, Multiple sites of adaptation lead to contrast encoding in the *Drosophila* olfactory system, Physiological Reports **4**, e12762 (2016).
55. J.-Y. Chen, E. Marachlian, C. Assisi, R. Huerta, B. H. Smith, F. Locatelli, and M. Bazhenov, Learning Modifies Odor Mixture Processing to Improve Detection of Relevant Components, The Journal of Neuroscience **35**, 179 (2015).

56. N. M. Chapochnikov, C. Pehlevan, and D. B. Chklovskii, Normative and mechanistic model of an adaptive circuit for efficient encoding and feature extraction, Proceedings of the National Academy of Sciences **120**, e2117484120 (2023).
57. A. Sengupta, C. Pehlevan, M. Tepper, A. Genkin, and D. Chklovskii, Manifold-tiling Localized Receptive Fields are Optimal in Similarity-preserving Neural Networks, in *Advances in Neural Information Processing Systems*, Vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018).
58. M. C. Ogg, J. M. Ross, M. Bendahmane, and M. L. Fletcher, Olfactory bulb acetylcholine release dishabituates odor responses and reinstates odor investigation, Nature communications **9**, 1868 (2018).
59. C. Linster and T. A. Cleland, Neuromodulation of olfactory transformations, Current opinion in neurobiology **40**, 170 (2016).
60. S. D. Shea, L. C. Katz, and R. Mooney, Noradrenergic induction of odor-specific neural habituation and olfactory memories, Journal of Neuroscience **28**, 10711 (2008).
61. T. A. Cleland and A. Borthakur, A systematic framework for olfactory bulb signal transformations, Frontiers in computational neuroscience **14**, 579143 (2020).
62. S. R. Odell, D. Clark, N. Zito, R. Jain, H. Gong, K. Warnock, R. Carrion-Lopez, C. Maixner, L. Prieto-Godino, and D. Mathew et al., Internal state affects local neuron function in an early sensory processing center to shape olfactory behavior in Drosophila larvae, Scientific Reports **12**, 15767 (2022).
63. R. M. Blazing and K. M. Franks, Odor coding in piriform cortex: mechanistic insights into distributed coding, Current Opinion in Neurobiology **64**, 96 (2020).
64. G. Reddy, B. I. Shraiman, and M. Vergassola, Sector search strategies for odor trail tracking, Proceedings of the National Academy of Sciences **119**, e2107431118 (2022).
65. N. Rigolli, G. Reddy, A. Seminara, and M. Vergassola, Alternation emerges as a multi-modal strategy for turbulent odor navigation, eLife **11**, e76989 (2022).
66. C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen et al., In-context Learning and Induction Heads (2022).
67. L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang, New generation deep learning for video object detection: A survey, IEEE Transactions on Neural Networks and Learning Systems **33**, 3195 (2022).
68. L. Eckert, M. S. Vidal-Saez, Z. Zhao, J. Garcia-Ojalvo, R. Martinez-Corral, and J. Gunawardena, Biochemically plausible models of habituation for single-cell learning, Current Biology **34**, 5646 (2024).
69. J.-B. Lalanne and P. François, Chemodetection in fluctuating environments: Receptor coupling, buffering, and antagonism, Proc Natl Acad Sci USA **112**, 1898 (2015).
70. E. R. Soucy, D. F. Albeanu, A. L. Fantana, V. N. Murthy, and M. Meister, Precision and diversity in an odor map on the olfactory bulb, Nature Neuroscience **12**, 210 (2009).
71. A. Apicella, Q. Yuan, M. Scanziani, and J. S. Isaacson, Pyramidal Cells in Piriform Cortex Receive Convergent Input from Distinct Olfactory Bulb Glomeruli, The Journal of Neuroscience **30**, 14255 (2010).
72. I. G. Davison and M. D. Ehlers, Neural Circuit Mechanisms for Pattern Detection and Feature Combination in Olfactory Cortex, Neuron **70**, 82 (2011).
73. P. H. Schönemann, A Generalized Solution of the Orthogonal Procrustes Problem, Psychometrika **31**, 1 (1966).
74. M. Dow, Explicit inverses of Toeplitz and associated matrices, ANZIAM J. **44**, E185 (2003).
75. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods **17**, 261 (2020).
76. DLMF, *NIST Digital Library of Mathematical Functions*, https://dlmf.nist.gov/, Release 1.2.4 of 2025-03-15 (2020), f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.
77. D. T. Gillespie, The mathematics of Brownian motion and Johnson noise, American Journal of Physics **64**, 225 (1996).
78. C. W. Gardiner, *Stochastic methods : a handbook for the natural and social sciences*, 4th ed., Springer complexity (Springer, Berlin, 2009).
79. N. Balakrishnan and W. W. S. Chen, *Handbook of Tables for Order Statistics from Lognormal Distributions with Applications* (Springer US, Boston, MA, 1999).

Supplementary Materials

1. Optimal models of manifold learning and predictive filtering

In this section, we detail the optimization problem we solved to delineate regimes of predictive filtering and manifold learning, as shown in Fig. 1E and Fig. S1.

A. Definition of the loss function.

We want to minimize the loss function

$$\mathcal{L}_{v,P} = \left\langle \left\| \mathbf{b}_t - \sum_{l=1}^{t-1} v_l \mathbf{b}_{t-l} - P(\mathbf{b}_t + \mathbf{x}) \right\|^2 \right\rangle_{\mathbf{b}\sim\mathcal{P},\, \mathbf{x}\sim\mathcal{Q}} \tag{23}$$

as a function of the scalar coefficients v_l (predictive filtering) and of the matrix P (manifold learning). In this section, we call $\mathbf{b}_{t'}$ the background OSN input vector at time t', and $\mathbf{x}$ the new odor (appearing at time t), instead of $\mathbf{s}_{\mathrm{b}}$ and $\mathbf{s}_{\mathrm{new}}$. The background is a linear combination of pre-defined odor vectors, $\hat{\mathbf{y}}_\rho$, weighted by stochastic concentrations, $\tilde{c}_{\rho,t'}$, so $\mathbf{b}_{t'} = \sum_{\rho=1}^{N_{\mathrm{B}}} \tilde{c}_{\rho,t'} \hat{\mathbf{y}}_\rho$. For simplicity, the concentrations are assumed i.i.d. and stationary with mean zero, variance $\langle \tilde{c}_{\rho,t'} \tilde{c}_{\lambda,t'} \rangle = \sigma^2 \delta_{\rho\lambda}$, and autocorrelation function $\langle \tilde{c}_{\rho,t'} \tilde{c}_{\lambda,t'\pm s} \rangle = C(s)\delta_{\rho\lambda}$, with $C(0) = \sigma^2$; together, these concentrations statistics define the background vector distribution $\mathcal{P}$. The new odor $\mathbf{x}$ comes from some distribution $\mathcal{Q}$ we assume has zero mean and finite covariance matrix $\langle \mathbf{x}\mathbf{x}^\intercal \rangle$.

A.1. Remarks on notation. In this calculation, sums and matrix products are applied over three different indices, denoting olfactory dimensions, time, and background odors, *e.g.*, $(P\mathbf{x})_i = \sum_j P_{ij} x_j$. To make notation more concise, we rewrite several sums as dot products. To clarify the indices on which these products are, we use boldface $\mathbf{x}$ on vectors in olfactory dimensions, and underlines $\underline{\ }$ ($\underline{\underline{\ }}$) for $\underline{\text{vectors}}$ ($\underline{\underline{\text{matrices}}}$) in time dimensions. We write out explicitly sums with Greek indices on background odor indices, $\sum_{\rho=1}^{N_{\mathrm{B}}}$.

A.2. Expanding the loss function terms. With the above assumptions on the background and new odor statistics, we can expand the square and write out the different terms in the loss function. First, using the statistical independence and zero mean property of $\mathbf{x}$ and $\mathbf{b}_{t'}$, the terms to evaluate are

$$\begin{aligned}\mathcal{L}_{v,P} = &\langle \mathbf{b}_t{}^\intercal \mathbf{b}_t \rangle + \sum_{l,m} v_l v_m \langle \mathbf{b}_{t-l}{}^\intercal \mathbf{b}_{t-m} \rangle + \langle \mathbf{b}_t{}^\intercal P^\intercal P \mathbf{b}_t \rangle + \langle \mathbf{x}^\intercal P^\intercal P \mathbf{x} \rangle \\ &- 2\sum_{l=1}^{t-1} v_l \langle \mathbf{b}_t{}^\intercal \mathbf{b}_{t-l} \rangle - 2\langle \mathbf{b}_t{}^\intercal P \mathbf{b}_t \rangle + 2\sum_{l=1}^{t-1} v_l \langle \mathbf{b}_{t-l}{}^\intercal P \mathbf{b}_t \rangle \ .\end{aligned}$$

We compute these terms more explicitly by using the background statistics defined above. The loss function is thus

$$\begin{aligned}\mathcal{L} = &N_{\mathrm{B}}\sigma^2 + N_{\mathrm{B}} \sum_{l,m=1}^{t-1} v_l v_m C(l-m) + \sigma^2 \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho^\intercal P^\intercal P \hat{\mathbf{y}}_\rho + \langle \mathbf{x}^\intercal P^\intercal P \mathbf{x} \rangle \\ &- 2N_{\mathrm{B}} \sum_{l=1}^{t-1} v_l C(l) - 2\sigma^2 \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho + 2\left(\sum_{l=1}^{t-1} v_l C(l)\right)\left(\sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho\right) \ .\end{aligned} \tag{24}$$

We did not need to assume that background odors were orthogonal to get this answer; the statistical independence of their concentrations $\tilde{c}_{\rho,t}$ removed cross-odor terms. Most terms involve only P or v; only the last term couples the two strategies together.

B. Solving for the optimal P and v.

B.1. Loss function derivatives and resulting optimum equations. We can now take the derivative of this loss function with respect to the parameters v_j and P_{ij}. After working out the derivatives of the different terms, the result is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial v_j} &= 2N_{\mathrm{B}} \sum_{l=1}^{t-1} v_l C(l-j) - 2N_{\mathrm{B}} C(j) + 2\left(\sum_\rho \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho\right) C(j) \\ \frac{\partial \mathcal{L}}{\partial P_{ij}} &= 2\sigma^2 \sum_{l=1}^{N_{\mathrm{S}}} P_{il} \hat{\mathbf{y}}_{\rho,l} \hat{\mathbf{y}}_{\rho,j} + 2\sum_{l=1}^{N_{\mathrm{S}}} P_{il} \langle x_l x_j \rangle - 2\left(\sigma^2 - \sum_{l=1}^{t-1} v_l C(l)\right) \hat{\mathbf{y}}_{\rho,i} \hat{\mathbf{y}}_{\rho,j}\end{aligned}$$

To shorten the notation of terms involving the autocorrelation function $C(s)$, we introduce the vectors $\underline{c} = (C(1), \dots, C(t-1))$ and $\underline{v} = (v_1, \dots, v_{t-1})$, and the matrix $\underline{\underline{C}}_{ij} = C(i-j)$. We note that $\underline{\underline{C}}$ is a Toeplitz matrix (*i.e.*, C_{ij} only depends on $i-j$), symmetric since $C(i-j) = C(j-i)$. These properties help to express its inverse explicitly in some cases [74].

Setting the derivatives to zero to find the optimum parameters, we thus have a set of vector and matrix equations for $\underline{v}$ and P, respectively:

$$0 = N_{\mathrm{B}} \underline{\underline{C}} \underline{v} - \left(N_{\mathrm{B}} - \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho \right) \underline{c} \tag{25}$$

$$0 = P \left(\sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal + \langle \mathbf{x}\mathbf{x}^\intercal \rangle \right) - (\sigma^2 - \underline{v}^\intercal \underline{c}) \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal \tag{26}$$

B.2. Solving for P in terms of $\underline{v}$. The best solution path is to first solve for P in terms of $\underline{v}^\intercal \underline{c}$, then solve for $\underline{u}$. We define the $N_{\mathrm{S}} \times N_{\mathrm{S}}$ symmetric matrix

$$M = \sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal + \langle \mathbf{x}\mathbf{x}^\intercal \rangle \tag{27}$$

which admits a spectral decomposition $M = U \Sigma U^\intercal$ and a Moore-Penrose pseudo-inverse $M^+ = U \Sigma^+ U^\intercal$, which is the actual inverse M^{-1} when M is invertible (*i.e.*, no zero eigenvalue in Σ). Equation 26 thus takes the form $PM = (\sigma^2 - \underline{v}^\intercal \underline{c}) \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal$, which can be inverted for $P_M = P U \Sigma \Sigma^+ U^\intercal$, the component of P in the subspace spanned by M's eigenvectors of non-zero eigenvalues. The P component in the null space of M, if any, is not constrained by this optimization problem, so we set it to zero, and take $P = P_M$. Hence,

$$P = (\sigma^2 - \underline{v}^\intercal \underline{c}) \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal M^+ \ . \tag{28}$$

B.3. Solving for $\underline{v}$. We can now insert the implicit solution for P in equation 25 for $\underline{v}$. We first evaluate the term

$$\sum_\rho \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho = \sum_\rho \hat{\mathbf{y}}_\rho^\intercal \left[(\sigma^2 - \underline{v}^\intercal \underline{c}) \sum_\lambda \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal M^+ \right] \hat{\mathbf{y}}_\rho = (\sigma^2 - \underline{v}^\intercal \underline{c}) N_{\mathrm{B}} m_y$$

where we have defined

$$m_y = \frac{1}{N_{\mathrm{B}}} \sum_{\rho,\lambda=1}^{N_{\mathrm{B}}} (\hat{\mathbf{y}}_\rho^\intercal \hat{\mathbf{y}}_\lambda) \hat{\mathbf{y}}_\lambda^\intercal M^+ \hat{\mathbf{y}}_\rho \ . \tag{29}$$

that background odor directions $\hat{\mathbf{y}}_\rho$ were orthogonal, but if that were the case, m_y would simplify to $\frac{1}{N_{\mathrm{B}}} \sum_\rho \hat{\mathbf{y}}_\rho^\intercal M^+ \hat{\mathbf{y}}_\rho$. Inserting in eq. 25, dividing by N_{B}, and isolating $\underline{v}$ by assuming that the autocorrelation matrix $\underline{\underline{C}}$ is invertible, we have

$$\underline{v} = \underline{\underline{C}}^{-1} \underline{c} - \sigma^2 m_y \underline{\underline{C}}^{-1} \underline{c} + m_y (\underline{v}^\intercal \underline{c}) \underline{\underline{C}}^{-1} \underline{c} \ .$$

This is still an implicit expression because $\underline{v}^\intercal \underline{c}$ appears on the right; taking the dot product of this expression with $\underline{c}$, we can isolate $\underline{v}^\intercal \underline{c}$, then reinsert in the implicit equation for $\underline{v}$ to arrive at an explicit solution (*i.e.*, in terms of $\mathcal{P}$, $\mathcal{Q}$ parameters),

$$\underline{v} = \frac{1 - \sigma^2 m_y}{1 - \gamma m_y} \underline{\underline{C}}^{-1} \underline{c} \tag{30}$$

where we have defined

$$\gamma = \underline{c}^\intercal \underline{\underline{C}}^{-1} \underline{c} \ . \tag{31}$$

B.4. Replacing $\underline{v}$ in the P solution. Having solved for $\underline{v}$, we can put it back in 28 to obtain an explicit solution for P,

$$P = \frac{\sigma^2 - \gamma}{1 - \gamma m_y} \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal M^+ \ . \tag{32}$$

C. Evaluating the loss function at the optimum. For simplicity, we first rewrite the loss function in eq. 24 using the underlined vector notation for $\underline{c}, \underline{v}, \underline{\underline{C}}$, giving

$$\mathcal{L}_{v,P} = N_{\mathrm{B}} \sigma^2 + N_{\mathrm{B}} \underline{v}^\intercal \underline{\underline{C}} \, \underline{v} + \sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho^\intercal P^\intercal P \hat{\mathbf{y}}_\rho + \langle \mathbf{x}^\intercal P^\intercal P \mathbf{x} \rangle - 2 N_{\mathrm{B}} \underline{v}^\intercal \underline{c} - 2\sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho + 2 \underline{v}^\intercal \underline{c} \sum_\rho \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho \ . \tag{33}$$

We need to evaluate multiple terms to insert the solutions for $\underline{v}$ and P in $\mathcal{L}$. We find several simplifications by using the fact that M and thus M^+ and P are symmetric (eqs. 27 and 32), commuting scalars resulting from intermediate dot products, and renaming indices when appropriate. We find the following terms,

$$\underline{v}^\intercal \underline{\underline{C}}\,\underline{v} = \left(\frac{1-\sigma^2 m_y}{1-\gamma m_y}\right)^2 \underline{c}^\intercal \underline{\underline{C}}^{-1}\underline{\underline{C}}\,\underline{\underline{C}}^{-1}\underline{c} = \left(\frac{1-\sigma^2 m_y}{1-\gamma m_y}\right)^2 \gamma$$

$$\sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho^\intercal P^\intercal P \hat{\mathbf{y}}_\rho = \left(\frac{\sigma^2-\gamma}{1-\gamma m_y}\right)^2 \sum_{\lambda,\mu} \hat{\mathbf{y}}_\mu^\intercal M^+ \left\{\sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal\right\} M^+ \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal \hat{\mathbf{y}}_\mu$$

$$\langle \mathbf{x}^\intercal P^\intercal P \mathbf{x}\rangle = \left(\frac{\sigma^2-\gamma}{1-\gamma m_y}\right)^2 \sum_{\lambda,\mu} \hat{\mathbf{y}}_\mu^\intercal M^+ \left\{\langle \mathbf{x}\mathbf{x}^\intercal\rangle\right\} M^+ \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal \hat{\mathbf{y}}_\mu$$

$$\sum_\rho \hat{\mathbf{y}}_\rho^\intercal P \hat{\mathbf{y}}_\rho = \frac{\sigma^2-\gamma}{1-\gamma m_y} \sum_\rho \hat{\mathbf{y}}_\rho^\intercal \sum_\lambda \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal M^+ \hat{\mathbf{y}}_\rho = \frac{\sigma^2-\gamma}{1-\gamma m_y} N_\mathrm{B} m_y \ .$$

The second and third terms can be combined by noticing they have the same form with sums over odor indices λ, μ, and combining the bracketed terms to find $\sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal + \langle \mathbf{x}\mathbf{x}^\intercal\rangle = M$. Then, using the definition of the pseudo-inverse, we have $M^+ M M^+ = M^+$, resulting in

$$\sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho^\intercal P^\intercal P \hat{\mathbf{y}}_\rho + \langle \mathbf{x}^\intercal P^\intercal P \mathbf{x}\rangle = \left(\frac{\sigma^2-\gamma}{1-\gamma m_y}\right)^2 N_\mathrm{B} m_y \ .$$

Combining these expressions in $\mathcal{L}$, we can cancel out a few terms with further algebra and factorize common expressions, finding a surprisingly simple form,

$$\mathcal{L}_{v,P} = N_\mathrm{B} \frac{(\sigma^2-\gamma)(1-\sigma^2 m_y)}{1-\gamma m_y} \ . \tag{34}$$

We notice that the loss seems to scale proportionally with the background subspace dimensions, N_B. Terms σ^2, γ do not depend on N_B but only on the autocorrelation and variance of odor concentrations. The only term that could depend on N_B is $m_y = \frac{1}{N_\mathrm{B}} \sum_{\rho,\lambda=1}^{N_\mathrm{B}} (\hat{\mathbf{y}}_\rho^\intercal \hat{\mathbf{y}}_\lambda) \hat{\mathbf{y}}_\lambda^\intercal M^+ \hat{\mathbf{y}}_\rho$, but we generally expect $N_\mathrm{B} m_y \sim N_\mathrm{B}$. This is especially clear if we assume orthogonality of the $\hat{\mathbf{y}}$, such that $m_y = \frac{1}{N_\mathrm{B}} \sum_{\rho=1}^{N_\mathrm{B}} \hat{\mathbf{y}}_\rho^\intercal M^+ \hat{\mathbf{y}}_\rho \sim \mathcal{O}(1)$.

Hence, there is no obvious tradeoff between the two strategies – predictive filtering and manifold learning – as a function of the background dimension. This makes sense *a posteriori*. Predictive filtering tries to anticipate N_B independent, identically distributed odors, hence the squared errors committed on each background component add up in variance. Meanwhile, the error in manifold learning increase with N_B because a fraction $\sim N_\mathrm{B}$ of the new odor, on average, will lie in the background subspace.

However, there is a tradeoff between the strategies as a function of the autocorrelation time, encoded in the parameter γ, and the dimensionality of the olfactory space, which enters m_y through the new odor statistics $\langle \mathbf{x}\mathbf{x}^\intercal\rangle$ in M. We expect γ to increase with the autocorrelation time scale, and m_y to increase with the olfactory space dimension N_S. Hence, as the autocorrelation time increases, γ diminishes the relative efficacy of manifold learning by reducing the denominator $1-\gamma m_y$, while increasing the importance of predictive filtering by reducing the numerator factor $\sigma^2-\gamma$. This tradeoff will be clearer in the special case studied in section F.

D. Summary of the general optimal solution. We recapitulate the optimization results here. The optimal $\underline{v}$ and P are

$$\underline{v} = \frac{1-\sigma^2 m_y}{1-\gamma m_y} \underline{\underline{C}}^{-1} \underline{c} \tag{30}$$

$$P = \frac{\sigma^2-\gamma}{1-\gamma m_y} \sum_{\rho=1}^{N_\mathrm{B}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal M^+ \tag{32}$$

and they give a minimum loss of

$$\mathcal{L}_{v,P} = N_\mathrm{B} \frac{(\sigma^2-\gamma)(1-\sigma^2 m_y)}{1-\gamma m_y} \tag{34}$$

where N_{B} is the number of i.i.d. background odors, σ^2 is the variance of each odor concentration $\tilde{c}_{\rho,t}$, and where we have defined background parameters

$$\gamma = \underline{c}^\intercal \underline{\underline{C}}^{-1} \underline{c} \tag{31}$$

$$m_y = \frac{1}{N_{\mathrm{B}}} \sum_{\rho,\lambda=1}^{N_{\mathrm{B}}} (\hat{\mathbf{y}}_\rho^\intercal \hat{\mathbf{y}}_\lambda) \hat{\mathbf{y}}_\lambda^\intercal M^+ \hat{\mathbf{y}}_\rho \tag{29}$$

$$\text{in which } M = \sigma^2 \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal + \langle \mathbf{x}\mathbf{x}^\intercal \rangle \tag{27}$$

$$\begin{aligned} \underline{c}_i &= C(i) = \langle \tilde{c}_{\rho,t'} \tilde{c}_{\rho,t' \pm i} \rangle \quad (i \in \{1, 2, \ldots, t-1\}) \\ \underline{\underline{C}}_{ij} &= C(i-j) = \langle \tilde{c}_{\rho,t' \pm i} \tilde{c}_{\rho,t' \pm j} \rangle \end{aligned} \tag{35}$$

E. Limiting cases: $P = 0$ and $\underline{v} = 0$.

E.1. Predictive filtering only: $P = 0$**.** When $P = 0$, we can directly solve eq. 25 for $\underline{v}$, finding

$$\underline{v}_{\,P=0} = \underline{\underline{C}}^{-1} \underline{c}$$

which yields a loss of

$$\mathcal{L}_v = N_{\mathrm{B}}(\sigma^2 - \gamma) \ . \tag{36}$$

As long as $\gamma < \sigma^2$ – which should be the case if the autocorrelation function decays with time – we have $\mathcal{L}_{v,P} < \mathcal{L}_v$ since $\frac{1-\sigma^2 m_y}{1-\gamma m_y} < 1$ in that case.

E.2. Manifold learning only: $\underline{v} = 0$**.** When $\underline{v} = 0$, we can directly solve eq. 26 in terms of M^+, resulting in

$$P_{v=0} = \sigma^2 \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal M^+$$

which yields a loss of

$$\mathcal{L}_P = N_{\mathrm{B}} \sigma^2 (1 - \sigma^2 m_y) \, . \tag{37}$$

As long as $m_y < \frac{1}{\sigma^2}$ – which should be the case since $M^+ \sim 1/\sigma^2$ and m_y is some projection of it on the background subspace – then $\mathcal{L}_{v,P} < \mathcal{L}_P$, since $\frac{1-\gamma/\sigma^2}{1-\gamma m_y} < 1$ in that case.

F. Special case: exponential kernel, $\hat{\mathbf{x}}$ uniform on hypersphere. To make these results more concrete, we now consider a simple case of background statistics where expressions such as γ, m_y, etc. can be computed analytically in terms of interpretable parameters. We consider an exponential autocorrelation function and new odors uniformly sampled on a hypersphere. This is the case plotted in Figs. 1E and S1.

F.1. Exponential autocorrelation kernel, to evaluate γ**.** We suppose that each odor concentration $\tilde{c}_{\rho,t'}$ is independent of other odors and forms a Gaussian process with exponential autocorrelation kernel $C(s) = \langle \tilde{c}_{\rho,t'} \tilde{c}_{\rho,t' \pm s} \rangle = \sigma^2 e^{-|s|/\tau}$ with autocorrelation time τ (*i.e.*, the Ornstein-Uhlenbeck process). A small τ corresponds to fast fluctuations compared to the time scale of learning. In this case, the symmetric Toeplitz matrix $\underline{\underline{C}}_{ij} = \sigma^2 (e^{-1/\tau})^{|i-j|}$ is a Kac-Murdock-Szegö matrix (form $A_{ij} = r^{|i-j|}$, $r \neq 1$), which has an explicit inverse, provided in [74, sec. 1.3]. This inverse is the tridiagonal matrix

$$\underline{\underline{C}}^{-1} = \frac{1}{\sigma^2} \frac{1}{1 - e^{-2\tau}} \begin{pmatrix} 1 & -e^{-1/\tau} & 0 & \ldots & 0 \\ -e^{-1/\tau} & 1 + e^{-2/\tau} & -e^{-1/\tau} & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & -e^{-1/\tau} & 1 + e^{-2/\tau} & -e^{-1/\tau} \\ 0 & \ldots & 0 & -e^{-1/\tau} & 1 \end{pmatrix}$$

It is not hard to check that this is indeed the inverse of $\underline{\underline{C}}$. This allows us to evaluate

$$\underline{\underline{C}}^{-1} \underline{c} = \begin{pmatrix} e^{-1/\tau} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{thus} \quad \gamma = \underline{c}^\intercal \underline{\underline{C}}^{-1} \underline{c} = \sigma^2 e^{-2/\tau}$$

F.2. New odors $\hat{\mathbf{x}}$ uniform, and $\hat{\mathbf{y}}_\rho$ orthogonal, to evaluate m_y. Moreover, we suppose that new odors take the form $\mathbf{x} = \tilde{c}_x \hat{\mathbf{x}}$, where $\tilde{c}_x$ has variance σ^2_{new} possibly different from the background odors, and where $\hat{\mathbf{x}}$ is uniformly sampled on the N_{S}-dimensional unit hypersphere. In that case, by symmetry, the new odor covariance matrix is $\langle \mathbf{x}\mathbf{x}^\intercal \rangle = \frac{\sigma^2_{\mathrm{new}}}{N_{\mathrm{S}}} \mathbb{I}$, with $\mathbb{I}$ the $N_{\mathrm{S}} \times N_{\mathrm{S}}$ identity matrix. Additionally, we assume that background odors are orthogonal to each other: $\hat{\mathbf{y}}_\rho^\intercal \hat{\mathbf{y}}_\lambda = \delta_{\rho\lambda}$.

These choices allow us to compute M^+ explicitly. We first note that $M = \langle \mathbf{x}\mathbf{x}^\intercal \rangle + \sigma^2 \sum_\rho \hat{\mathbf{y}}\hat{\mathbf{y}}_\rho$ has full rank, and thus $M^+ = M^{-1}$. We define the rescaled olfactory dimension

$$\tilde{N}_{\mathrm{S}} = \frac{\sigma^2}{\sigma^2_{\mathrm{new}}} N_{\mathrm{S}} \tag{38}$$

to simplify expressions (equal to N_{S} if the background and new odors have the same concentration variance, $\sigma^2_{\mathrm{new}} = \sigma^2$). We can compute that inverse by repeatedly applying the Sherman-Morrison formula to the sequence of matrices $M_k = \frac{1}{\tilde{N}_{\mathrm{S}}} \mathbb{I} + \sum_{\rho=1}^{k} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal$, $k \leq N_{\mathrm{B}}$. Proceeding by induction, we eventually find the inverse of the full matrix $M^{-1} = \sigma^{-2} M^{-1}_{N_{\mathrm{B}}}$,

$$M^+ = M^{-1} = \frac{\tilde{N}_{\mathrm{S}}}{\sigma^2} \left(\mathbb{I} - \frac{\tilde{N}_{\mathrm{S}}}{\tilde{N}_{\mathrm{S}} + 1} \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal \right) .$$

We can thus evaluate the matrix product appearing in the optimal P solution,

$$\sum_\lambda \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal M^+ = \frac{\tilde{N}_{\mathrm{S}}}{\sigma^2} \sum_\lambda \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal - \frac{\tilde{N}_{\mathrm{S}}^2}{\sigma^2 (\tilde{N}_{\mathrm{S}} + 1)} \sum_\lambda \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal = \frac{\tilde{N}_{\mathrm{S}}}{\tilde{N}_{\mathrm{S}} + 1} \frac{1}{\sigma^2} \sum_\rho \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal ,$$

using $\hat{\mathbf{y}}_\lambda^\intercal \hat{\mathbf{y}}_\rho = \delta_{\lambda\rho}$, which also allows us to compute

$$m_y = \frac{1}{N_{\mathrm{B}}} \sum_{\rho,\mu} \hat{\mathbf{y}}_\rho^\intercal \hat{\mathbf{y}}_\mu \hat{\mathbf{y}}_\mu^\intercal M^+ \hat{\mathbf{y}}_\rho = \frac{\tilde{N}_{\mathrm{S}}}{\tilde{N}_{\mathrm{S}} + 1} \frac{1}{\sigma^2} \frac{1}{N_{\mathrm{B}}} \sum_{\rho,\lambda} \hat{\mathbf{y}}_\rho^\intercal \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\intercal \hat{\mathbf{y}}_\rho = \frac{1}{\sigma^2} \frac{\tilde{N}_{\mathrm{S}}}{\tilde{N}_{\mathrm{S}} + 1} .$$

F.3. Optimal $\underline{v}$, P and loss function in this background choice. Inserting the above expressions for γ, M^+, m_y, etc. into the general optimal solution, we find

$$v_1 = \frac{1}{1 + \tilde{N}_{\mathrm{S}}(1 - e^{-2/\tau})} e^{-1/\tau} \;\; ; \quad v_{j>1} = 0$$
$$P = \frac{\tilde{N}_{\mathrm{S}}(1 - e^{-2/\tau})}{1 + \tilde{N}_{\mathrm{S}}(1 - e^{-2/\tau})} \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal$$

and a minimum loss of

$$\mathcal{L}_{v,P} = N_{\mathrm{B}} \sigma^2 \frac{1 - e^{-2/\tau}}{1 + \tilde{N}_{\mathrm{S}}(1 - e^{-2/\tau})} \tag{39}$$

Here, we see clearly that there is no tradeoff as a function of N_{B}: both strategies have an error that increases proportionally to N_{B}. At least, we clearly see the transition from predictive filtering to manifold learning as $\tilde{N}_{\mathrm{S}}$ decreases or τ decreases. For small correlation times, $1 - e^{-2/\tau} \approx 1$, so the main reduction of the loss comes from the $\frac{1}{\tilde{N}_{\mathrm{S}}+1}$ factor (also the case if $\tilde{N}_{\mathrm{S}}$ is large). Meanwhile, for large correlation τ, the reduction comes from the numerator $1 - e^{-2/\tau} \approx 0$, while the $\tilde{N}_{\mathrm{S}}$ term in the denominator is rendered ineffective – the same is true if $\tilde{N}_{\mathrm{S}}$ is small.

F.4. Special cases $P = 0$ and $\underline{v} = 0$ in this background choice. For further comparison, the optimal solutions and loss function in the pure predictive filtering case ($P = 0$) with our specific background choice are:

$$v_{1,P=0} = e^{-1/\tau} \;\; ; \quad v_{j\geq 2,P=0} = 0$$
$$\mathcal{L}_v = N_{\mathrm{B}} \sigma^2 (1 - e^{-2/\tau}) \; . \tag{40}$$

In the pure manifold learning case ($\underline{v} = 0$), these are rather

$$P_{v=0} = \sigma^2 \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal M^+ = \frac{\tilde{N}_{\mathrm{S}}}{\tilde{N}_{\mathrm{S}} + 1} \sum_{\rho=1}^{N_{\mathrm{B}}} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\intercal \tag{41}$$

$$\mathcal{L}_P = \frac{N_{\mathrm{B}} \sigma^2}{\tilde{N}_{\mathrm{S}} + 1} \; . \tag{42}$$

F.5. Plots of the loss function versus $\tilde{N}_{\mathrm{S}}$ and τ for the different strategies. We notice that the loss is proportional to $N_{\mathrm{B}}\sigma^2$ in all strategies for our special background choice, hence we can illustrate the relative efficacy of predictive filtering and manifold learning by plotting $\mathcal{L}/(N_{\mathrm{B}}\sigma^2)$ as a function of $\tilde{N}_{\mathrm{S}}$ and τ, the two background parameters for which there is actually a transition between the two strategies. Fig. S1A-B shows the single-strategy losses $\mathcal{L}_v$ (eq. 40) and $\mathcal{L}_P$ (eq. 42) compared to the loss for both strategies applied simultaneously, $\mathcal{L}_{v,P}$ (eq. 39).

We see that for even a small olfactory space dimension, $\tilde{N}_{\mathrm{S}} = 50$ as in the fruit fly, manifold learning performs much better than predictive filtering even for moderately long correlation time scales, as discussed in the main text. When $\mathcal{L}_{v,P}$ is close to either $\mathcal{L}_P$ or $\mathcal{L}_v$, it means that the corresponding strategy contributes most of the loss reduction. We can thus draw a "phase" diagram of where habituation is dominated by manifold learning (large $\tilde{N}_{\mathrm{S}}$, small τ) or by predictive filtering (small $\tilde{N}_{\mathrm{S}}$, large τ). The (smooth) transition occurs when $\mathcal{L}_P = \mathcal{L}_v$, which happens at $1 - e^{-2/\tau} = \frac{1}{\tilde{N}_{\mathrm{S}}+1}$, which is at $\tau \approx 2(\tilde{N}_{\mathrm{S}} + 1)$ for large $\tilde{N}_{\mathrm{S}}$. This is shown in Fig. 1E. There is a large region (in red) where manifold learning is the most effective strategy.

2. Simulating background odor fluctuations

A. Turbulent statistics. We use the statistics of whiff durations $p_{t_{\mathrm{w}}}(t_{\mathrm{w}})$, blank durations $p_{t_{\mathrm{b}}}(t_{\mathrm{b}})$, and whiff concentrations $p_c(c)$ derived in [9], with the exponents and statistics for an atmospheric boundary layer. We simulate each background odor concentration as a telegraph-like process, as illustrated in Fig. 1, by drawing a next blank duration at the end of a whiff, a next whiff duration at the end of a blank, and a concentration for each whiff from p_c. As the simulation advances per time steps $\Delta t = 10$ ms, we keep track of how much time is left in the current whiff or blank as well as of the current concentration, and update when that time runs out. This method neglects intra-whiff concentration fluctuations, as well as correlations between successive whiff and blank durations, but captures the main challenges of varying whiff concentrations and power-law (long-tailed) distributions of durations. We now describe these distributions and how we numerically sample from them.

For the durations of whiffs and blanks, the distribution is a power law with exponent $-3/2$ and a lower cutoff at τ_{b}, τ_{w}. For numerical stability, we prevent abnormally large durations by imposing also an upper cutoff at $T_{\mathrm{max,b}}$, $T_{\mathrm{max,w}}$, and we use sharp cutoffs, corresponding to a probability density function

$$p_{t_{\mathrm{x}}}(t_{\mathrm{x}}) = \begin{cases} 0 & \text{if } t_{\mathrm{x}} < \tau_{\mathrm{x}} \text{ or } t_{\mathrm{x}} > T_{\mathrm{max,x}} \\ \dfrac{1}{A_{\mathrm{x}}}\left(\dfrac{t_{\mathrm{x}}}{\tau_{\mathrm{x}}}\right)^{-3/2} & \text{if } \tau_{\mathrm{x}} \leq t_{\mathrm{x}} \leq T_{\mathrm{max,x}} \end{cases} \qquad (\mathrm{x} = \mathrm{b \text{ or } w}) \ , \tag{43}$$

where $A_{\mathrm{x}} = 2\tau_{\mathrm{x}}\left(1 - (T_{\mathrm{max,x}}/\tau_{\mathrm{x}})^{-1/2}\right)$ is a normalization constant. From these distributions, the average duration of a whiff or blank is the geometric average of the limits, since

$$\langle t_{\mathrm{x}} \rangle = \int_{\tau_{\mathrm{x}}}^{T_{\mathrm{max,x}}} \left(\frac{t}{\tau_{\mathrm{x}}}\right)^{-3/2} \frac{t}{A_{\mathrm{x}}} \mathrm{d}t = \sqrt{\tau_{\mathrm{x}} T_{\mathrm{max,x}}} \quad (\mathrm{x} = \mathrm{b \text{ or } w}) \ ,$$

so the probability χ to be in a whiff is

$$\chi = \frac{\langle t_{\mathrm{w}} \rangle}{\langle t_{\mathrm{w}} \rangle + \langle t_{\mathrm{b}} \rangle} = \left(1 + \sqrt{\frac{\tau_{\mathrm{b}} T_{\mathrm{max,b}}}{\tau_{\mathrm{w}} T_{\mathrm{max,w}}}}\right)^{-1} .$$

The probability distribution of whiff concentrations c is therefore 0 with probability $1 - \chi$ (illustrated by the point at $c = 0$ in Fig. 1C, left) or, with probability χ, the conditional distribution p_c given there is a whiff. Hence, $p_c(c) = (1 - \chi)\delta(c) + \chi p_c(c|\mathrm{whiff})$. The conditional distribution has a tail $p_c \sim e^{-c/c_0}/c$, where c_0 is a typical concentration scale, and a probability plateau near $c = 0$. As illustrated in Fig. 1C, left, we use a sharp transition at $\alpha_c c_0$ for some $\alpha_c < 1$, with a uniform probability on the range below, $(0, \alpha_c c_0]$, corresponding to a probability density function

$$p_c(c|\mathrm{whiff}) = \begin{cases} \dfrac{1}{A_\alpha} \dfrac{e^{-c/c_0}}{c} & \text{if } c \geq \alpha_c c_0 \\ \dfrac{1}{A_\alpha} \dfrac{e^{-\alpha_c}}{\alpha_c c_0} & \text{if } c < \alpha_c c_0 \end{cases} , \tag{44}$$

where $A_\alpha = e^{-\alpha_c} + E_1(\alpha_c)$ is a normalization constant and E_1 is the exponential integral,

$$E_1(x) = \int_x^\infty \mathrm{d}u \, \frac{e^{-u}}{u} \ .$$

Of note, the average whiff concentration then has the analytical expression

$$\langle c \rangle_{\mathrm{whiff}} = \int_0^\infty \mathrm{d}c \, c \, p_c(c|\mathrm{whiff}) = \frac{(1 + \alpha_c/2)c_0 e^{-\alpha_c}}{A_\alpha} \ .$$

To sample from these distributions during a simulation, we use the inverse transform method: given a random $\mathrm{uniform}(0,1)$ sample r, we generate a sample of a random variable X following the cumulative distribution function (cdf) F_X as $x = F_X^{-1}(r)$. The cdf for the whiff or blank durations is

$$F_t(t_\mathrm{x}) = \begin{cases} 0 & \text{if } t_\mathrm{x} \leq \tau_\mathrm{x} \\ \dfrac{\left(1-(t_\mathrm{x}/\tau_\mathrm{x})^{-1/2}\right)}{1-(T_{\max,\mathrm{x}}/\tau_\mathrm{x})^{-1/2}} & \text{if } \tau_\mathrm{x} < t_\mathrm{x} < T_{\max,\mathrm{x}} \\ 1 & \text{if } T_{\max,\mathrm{x}} \leq t_\mathrm{x} \end{cases} \qquad (\mathrm{x} = \mathrm{b \text{ or } w}) \ .$$

Taking the inverse, we generate t_x from uniform samples r as

$$t_\mathrm{x} = \frac{\tau_\mathrm{x}}{\left[1 - r\left(1-(T_{\max,\mathrm{x}}/\tau_\mathrm{x})^{-1/2}\right)\right]^2} \qquad (\mathrm{x} = \mathrm{b \text{ or } w}) \ . \tag{45}$$

As a check, notice that $t_\mathrm{x} = \tau_\mathrm{x}$, the lower cutoff, when $r = 0$, and $t_\mathrm{x} = T_{\max,\mathrm{x}}$, the upper cutoff, when $r = 1$.

The cdf for the conditional whiff concentrations is

$$F_c(c|\mathrm{whiff}) = \begin{cases} 0 & \text{if } c < 0 \\ \dfrac{c}{\alpha_c A_\alpha c_0 e^{\alpha_c}} & \text{if } 0 \leq c \leq \alpha_c c_0 \\ 1 - \frac{1}{A_\alpha} E_1(c/c_0) & \text{if } c > \alpha_c c_0 \end{cases} \ .$$

Hence, given a random uniform sample r, we generate a sample c as

$$c = F_c^{-1}(r) = \begin{cases} \alpha_c A_\alpha c_0 e^{\alpha_c} r & \text{if } r \leq F_c(\alpha_c c_0) = \frac{1}{e^{\alpha_c} A_\alpha} \\ c_0 E_1^{-1}(A_\alpha(1-r)) & \text{if } \frac{1}{e^{\alpha_c} A_\alpha} < r < 1 \end{cases} \tag{46}$$

where E_1^{-1} is the inverse exponential integral. This inverse function does not have an analytical closed form, so we evaluate $y = E_1^{-1}(x)$ at a given x numerically by solving the equation $E_1(y) - x = 0$ for y, using Brent's method [75] with suitable bounds on the solution. For numerical accuracy, for $x < 1$, we solve in log scale, $\log(E_1(y)) - \log(x) = 0$, to expand the range of $E_1(y)$ values. For larger x, y becomes very small (*e.g.*, $E_1(2) = 0.0489$), so we solve for $z = \log(y)$. For $x > 30$, y is small enough ($y \sim 10^{-13}$) to use the approximation $E_1(y) = -\gamma - \log(y) - \mathcal{O}(y)$ [76, 6.6.1], where $\gamma = 0.577\ldots$ is the Euler-Mascheroni constant, so the equation is inverted directly: $y = e^{-\gamma - x}$.

Hence, overall, for each update when a whiff starts, we use two random $\mathrm{uniform}(0,1)$ samples, one for t_w (using Eq. (45)), one for c (using Eq. (46)); only one sample is needed when a blank starts, for t_w (Eq. (45)). In our simulations, we use the following typical parameter values, the same for all background odors: $\tau_\mathrm{b} = \tau_\mathrm{w} = 10\,\mathrm{ms} = 1$ time step for the lower whiff or blank duration cutoff, $T_{\max,\mathrm{w}} = 5000\,\mathrm{ms}$ and $T_{\max,\mathrm{b}} = 8000\,\mathrm{ms}$ for the maximum whiff and blank durations respectively, $c_0 = 0.6$ for the (arbitrary) concentration scale, and $\alpha_c = 0.5$ for the lower whiff concentration cutoff.

Moreover, to sample a concentration from the stationary distribution, we draw a first $\mathrm{uniform}(0,1)$ sample r_1 to determine whether there is a whiff, with probability χ (whiff if $r \leq \chi$), or a blank ($c = 0$). Then, if in a whiff, draw a second uniform sample r_2 to generate a concentration c using Eq. (46).

B. Univariate Ornstein-Uhlenbeck process. To simulate a univariate Ornstein-Uhlenbeck (O-U) process $\overline{\nu}(t)$ numerically, we use an exact update rule for finite time steps Δt, derived from the analytical solution of the O-U process. Taking the last time step as a new deterministic initial condition [77, eq. 2.47],

$$\overline{\nu}(t+\Delta t) = \overline{\nu}(t) e^{-\Delta t/\tau_\mathrm{b}} + \sqrt{\sigma^2\left(1 - e^{-2\Delta t/\tau_\mathrm{b}}\right)}\,\xi(t) \tag{47}$$

where $\xi(t)$ is white noise. The coefficients of $\overline{\nu}(t)$ and $\xi(t)$ can be computed in advance. This rule ensures a steady-state distribution of $\overline{\nu}$ with the desired variance σ^2 even when the simulation time step is on the order of τ_b.

C. Multivariate Ornstein-Uhlenbeck process. The multivariate Langevin equation for the Ornstein-Uhlenbeck process with zero stationary mean, $\overline{\boldsymbol{\nu}}(t)$, is covered in [78, sec. 4.5.6], and can be simulated exactly using the same trick as in the univariate case, Eq. (47). In practice, we used independent and identically distributed background odors in this paper, so the matrices A and B, were diagonal, and the general simulation method effectively reduced to simulating N_B zero-mean univariate processes in parallel (using Eq. (47)), then adding the desired mean vector $\overline{\boldsymbol{\nu}}_0$ to it.

D. Weakly non-Gaussian fluctuating background. We simulate a multivariate Ornstein-Uhlenbeck process, $\mathbf{g}$ as in section 2C, with zero mean and identically distributed variables with stationary variance σ_g^2. Then, we take $c_\gamma = g_0 + g_\gamma + \epsilon g_\gamma^2$, where ϵ should be chosen small and g_0 is the desired zeroths-order mean concentration. Then, the concentrations have the following moments:

$$\begin{aligned} \langle c \rangle &= g_0 + \epsilon \sigma_g^2 \\ \mathrm{Var}\left[c\right] &= \sigma_g^2 + 2\epsilon^2 \sigma_g^4 \\ \langle (c - \langle c \rangle)^3 \rangle &= 6\epsilon \sigma_g^4 + 8\epsilon^3 \sigma_g^6 \end{aligned}$$

These results are straightforward to obtain by expanding c^2 and c^3 and using higher moments of the Gaussian distribution as required, *i.e.*, $\langle g_\gamma^{2k} \rangle = \frac{(2k)!}{2^k k!} \sigma_g^{2k}$. The important outcome is that the third moment is of order ϵ.

E. Log-normal background fluctuations. As above, we simulate identically distributed O-U variables with variance σ_g^2, add a mean g_0 to them, then use them as the $\log_{10}$ of the concentrations, transforming them according to $c_\gamma = 10^{g_\gamma + g_0}$. Then, from the log-normal distribution properties [79], the concentrations themselves have a log-normal distribution with moments

$$\begin{aligned} \langle c \rangle &= 10^{g_0 + \frac{1}{2}\sigma_g^2 \ln 10} \\ \mathrm{Var}\left[c\right] &= \left(10^{\sigma_g^2 \ln 10} - 1\right) 10^{2g_0 + \sigma_g^2 \ln 10} \\ \langle (c - \langle c \rangle)^3 \rangle &= \mathrm{Var}\left[c\right]^{3/2} \left(10^{\sigma_g^2 \ln 10} + 2\right) \sqrt{10^{\sigma_g^2 \ln 10} - 1} \ . \end{aligned}$$

We test habituation to this background with IBCM and BioPCA networks in Fig. S6.

F. Numerical stability. Numerical integration of the W equations displays instabilities when $\overline{\mathbf{h}}$ reaches large magnitudes, *e.g.*, when increasing the scale parameter Λ (Fig. S11 and section 8). To ensure these are numerical errors rather than a true dynamical instability of the fixed points, we perform a nonlinear numerical stability analysis of the Euler integrator. This integrator, applied to the W matrix equation, is effectively a discrete mapping,

$$\begin{aligned} W_{t+1} &= W_t + \Delta t(\alpha \mathbf{y}\overline{\mathbf{h}}^\intercal - \beta W_t) = W_t + \Delta t(\alpha \mathbf{s}\overline{\mathbf{h}}^\intercal - \alpha W_t \overline{\mathbf{h}}\overline{\mathbf{h}}^\intercal - \beta W_t) \\ &= B\Delta t + W_t(\mathbb{I} + \Delta t J) \quad \text{where } B = \alpha \mathbf{s}\overline{\mathbf{h}}^\intercal \ , \ J = -\alpha \overline{\mathbf{h}}\overline{\mathbf{h}}^\intercal - \beta \mathbb{I} \end{aligned}$$

We consider the worst-case scenario, when $\overline{\mathbf{h}}$ reaches its maximal magnitude encountered in a simulation, $\overline{\mathbf{h}} = \mathbf{h}_{\max}$ and $J_{\max} = -\alpha \mathbf{h}_{\max}\mathbf{h}_{\max}^\intercal - \beta \mathbb{I}$, and iterate the map in this case, which gives

$$W_{t+N} = \delta t B \sum_{n=0}^{N-1} (\mathbb{I} + \Delta t J_{\max})^n + W_t(\mathbb{I} + \Delta t J_{\max})^N$$

Consequently, the stability depends on the eigenvalues of $A = \mathbb{I} + \Delta t J_{\max}$, but not on $\Delta t B$ (because the latter is not raised to some power). Indeed, first note that the matrix A is symmetric and diagonalizable as $A = UDU^\dagger$ with $D = \mathrm{diag}(\lambda_i)$. Then, if A has at least one eigenvalue with magnitude $|\lambda_j| > 1$, then $A^N = U\mathrm{diag}(\lambda_i^N)U^\dagger$ will have a diverging component as the mapping is iterated (as N increases). To find the threshold where this happens, we can read out the eigenvalues from the expression of $A = (1 - \beta\Delta t)\mathbb{I} - \alpha\Delta t \|\mathbf{h}_{\max}\|^2 \hat{\mathbf{h}}_{\max}\hat{\mathbf{h}}_{\max}^\intercal$,

$$\lambda_i = \begin{cases} 1 - \beta\Delta t - \|\mathbf{h}_{\max}\|^2 \alpha \Delta t & \text{once} \\ 1 - \beta \Delta t & N_{\mathrm{I}} - 1 \text{ times} \end{cases}$$

To see this, consider the rotation matrix R that aligns $\mathbf{h}_{\max}$ with one of the unit vectors, *e.g.*, $(1, 0, \ldots 0)$: $RAR^\intercal$ is then diagonal with $\lambda_1 = 1 - \beta\Delta t - \|\mathbf{h}_{\max}\|^2 \alpha \Delta t$ in one row, $\lambda_2 = 1 - \beta\Delta t$ on the others. For large LN activity, λ_1 is first to reach a magnitude > 1, by becoming negative. Hence, we can predict that numerical instabilities arise in the Euler integrator when

$$\lambda_1 < -1 \Rightarrow \|\mathbf{h}_{\max}\|^2 > \frac{2 - \beta\Delta t}{\alpha \Delta t} \quad \text{or} \quad \Delta t > \frac{2}{\beta + \alpha \|\mathbf{h}_{\max}\|^2} \ . \tag{48}$$

Given a simulation of the M weights and $\overline{\mathbf{h}}$ (which are not destabilized), we can thus anticipate whether the W integration should be unstable by extracting $\|h_{\max}\|$ from the simulation. The vertical lines in Fig. S11 indicate the smallest Λ value for which this threshold is reached, in each model, in the turbulent background considered. They coincide well with the observed drops in model performance, confirming these are due to Euler integrator instabilities. A linear stability analysis of the W equations (computing the Jacobian of the ODE near the fixed point, etc.) confirms that the W is linearly stable for any Λ. Hence, the observed divergences are numerical limitations rather than true model performance drops and would be remedied by decreasing the time step, Δt.

3. Average subtraction model

In this short section, we examine the average subtraction model from [36], and explain how it is insufficient against fluctuating backgrounds. In our notation, it corresponds to a vector $\mathbf{w}$ of inhibitory weights learned as

$$\frac{d\mathbf{w}}{dt} = \alpha\mathbf{s}(t) - (\alpha+\beta)\mathbf{w}(t) \tag{49}$$

and a PN response $\mathbf{y}(t) = \mathbf{s}(t) - \mathbf{w}$ (the LN activity is fixed to 1). If this network is exposed to a constant background odor $\mathbf{s}_{\mathrm{b},0}$, the inhibition vector $\mathbf{w}$ converges to $\mathbf{w}_{\mathrm{ss}} = \frac{\alpha}{\alpha+\beta}\mathbf{s}_{\mathrm{b},0}$, so the background is then perfectly subtracted. However, this strategy fails if the background vector $\mathbf{s}_{\mathrm{b}}(t)$ fluctuates randomly over time. In this case, equation 49 amounts to computing the average background over a time window of duration $\frac{1}{\alpha+\beta}$, as seen from the formal solution of the stochastic equation,

$$\mathbf{w}(t) = \alpha \int_0^t dt' e^{-(\alpha+\beta)(t-t')}\mathbf{s}_{\mathrm{b}}(t') \ ,$$

assuming $\mathbf{s}_{\mathrm{b}}(t)$ started at a given initial value $\mathbf{s}_{\mathrm{b},0}$. From the formal solution, assuming the background is a stationary process, the steady-state average value of $\mathbf{w}$ is therefore proportional to the average background, $\langle\mathbf{w}\rangle = \frac{\alpha}{\alpha+\beta}\langle\mathbf{s}_{\mathrm{b}}\rangle$. If the background process has an autocorrelation time scale much faster than the learning rate, $\tau_{\mathrm{b}} \ll \frac{1}{\alpha}$, with its elements approximately obeying $\langle\Delta s_i(t_1)\Delta s_i(t_2)\rangle = \sigma_{s_i}^2 e^{-|t_2-t_1|/\tau_{\mathrm{b}}}$, then to leading order, the inhibitory weights have a small variance

$$\mathrm{Var}\left[w_i\right] = \alpha\tau_{\mathrm{b}}\frac{\alpha}{\alpha+\beta}\sigma_{s_i}^2 \ .$$

Hence, the inhibitory weights do not fluctuate much around the constant average background, $\langle s_{\mathrm{b}}\rangle$, since the factor $\alpha\tau_{\mathrm{b}} \ll 1$. Therefore, this model computes the average background, scaled by $\frac{\alpha}{\alpha+\beta}$, and subtracts this fixed quantity from $\mathbf{s}_{\mathrm{b}}(t)$ to obtain the projection neuron activity $\mathbf{y}(t) = \mathbf{s}_{\mathrm{b}}(t) - \mathbf{w}$. Consequently, the variance of the PN activity, $\mathbf{y}(t)$, is not reduced compared to the variance of the background:

$$\begin{aligned} \langle\mathbf{y}\rangle = \langle\mathbf{s}_{\mathrm{b}}\rangle - \langle\mathbf{w}\rangle = \frac{\beta}{\alpha+\beta}\langle\mathbf{s}_{\mathrm{b}}\rangle \quad &\text{(reduction of the average)} \\ \text{but} \ \ \mathrm{Var}\left[s_i\right] \sim \sigma_{s_i}^2 \ \ \forall i \quad &\text{(no reduction of the fluctuations)} \ . \end{aligned} \tag{50}$$

These large remaining fluctuations in PN activity mix with new odors appearing in the landscape and thus hinder their recognition; this effect explains why the average subtraction model does not provide a significant reduction in PN response to the background, or improvement in new odor recognition compared to the absence of habituation in Figures 2, 5, S2, and others.

4. Analytical solution of the IBCM model's average fixed points

To understand what projections are learned by IBCM neurons, we derive analytical expressions for the synaptic weights $\mathbf{m}$ of an IBCM neuron at stationary state. We first establish approximate fixed point equations for these weights averaged over fast fluctuations of the background process $\mathbf{s}_{\mathrm{b}}(t)$, assuming perfect separation of time scales between $\mathbf{s}$, Θ, and $\mathbf{m}$. Then, we obtain exact solutions to these approximate equations.

A. Establishing the IBCM fixed point equations. Let's first recall the stochastic differential equations describing the synaptic weights learning of an IBCM neuron in a network with feedforward lateral coupling (Methods). For tractability, we assume the activation function ϕ is the identity function (instead of a nonlinearity like $\tanh$), such that $\overline{\mathbf{h}} = LM\mathbf{s}$. We also set the decay term $-\varepsilon\mu\mathbf{m}_i$ to zero. Then, for each neuron i,

$$\frac{\mathrm{d}\mathbf{m}_i}{\mathrm{d}t} = \mu_{\overline{\Theta}_i}\overline{h}_i\left(\overline{h}_i - \overline{\Theta}_i\right)\mathbf{s}(t) - \eta\sum_{j\neq i}\mu_{\overline{\Theta}_j}\overline{h}_j\left(\overline{h}_j - \overline{\Theta}_j\right)\mathbf{s}(t) \tag{51}$$

$$\frac{\mathrm{d}\overline{\Theta}_i}{\mathrm{d}t} = \frac{1}{\tau_\Theta}((\overline{h}_i)^2 - \overline{\Theta}_i) \ . \tag{52}$$

The reduced activity of interneuron i is $\overline{h}_i = \overline{\mathbf{m}}_i \cdot \mathbf{s}(t)$, where the reduced weights $\overline{\mathbf{m}}_i = \mathbf{m}_i - \eta\sum_{j\neq i}\mathbf{m}_j$. Hence, $\overline{h}_i$ varies in time on two separate scales: rapidly with sensory inputs $\mathbf{s}(t)$ fluctuating on time scale τ_{b}, and gradually as the synaptic weights $\overline{\mathbf{m}}_i$ are learned.

To make analytical progress, we focus on averages over fast background fluctuations, denoted by brackets $\langle\cdot\rangle$, while the slow dynamical variables (M, $\overline{\Theta}$) remain unchanged. Moreover, we make a quasi-static approximation on the thresholds: we assume τ_Θ is slow enough to average over background fluctuations, yet fast enough that $\overline{m}_i$ remains unchanged over the averaging time window (*i.e.*, we neglect correlations between $\overline{m}$ and $\overline{\Theta}$). Hence, as in the original IBCM models [43], we assume a perfect separation between the background, threshold, and synaptic weight time scales: $\tau_{\mathrm{b}} \ll \tau_{\Theta} \ll \frac{1}{\mu}$.

Thus, averaging equation 52, and setting $\frac{\mathrm{d}\langle\overline{\Theta}_i\rangle}{\mathrm{d}t} = 0$, we find

$$\langle\overline{\Theta}_i\rangle = \langle\overline{h}_i^2\rangle$$

and we replace $\overline{\Theta}_i$ in the $\mathbf{m}$ equation 51 with this average. Averaging that equation as well, and setting $\frac{\mathrm{d}\langle\overline{\mathbf{m}}_i\rangle}{\mathrm{d}t} = 0$, we find the IBCM fixed point equations:

$$0 = \mu_{\langle\overline{\Theta}_i\rangle} \left\langle \overline{h}_i \left(\overline{h}_i - \langle\overline{\Theta}_i\rangle\right) \mathbf{s}(t)\right\rangle - \eta \sum_{j \neq i} \mu_{\langle\overline{\Theta}_j\rangle} \left\langle \overline{h}_j \left(\overline{h}_j - \langle\overline{\Theta}_j\rangle\right) \mathbf{s}(t)\right\rangle \quad i \in \{1, 2, \ldots, N_{\mathrm{I}}\} \ . \tag{53}$$

Defining terms $\varphi_i = \mu_{\langle\overline{\Theta}_i\rangle} \left\langle \overline{h}_i \left(\overline{h}_i - \langle\overline{\Theta}_i\rangle\right) \mathbf{s}(t)\right\rangle$ and combining them in a N_{I}-dimensional vector φ, this system of equations can be written in matrix form

$$0 = L\varphi$$

where L is the $N_{\mathrm{I}} \times N_{\mathrm{I}}$ matrix of feedforward inhibitory coupling between interneurons, with 1 on the diagonal and $-\eta$ everywhere else. This L is a circulant matrix (each row is a cyclic permutation of the previous row by one element to the right); hence, except in the pathological cases $\eta = -1$ or $\eta = \frac{1}{N_{\mathrm{I}}-1}$ (for which some of its eigenvalues are zero), L is invertible and the unique solution is $\varphi = L^{-1}0 = 0$. Therefore, in general, the fixed points of the IBCM network are found by setting each φ_i term to zero individually

$$0 = \mu_{\langle\overline{\Theta}_i\rangle} \left\langle \overline{h}_i \left(\overline{h}_i - \langle\overline{\Theta}_i\rangle\right) \mathbf{s}(t)\right\rangle \ \forall i \in \{1, 2, \ldots, N_{\mathrm{I}}\} \ . \tag{54}$$

Hence, in terms of the reduced synaptic weights $\overline{\mathbf{m}}_i$ and activities $\overline{h}_i$, the fixed point equations for the network of IBCM neurons decouple to take the same form as that of a single IBCM neuron.

Before proceeding to solve this set of equations, we note that the actual synaptic weights $\mathbf{m}_i$ can be found from the $\overline{\mathbf{m}}_i$ solutions by inverting the matrix equation

$$\overline{M} = LM \Rightarrow M = L^{-1}\overline{M}$$

where $\overline{M}$ contains the reduced $\overline{\mathbf{m}}_i$ in its rows. From the eigenvectors and eigenvalues of the circulant matrix L, the inverse matrix elements are

$$L^{-1}_{ij} = \begin{cases} \frac{(N_{\mathrm{I}}-2)\eta-1}{(N_{\mathrm{I}}-1)\eta^2+(N_{\mathrm{I}}-2)\eta-1} & \text{on the diagonal} \\ \frac{-\eta}{(N_{\mathrm{I}}-1)\eta^2+(N_{\mathrm{I}}-2)\eta-1} & \text{off-diagonal} \end{cases}$$

so the synaptic weights $\mathbf{m}_i$ can be recovered from the reduced ones, if necessary, as

$$\mathbf{m}_i = \frac{[(N_{\mathrm{I}}-2)\eta - 1]\overline{\mathbf{m}}_i - \eta \sum_{k \neq j} \overline{\mathbf{m}}_k}{(N_{\mathrm{I}}-1)\eta^2 + (N_{\mathrm{I}}-2)\eta - 1} \quad \text{if } \eta \neq \frac{1}{N_{\mathrm{I}}-1} \text{ or } -1 \ . \tag{55}$$

B. IBCM fixed point equations for i.i.d. background concentrations. We will express the solutions to the fixed point equations in terms of the alignments, or dot products, of the IBCM synaptic weights with the background odors,

$$\overline{h}_{i\gamma} = \overline{\mathbf{m}}_i \cdot \hat{\mathbf{s}}_\gamma \ . \tag{56}$$

and we imply these are averaged over fast background fluctuations (*i.e.*, we really look at $\langle\overline{\mathbf{m}}_i\rangle \cdot \hat{\mathbf{s}}_\gamma$), The odor vectors $\hat{\mathbf{s}}_\gamma$ do not have to be orthogonal, but we assume they form a linearly independent set. Specifying these dot products is sufficient to obtain the complete solution, since the IBCM dynamics only update weights in the background subspace of $\mathbf{s}(t)$ (eq. 51). The other directions are reduced to zero by the slow decay term, $-\delta\mu\mathbf{m}_i$, that we have added to the full dynamics (Methods). For the rest of this section, we work with reduced variables and drop the overlines on $\overline{\Theta}_i$, $\overline{\mathbf{m}}_i$, and $\overline{h}_{i\gamma}$ to simplify notation. We also drop the IBCM neuron index i since the fixed point equations eq. 54 are identical and solved independently for each neuron.

Since we have decoupled the Θ and $\mathbf{m}$ fluctuations in Eq. (53), we can divide by the learning rate and write

$$0 = \langle h(t)^2 \mathbf{s}(t)\rangle - \langle\Theta\rangle \langle h\mathbf{s}(t)\rangle$$

We replace $\langle\Theta\rangle = \langle h^2\rangle$, using our quasi-static approximation. We now assume that the odor concentrations $h_\gamma(t)$ are independent, identically distributed stationary processes. We write $h_\gamma(t) = \langle c\rangle + \tilde{c}_\gamma(t)$ where $\langle\tilde{c}_\gamma\rangle = 0$. The concentrations have mean $\langle c\rangle$, variance $\langle\tilde{c}_\gamma^2\rangle = \sigma^2$, and third moment $\langle\tilde{c}^3\rangle = m_3$. We also let $\mathbf{s}(t) = \mathbf{s}_{\mathrm{d}} + \sum_{\gamma=1}^{N_{\mathrm{B}}} c_\gamma(t)\hat{\mathbf{s}}_\gamma$ and $h(t) = h_{\mathrm{d}} + \sum_\gamma h_\gamma \tilde{c}_\gamma(t)$, where we have defined $\mathbf{s}_\mathrm{d} = \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$ and $h_\mathrm{d} = \mathbf{m} \cdot \mathbf{s}_\mathrm{d}$. We proceed to compute the averages appearing in the fixed point equation:

$$\begin{aligned}
\langle h(t)^2 \rangle &= \langle (\mathbf{m} \cdot \mathbf{s})^2 \rangle = \sum_{\gamma,\rho} \langle c_\gamma(t) c_\rho(t) \rangle \, h_\gamma h_\rho = h_\mathrm{d}^2 + \sigma^2 u^2 \\
\langle h(t) \mathbf{s}(t) \rangle &= h_\mathrm{d} \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma + \sum_{\gamma,\rho} \langle \tilde{c}_\gamma \tilde{c}_\rho \rangle \, h_\rho \hat{\mathbf{s}}_\gamma = \sum_\gamma (\langle c \rangle \, h_\mathrm{d} + \sigma^2 h_\gamma) \hat{\mathbf{s}}_\gamma \\
\langle h(t)^2 \mathbf{s}(t) \rangle &= h_\mathrm{d}^2 \mathbf{s}_\mathrm{d} + 2 h_\mathrm{d} \sum_{\rho\gamma} \langle \tilde{c}_\rho \tilde{c}_\gamma \rangle \, h_\rho \hat{\mathbf{s}}_\gamma + \mathbf{s}_\mathrm{d} \sum_{\rho,\gamma} h_\rho h_\gamma \langle \tilde{c}_\rho \tilde{c}_\gamma \rangle + \sum_{\rho,\lambda,\gamma} \langle \tilde{c}_\rho \tilde{c}_\lambda \tilde{c}_\gamma \rangle \, h_\rho h_\lambda \hat{\mathbf{s}}_\gamma \\
&= h_\mathrm{d}^2 \mathbf{s}_\mathrm{d} + 2 h_\mathrm{d} \sigma^2 \sum_\gamma h_\gamma \hat{\mathbf{s}}_\gamma + \sigma^2 u^2 \mathbf{s}_\mathrm{d} + m_3 \sum_\gamma h_\gamma^2 \hat{\mathbf{s}}_\gamma
\end{aligned}$$

where we have defined

$$h_\mathrm{d} = \mathbf{m} \cdot \mathbf{s}_\mathrm{d} = \langle c \rangle \sum_\gamma h_\gamma \quad \text{and} \quad u^2 = \sum_{\gamma=1}^{N_\mathrm{B}} h_\gamma^2 \tag{57}$$

Combining and expanding $\mathbf{s}_\mathrm{d} = \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$, we have

$$0 = \sum_{\gamma=1}^{N_\mathrm{B}} \left[\langle c \rangle \, (h_\mathrm{d}^2 + \sigma^2 u^2)(1 - h_\mathrm{d}) - \sigma^2 h_\gamma (h_\mathrm{d}^2 + \sigma^2 u^2 - 2 h_\mathrm{d}) + m_3 h_\gamma^2 \right] \hat{\mathbf{s}}_\gamma$$

Since the $\hat{\mathbf{s}}_\gamma$ odors are linearly independent, we have, for each IBCM neuron, a set of N_B equations specifying the neuron's alignment with each odor, h_γ. These are the fixed point equations to solve for the h_γs:

$$0 = \langle c \rangle \, (h_\mathrm{d}^2 + \sigma^2 u^2)(1 - h_\mathrm{d}) - \sigma^2 h_\gamma (h_\mathrm{d}^2 + \sigma^2 u^2 - 2 h_\mathrm{d}) + m_3 h_\gamma^2 \qquad \forall \, \gamma \in \{1, 2, \ldots, N_\mathrm{B}\} \, , \forall \, \text{neuron} \tag{58}$$

These equations are cubic polynomials in the h_γs, since $h_\mathrm{d} = \langle c \rangle \sum_\gamma h_\gamma$ and $u^2 = \sum_\gamma h_\gamma^2$. There is always a fixed point at $h_\gamma = 0 \, \forall \, \gamma$, but it is unstable; trajectories initialized near the origin move away from it during habituation.

C. Solution for a zero third moment background processes. We first examine solutions when $m_3 = 0$, *e.g.*, Gaussian backgrounds, since the solutions simplify greatly in that case. The equations then have two terms,

$$0 = \langle c \rangle \, (h_\mathrm{d}^2 + \sigma^2 u^2)(1 - h_\mathrm{d}) - \sigma^2 h_\gamma (h_\mathrm{d}^2 + \sigma^2 u^2 - 2 h_\mathrm{d})$$

which can be made individually zero by setting $h_\mathrm{d} = 1$ and $\sigma^2 u^2 = 1$. Thus, these fixed points are defined by two constraints,

$$\sum_\gamma \overline{h}_\gamma = \frac{1}{\langle c \rangle} \tag{59}$$

$$\sum_\gamma \overline{h}_\gamma^2 = \frac{1}{\sigma^2} \tag{60}$$

which correspond, geometrically, to the intersection of a hyperplane where coordinates sum to one, and a hypersphere of radius $1/\sigma^2$, respectively. All $\mathbf{m}$ weights on the resulting $(N_\mathrm{B} - 2)$-dimensional surface in the background subspace are non-isolated fixed points when $N_\mathrm{B} > 2$; Fig. S5A shows a three-dimensional background example, with each neuron converging to a point on the ring defined by these equations. For a two-dimensional background, there can be zero, one, or two isolated fixed points, while for a one-dimensional background, these equations do not apply.

The fixed point equation for a Gaussian background admits another set of solutions, where the two terms are not individually zero. Isolating h_γ, we find it must have the same value for every odor, $h_\gamma = h_0$. Then, $h_\mathrm{d} = N_\mathrm{B} h_0$, $u^2 = N_\mathrm{B} h_0^2$, and we can solve to find

$$h_\gamma = h_0 = \frac{\langle c \rangle + 2\sigma^2 / (N_\mathrm{B} + \sigma^2)}{\sigma^2 + N_\mathrm{B} \langle c \rangle} \quad \forall \, \gamma \in \{1, 2, \ldots, N_\mathrm{B}\} \, .$$

For a one-odor background, this solution would be the stable fixed point; for higher dimensions, we find in practice that it is unstable (see also D.1).

D. Solution for a general background with non-zero third moment. We return to solving the full fixed point equations, eq. 58, when $m_3 \neq 0$. Now, because of the $m_3 h_\gamma$ term, setting $h_\mathrm{d} = 1$ and $u^2 = 1/\sigma^2$ does not satisfy the equations, so we must solve for individual h_γs. This leads to a finite number of isolated fixed points: as shown in Fig. S5B, the third moment of the background breaks the degeneracy seen in the Gaussian case.

First, we notice that the dot products $\overline{h}_\gamma$ of an IBCM neuron can only take at most two different values at steady-state. To show this, we take the difference between the equation for h_γ and some other h_α, which gives

$$0 = m_3(h_\gamma + h_\alpha)(h_\gamma - h_\alpha) - (h_\mathrm{d}^2\sigma^2 + \sigma^4 u^2 - 2h_\mathrm{d}\sigma^2)(h_\gamma - h_\alpha) \ .$$

Either $h_\gamma = h_\alpha$, or if they have different values, then they are related by the constraint

$$h_\gamma + h_\alpha = \frac{h_\mathrm{d}^2\sigma^2 + \sigma^4 u^2 - 2h_\mathrm{d}\sigma^2}{m_3} \ . \tag{61}$$

Note that constraint 61 is removed for Gaussian backgrounds $m_3 = 0$, explaining the non-isolated fixed points in that special case. For any pair γ, α, the r.h.s. is the same; if we imagine a third dot product h_β, then $h_\gamma + h_\alpha = h_\gamma + h_\beta \Rightarrow h_\alpha = h_\beta$, implying that there cannot be a third distinct value. Therefore, either all h_γs are equal, or they each take one of two possible values.

D.1. All h_γs are equal. First, consider the case where all $h_\gamma = y$, a unique dot product value. Then $u^2 = N_\mathrm{B} y^2$ and $h_\mathrm{d} = N_\mathrm{B} \langle c \rangle y$. Inserting in Eq. (58), we can factor out y^2, giving

$$0 = y^2 \left[N_\mathrm{B}^2 \langle c \rangle^3 + 3\sigma^2 N_\mathrm{B} \langle c \rangle + m_3 - y(N_\mathrm{B}^3 \langle c \rangle^4 + 2\sigma^2 N_\mathrm{B}^2 \langle c \rangle^2 + \sigma^4 N_\mathrm{B}) \right] \ .$$

So, either $y = 0$, which is an unstable fixed point, or

$$y = \frac{N_\mathrm{B}^2 \langle c \rangle^3 + 3\sigma^2 N_\mathrm{B} \langle c \rangle + m_3}{N_\mathrm{B}^3 \langle c \rangle^4 + 2\sigma^2 N_\mathrm{B}^2 \langle c \rangle^2 + \sigma^4 N_\mathrm{B}} \ . \tag{62}$$

We conjecture that this fixed point is always unstable, based on the linear stability analysis of section 4F below. Fig. S3 shows that in the background process examples we considered, it is a saddle point, approached before the IBCM neuron becomes selective for a background odor.

D.2. Two different h_γ values (general case. Second, consider the case where $\mathbf{m}$ has a dot product equal to y_1 with k_1 odors, and equal to y_2 with the remaining $k_2 = N_\mathrm{B} - k_1$ odors. We let $y_1 > y_2$ by convention. The values y_1 and y_2 will depend on the repartition k_1, k_2, but there will be a unique pair y_1, y_2 for each choice of k_1, k_2. Moreover, in this case,

$$h_\mathrm{d} = \langle c \rangle (k_1 y_1 + k_2 y_2) \tag{63}$$

$$u^2 = k_1 y_1^2 + k_2 y_2^2 \ . \tag{64}$$

Then, the set of N_B equations 58 really reduces to two equations, one for all the $h_\gamma = y_1$ and the other for y_2, which are symmetric under $1 \leftrightarrow 2$.

We start from Eq. (61) – the difference between the two values of h_γ. We rewrite the equation in terms of y_1 and y_2 as

$$0 = h_\mathrm{d}^2 - 2h_\mathrm{d} + \sigma^2 u^2 - \frac{m_3}{\sigma^2}(y_1 + y_2) \tag{65}$$

by letting one of the h_γs be equal to y_1 and the other, to y_2 . We use this equation to replace, where appropriate, the following term in the fixed point equation 58,

$$h_\mathrm{d}^2 + \sigma^2 u^2 = 2h_\mathrm{d} + \frac{m_3}{\sigma^2}(y_1 + y_2)$$

for, say, $h_\gamma = y_1$ (using y_2 would not make a difference), resulting in

$$0 = \sigma^2 u^2 - h_\mathrm{d}^2 - \frac{m_3}{\sigma^2} h_\mathrm{d}(y_1 + y_2) - \frac{m_3}{\langle c \rangle} y_1 y_2 \tag{66}$$

Together, equations 65 and 66 form our system of equations to solve for y_1 and y_2. Obtaining the latter was the crucial simplification to make, because it eliminates terms linear or cubic in y_i, allowing us to easily isolate y_2 in terms of y_1 (or vice-versa)[1]. Indeed, writing h_d and u^2 in terms of the y_i (equations 63-64), we find it takes the simple, symmetric form

$$0 = a_1 y_1^2 - b y_1 y_2 + a_2 y_2^2$$

[1] We have managed to reduce the degree from cubic to quadratic because the substitutions eliminated one root $y_1 = y_2 = 0$, which was not interesting.

where

$$\begin{aligned} a_i &= \sigma^2 k_i - \langle c \rangle^2 k_i^2 - \frac{m_3 \langle c \rangle}{\sigma^2} k_i \quad (i \in \{1, 2\}) \\ b &= 2 \langle c \rangle^2 k_1 k_2 + \frac{m_3 \langle c \rangle}{\sigma^2}(k_1 + k_2) + \frac{m_3}{\langle c \rangle} \ . \end{aligned} \tag{67}$$

This equation makes clear the symmetry of solutions under exchange of labels 1 ↔ 2, confirming that we can keep only the roots where $y_1 > y_2$, knowing that other roots would be found by exchanging k_1 and k_2; in other words, the roots y_1', y_2' for $k_1' = k_2$, $k_2' = k_1$, are the solutions for k_1, k_2 with $y_2 > y_1$ (this can be checked explicitly with the solution below). For now, we write y_2 in terms of y_1,

$$y_2 = \left(\frac{b \pm \sqrt{b^2 - 4a_1 a_2}}{2a_2} \right) y_1 = \alpha_{\pm} y_1 \ . \tag{68}$$

Formally, the numerator should be $by_1 \pm \sqrt{b^2 - 4a_1a_2}|y_1|$, but we can absorb the absolute value into ±, compute the solution for both α values, and keep the one with $y_1 > y_2$ at the end. We now insert $y_2 = \alpha y_1$ into Eq. (65); another root $y_1 = y_2 = 0$ can be factored out, and we find the non-trivial solution

$$y_1 = \frac{2 \langle c \rangle (k_1 + \alpha k_2) + \frac{m_3}{\sigma^2}(1 + \alpha)}{\langle c \rangle^2 (k_1 + \alpha k_2)^2 + \sigma^2 (k_1 + \alpha^2 k_2)} \ . \tag{69}$$

Equations 67, 68, and 69 form our analytical solution for the (approximate) fixed points of IBCM neurons in terms of the dot products h_γ taking values y_1 and y_2.

Condition of existence of non-trivial fixed points The fixed points with y_1 and y_2 given by equations 68-69 will only exist in $\mathbb{R}$ if the discriminant $b^2 - 4a_1a_2$ in α is non-negative. Writing this discriminant explicitly,

$$b^2 - 4a_1 a_2 = 4\sigma^2 \langle c \rangle^2 k_1 k_2 \left(N_\mathrm{B} - \frac{\sigma^2}{\langle c \rangle^2} \right) + 12 m_3 \langle c \rangle k_1 k_2 + m_3^2 \left(\frac{\langle c \rangle^2}{\sigma^4} (k_1 - k_2)^2 + \frac{1}{\langle c \rangle^2} + \frac{2N_\mathrm{B}}{\sigma^2} \right) \ . \tag{70}$$

In general, this expression will be > 0, unless there is very high variance and low average concentration, $\sigma^2 > N_\mathrm{B} \langle c \rangle^2$, and vanishing third moment $m_3 \to 0$. This would be an unnaturalistic setting corresponding to Gaussian, zero-average background fluctuations.

E. Summary of the general fixed point solutions.

The fixed point solution for i.i.d. odor concentrations with mean $\langle c \rangle$, variance σ^2, third moment m_3, is summarized here. The fixed points of $\overline{\mathbf{m}}$ are characterized by their dot products with the N_B background odors, $\overline{h}_\gamma = \overline{\mathbf{m}} \cdot \hat{\mathbf{s}}_\gamma$. There are 2^{N_B} fixed points in total: one with all $h_\gamma = 0$, one with all h_γs equal to (subsection D.1)

$$\overline{h}_\gamma = \frac{N_\mathrm{B}^2 \langle c \rangle^3 + 3\sigma^2 N_\mathrm{B} \langle c \rangle + m_3}{N_\mathrm{B}^3 \langle c \rangle^4 + 2\sigma^2 N_\mathrm{B}^2 \langle c \rangle^2 + \sigma^4 N_\mathrm{B}} \ , \tag{62}$$

and $2^{N_\mathrm{B}} - 2$ where k_1 dot products are equal to y_1, and $k_2 = N_\mathrm{B} - k_1$ are equal to y_2, with $\binom{N_\mathrm{B}}{k_1}$ choices for each possible $k_1 \in \{1, \ldots, N_\mathrm{B} - 1\}$. The values y_1 and y_2, where by convention $y_1 > y_2$, are calculated as follows:

$$\begin{aligned} a_i &= \sigma^2 k_i - \langle c \rangle^2 k_i^2 - \frac{m_3 \langle c \rangle}{\sigma^2} k_i \quad (i \in \{1, 2\}) \\ b &= 2 \langle c \rangle^2 k_1 k_2 + \frac{m_3 \langle c \rangle}{\sigma^2}(k_1 + k_2) + \frac{m_3}{\langle c \rangle} && (67) \\ \alpha &= \frac{b \pm \sqrt{b^2 - 4a_1 a_2}}{2a_2} && (68) \\ &\text{Compute solutions for each sign in } \alpha \\ y_1 &= \frac{2 \langle c \rangle (k_1 + \alpha k_2) + \frac{m_3}{\sigma^2}(1 + \alpha)}{\langle c \rangle^2 (k_1 + \alpha k_2)^2 + \sigma^2 (k_1 + \alpha^2 k_2)} && (69) \\ y_2 &= \alpha y_1 \\ &\text{Keep the pair where } y_1 > y_2 \end{aligned}$$

Our linear stability analysis, below, suggests that the *stable* fixed points of an IBCM neuron have one dot product equal to y_1 and all others equal to y_2. This means the neuron becomes selective: it is specifically responding to one background odor.

There are N_{B} such fixed points. We can therefore call $\mathbf{m} \cdot \hat{\mathbf{s}}_\gamma = y_1 \equiv h_{\mathrm{sp}}$ (specific) for that odor, and $y_2 \equiv h_{\mathrm{ns}}$ (non-specific) for all other odors. We observe close quantitative agreement between these fixed point predictions and numerical simulations in the weakly non-Gaussian background, Fig. 3C. For log-normal (Fig. S6D-E) and turbulent backgrounds (Fig. 4A-B), we also observe that IBCM neurons align with individual background odors, but due to stronger correlations between $\mathbf{m}$ and Θ, the exact dot product values do not exactly match equations 67-69.

F. Linear stability analysis of IBCM fixed points. To support our empirical results on IBCM neuron selectivity, we linearize the dynamical equations 51-52 around a fixed point and compute the Jacobian matrix. Then, we evaluate its eigenvalues, at least numerically, for every fixed point, and check which fixed points are stable in a number of examples. Here, we perform this analysis for a single neuron, and assume that weak coupling with other neurons in the network does not fundamentally affect the stability of single-neuron fixed points. We also assume a constant learning rate $\mu_\Theta = \mu$, as would approximately be the case at steady-state in the Law and Cooper variant used for full simulations. Hence, the dynamical equations are

$$\left\langle \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} \right\rangle = \mu \left\langle h^2 \mathbf{s}_{\mathrm{b}} \right\rangle - \mu \left\langle \Theta \right\rangle \left\langle h \mathbf{s}_{\mathrm{b}} \right\rangle$$

$$\left\langle \frac{\mathrm{d}\Theta}{\mathrm{d}t} \right\rangle = \frac{1}{\tau_\Theta} \left(\left\langle h^2 \right\rangle - \left\langle \Theta \right\rangle \right)$$

Computing the Jacobian entries, recalling that $h = \mathbf{m} \cdot \mathbf{s}$ and thus $\frac{\partial h}{\partial m_i} = s_i$, we find

$$\frac{\partial}{\partial m_i} \left\langle \frac{\mathrm{d}m_j}{\mathrm{d}t} \right\rangle = 2\mu \left\langle h s_i s_j \right\rangle - \mu \Theta \left\langle s_i s_j \right\rangle$$

$$\frac{\partial}{\partial \Theta} \left\langle \frac{\mathrm{d}m_j}{\mathrm{d}t} \right\rangle = -\mu \left\langle h s_j \right\rangle$$

$$\frac{\partial}{\partial m_i} \left\langle \frac{\mathrm{d}\Theta}{\mathrm{d}t} \right\rangle = \frac{2}{\tau_\Theta} \left\langle h s_i \right\rangle$$

$$\frac{\partial}{\partial \Theta} \left\langle \frac{\mathrm{d}\Theta}{\mathrm{d}t} \right\rangle = -\frac{1}{\tau_\Theta} \ .$$

These derivatives form the different blocks of the Jacobian matrix, which is, in vector notation,

$$D\mathbf{f}(\mathbf{m}, \Theta) = \left(\begin{array}{c|c} \begin{matrix} 2\mu \left\langle h \mathbf{s}_{\mathrm{b}} \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle \\ -\mu \Theta \left\langle \mathbf{s}_{\mathrm{b}} \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle \end{matrix} & \frac{2}{\tau_\Theta} \left\langle h \mathbf{s}_{\mathrm{b}} \right\rangle \\ \hline -\mu \left\langle h \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle & -\frac{1}{\tau_\Theta} \end{array} \right) \ . \tag{71}$$

This expression is general. Now, computing more explicitly the expectation values for i.i.d. concentrations as in previous subsections,

$$\left\langle h \mathbf{s}_{\mathrm{b}} \right\rangle = h_{\mathrm{d}} \mathbf{s}_{\mathrm{d}} + \sigma^2 \sum_\gamma h_\gamma \hat{\mathbf{s}}_\gamma$$

$$\left\langle \mathbf{s}_{\mathrm{b}} \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle = \mathbf{s}_{\mathrm{d}} \mathbf{s}_{\mathrm{d}}^{\mathsf{T}} + \sigma^2 \sum_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^{\mathsf{T}}$$

$$\left\langle h \mathbf{s}_{\mathrm{b}} \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle = h_{\mathrm{d}} \left\langle \mathbf{s}_{\mathrm{b}} \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle + \sigma^2 \sum_\gamma h_\gamma (\mathbf{s}_{\mathrm{d}} \hat{\mathbf{s}}_\gamma^{\mathsf{T}} + \hat{\mathbf{s}}_\gamma \mathbf{s}_{\mathrm{d}}^{\mathsf{T}}) + 2 m_3 \sum_\gamma h_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^{\mathsf{T}}$$

The last two lines, along with $\Theta = h_{\mathrm{d}}^2 + \sigma^2 u^2$, means that the main block of the matrix is

$$2\mu \left\langle h \mathbf{s}_{\mathrm{b}} \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle - \mu \Theta \left\langle \mathbf{s}_{\mathrm{b}} \mathbf{s}_{\mathrm{b}}^{\mathsf{T}} \right\rangle = \mu \left[(2h_{\mathrm{d}} - h_{\mathrm{d}}^2 - \sigma^2 u^2)(\mathbf{s}_{\mathrm{d}} \mathbf{s}_{\mathrm{d}}^{\mathsf{T}} + \sigma^2 \sum_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^{\mathsf{T}}) + 2 m_3 \sum_\gamma h_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^{\mathsf{T}} + 2\sigma^2 \sum_\gamma h_\gamma (\mathbf{s}_{\mathrm{d}} \hat{\mathbf{s}}_\gamma^{\mathsf{T}} + \hat{\mathbf{s}}_\gamma \mathbf{s}_{\mathrm{d}}^{\mathsf{T}}) \right]$$

We have defined here $\mathbf{s}_{\mathrm{d}} = \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$. We see that these moments depend on the specific odor components $\hat{\mathbf{s}}_\gamma$, making an analytical calculation of eigenvalues hard in general. However, these expressions can be evaluated easily at the analytical fixed points in several examples to check the stability in these cases; we show the eigenvalues for weakly non-Gaussian, log-normal, and turbulent background statistics in Fig. S3. In all these examples, we find that the only stable fixed points are those where the neuron has one dot product $h_\gamma = h_{\mathrm{sp}}$ (specific) and the $N_{\mathrm{B}} - 1$ others are equal to h_{ns} (non-specific). This property is robust against OSN noise, as shown in Fig. S8. In that case, the IBCM and BioPCA models still perform habituation to the true background subspace and new odor recognition, until the OSN noise becomes comparable in magnitude with odor signals (Fig. S8H).

G. Analytical W weights with IBCM neurons. We can also derive an analytical expression for the average, steady-state values of the inhibitory weights W when the projection weights M converge to the IBCM fixed point derived above. We assume there is at least one neuron per odor, $N_\mathrm{I} \geq N_\mathrm{B}$. We call γ_j the background odor to which IBCM neuron j is specific; thus, $\mathbf{m}_j \cdot \hat{\mathbf{s}}_\gamma = h_\mathrm{sp}$ if $\gamma = \gamma_j$, h_ns otherwise. There will be in general some number n_γ of neurons specific to odor γ.

We start by working on the W equation in matrix form,

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \alpha \mathbf{y}(t)\overline{\mathbf{h}}(t)^\intercal - \beta W \ . \tag{72}$$

We define the (constant) matrix Γ, whose columns are the $\hat{\mathbf{s}}_\gamma$, and $\mathbf{c}$ the vector of odor concentrations. Then, $\mathbf{s}_\mathrm{b}(t) = \Gamma\mathbf{c}$. We also define the matrix $H = LM\Gamma$, in which row j gives the alignment of IBCM neuron j with each odor, $\overline{h}_{j\gamma}$. Since each neuron is selective for one odor, each row contains h_sp once and h_ns $N_\mathrm{B} - 1$ times. Averaging the W equation over fast c fluctuations, and neglecting correlations between c, W, and H, we have

$$\left\langle \frac{\mathrm{d}W}{\mathrm{d}t} \right\rangle = \alpha(\Gamma - \langle W\rangle \langle H\rangle) \langle \mathbf{c}\mathbf{c}^\intercal\rangle \langle H\rangle^\intercal - \beta \langle W\rangle \ .$$

Average signs on W and H are implied below. We define $N = \langle \mathbf{c}\mathbf{c}^\intercal\rangle$ and evaluate it for i.i.d. odors,

$$N = \langle \mathbf{c}\mathbf{c}^\intercal\rangle = \langle c\rangle^2 O_{N_\mathrm{B}} + \sigma^2 \mathbb{I}_{N_\mathrm{B}} \ , \tag{73}$$

where $\mathbb{I}$ is the $N_\mathrm{B} \times N_\mathrm{B}$ identity matrix and O_{N_B} is a $N_\mathrm{B} \times N_\mathrm{B}$ matrix filled with ones. We now set $\mathrm{d}W/\mathrm{d}t = 0$ and focus on single columns of the equation, with the notation $\mathbf{w}_j$ for column j of W, and $\mathbf{h}_j$ for column j of $H^\intercal$ ($\mathbf{h}_j^\intercal$ is the row j of H). The set of equations to solve for the $\mathbf{w}_j$ is then

$$\frac{\beta}{\alpha}\mathbf{w}_j + WHN\mathbf{h}_j = \Gamma N \mathbf{h}_j \tag{74}$$

We notice here that all IBCM neurons with the same specificity γ_j will have the same $\mathbf{h}_j$, thus all columns $\mathbf{w}_j$ with the same γ_j will be identical, and we can denote them by $\mathbf{w}_{\gamma_j}$. This allows to rewrite sums over columns as sums over components, for instance $\sum_j \mathbf{w}_j = \sum_\gamma n_\gamma \mathbf{w}_\gamma$. We moreover notice that

$$\mathbf{h}_k^\intercal \mathbf{h}_j = \begin{cases} h_\mathrm{sp}^2 + (N_\mathrm{B} - 1)h_\mathrm{ns}^2 = u^2 & \text{if } \gamma_j = \gamma_k \\ 2h_\mathrm{sp}h_\mathrm{ns} + (N_\mathrm{B} - 2)h_\mathrm{ns}^2 = h_\mathrm{ns}(\overline{h}_\mathrm{d}/\langle c\rangle + h_\mathrm{sp} - h_\mathrm{ns}) & \text{else} \end{cases} \tag{75}$$

and that

$$\mathbf{h}_k^\intercal O_{N_\mathrm{B}} \mathbf{h}_j = \mathbf{h}_k^\intercal \mathbf{v}(h_\mathrm{sp} + (N_\mathrm{B} - 1)h_\mathrm{ns}) = (h_\mathrm{sp} + (N_\mathrm{B} - 1)h_\mathrm{ns})^2 = \frac{\overline{h}_\mathrm{d}^2}{\langle c\rangle^2} \tag{76}$$

where we defined $\mathbf{v}$, a vector filled with ones, and recognized $\overline{h}_\mathrm{d} = \langle c\rangle \sum_\gamma \overline{\mathbf{m}} \cdot \hat{\mathbf{s}}_\gamma = \langle c\rangle^2 (h_\mathrm{sp} + (N_\mathrm{B} - 1)h_\mathrm{ns})$ for all neurons specific to one odor. We also need to compute

$$N\mathbf{h}_j = \sigma^2 \mathbf{h}_j + \langle c\rangle^2 (h_\mathrm{sp} + (N_\mathrm{B} - 1)h_\mathrm{ns})\mathbf{v} = \sigma^2\mathbf{h}_j + \langle c\rangle \overline{h}_\mathrm{d}\mathbf{v}$$

where we used Eq. (73). We use the above results to evaluate the two terms involving N in Eq. (74). With some algebra to combine coefficients efficiently, we find

$$WHN\mathbf{h}_j = \sigma^2(h_\mathrm{sp} - h_\mathrm{ns})^2 n_{\gamma_j}\mathbf{w}_{\gamma_j} + A\sum_\gamma n_\gamma \mathbf{w}_\gamma \tag{77}$$

$$\Gamma N\mathbf{h}_j = \sigma^2(h_\mathrm{sp} - h_\mathrm{ns})\hat{\mathbf{s}}_{\gamma_j} + \left(\frac{\sigma^2}{\langle c\rangle}h_\mathrm{ns} + \overline{h}_\mathrm{d}\right)\mathbf{s}_\mathrm{d} \tag{78}$$

where we have used the average background expression, $\mathbf{s}_\mathrm{d} = \langle c\rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$, and defined a coefficient independent of j,

$$A = \overline{h}_\mathrm{d}^2 + \frac{\sigma^2}{\langle c\rangle}\overline{h}_\mathrm{d}h_\mathrm{ns} + \sigma^2 h_\mathrm{ns}(h_\mathrm{sp} - h_\mathrm{ns}) \ . \tag{79}$$

We now insert Eq. (77) and Eq. (78) into the equation Eq. (74) for $\mathbf{w}_j$ (or equivalently, $\mathbf{w}_{\gamma_j}$), to find

$$B_{\gamma_j}\mathbf{w}_j + A\sum_\gamma n_\gamma \mathbf{w}_\gamma = \sigma^2(h_\mathrm{sp} - h_\mathrm{ns})\hat{\mathbf{s}}_{\gamma_j} + \left(\frac{\sigma^2}{\langle c\rangle}h_\mathrm{ns} + \overline{h}_\mathrm{d}\right)\mathbf{s}_\mathrm{d} \tag{80}$$

where we have defined another coefficient, different for each j in general, unless all n_{γ_j} are equal,

$$B_{\gamma_j} = \frac{\beta}{\alpha} + n_{\gamma_j} \sigma^2 (h_{\mathrm{sp}} - h_{\mathrm{ns}})^2 \ . \tag{81}$$

Now, to solve Eq. (80) for $\mathbf{w}_{\gamma_j}$, we look at the difference between the equations for columns j and k, with $\gamma_k \neq \gamma_j$; this eliminates common terms and allows us to isolate $\mathbf{w}_{\gamma_k}$ in terms of $\mathbf{w}_{\gamma_j}$ and the known vectors $\hat{\mathbf{s}}_\gamma$ only:

$$\begin{aligned} B_{\gamma_j}\mathbf{w}_{\gamma_j} - B_{\gamma_k}\mathbf{w}_{\gamma_k} &= \sigma^2(h_{\mathrm{sp}} - h_{\mathrm{ns}})(\hat{\mathbf{s}}_{\gamma_j} - \hat{\mathbf{s}}_{\gamma_k}) \\ \Rightarrow \mathbf{w}_{\gamma_k} &= \frac{B_{\gamma_j}}{B_{\gamma_k}}\mathbf{w}_{\gamma_j} + \frac{\sigma^2(h_{\mathrm{sp}} - h_{\mathrm{ns}})}{B_{\gamma_k}}(\hat{\mathbf{s}}_{\gamma_k} - \hat{\mathbf{s}}_{\gamma_j}). \end{aligned}$$

Doing this for each $\gamma_k \neq \gamma_j$, and noticing the expression reduces to $\mathbf{w}_{\gamma_j}$ for $\gamma = \gamma_j$, we express $\sum_\gamma n_\gamma \mathbf{w}_\gamma$ in terms of $\mathbf{w}_{\gamma_j}$ only, and insert into eq. Eq. (80) to isolate $\mathbf{w}_{\gamma_j}$. Dividing the result by $B_{\gamma_j}\left(1 + A\sum_\gamma \frac{n_\gamma}{B_\gamma}\right)$, and rearranging with further algebra gives our final expression for columns of the matrix W,

$$\mathbf{w}_{\gamma_j} = \frac{\sigma^2(h_{\mathrm{sp}} - h_{\mathrm{ns}})}{B_{\gamma_j}}\hat{\mathbf{s}}_{\gamma_j} + \frac{\overline{h}_{\mathrm{d}} + \frac{\sigma^2}{\langle c \rangle}h_{\mathrm{ns}}}{B_{\gamma_j}(1+AK)}\mathbf{s}_{\mathrm{d}} - \frac{A\sigma^2(h_{\mathrm{sp}} - h_{\mathrm{ns}})}{B_{\gamma_j}(1+AK)}\sum_\gamma \frac{n_\gamma}{B_\gamma}\hat{\mathbf{s}}_\gamma \ , \tag{82}$$

where we have defined $K = \sum_\rho \frac{n_\rho}{B_\rho}$. These are the analytical expressions that we compare to numerical simulations of the W weights in Figures 3D (weakly non-Gaussian) and S6F (log-normal), reaching close agreement at steady-state. For backgrounds with slower, stronger fluctuations where the correlations between M, W, and $\mathbf{s}(t)$ are not entirely negligible (*e.g.*, turbulent statistics), the agreement would be less accurate. The numbers of neurons selecting each odor, n_{γ_j}, are inferred from the M weights numerical results, then used to evaluate Eq. (82).

Unfortunately, the expression for the instantaneous $\mathbf{y}(t) = \mathbf{s}(t) - WLM\mathbf{s}(t)$ with steady-state M, M weights is quite cumbersome in general, due to the B_γ. However, we find a more compact expression for the average PN response: after simplifying some terms,

$$\langle \mathbf{y} \rangle = \langle c \rangle \sum_\gamma \frac{\hat{\mathbf{s}}_\gamma}{B_\gamma}\frac{1}{1+AK}\left(\frac{\beta}{\alpha} - \sigma^2 h_{\mathrm{ns}}(h_{\mathrm{sp}} - h_{\mathrm{ns}})(n_\gamma N_{\mathrm{B}} - KB_\gamma)\right) \ . \tag{83}$$

The factor $n_\gamma N_{\mathrm{B}} - KB_\gamma$ is zero when all n_γs are equal; hence, the second term represents the bias incurred by having an uneven distribution of IBCM neurons across odor components. The threshold n^* at which $n_\gamma N_{\mathrm{B}} - KB_\gamma = 0$ for some n_γ is $n^* = \frac{\beta/\alpha K}{N_{\mathrm{B}} - K\sigma^2(h_{\mathrm{sp}} - h_{\mathrm{ns}})^2}$; since we can show (using a Lagrange multiplier to enforce $\sum_\rho n_\rho = n_I$) that K is maximized by having a uniform distribution $n_\gamma = n_I/N_{\mathrm{B}}$, the threshold $n^* < n_I/N_{\mathrm{B}}$; hence, all components which have $n_\gamma > n_I/N_{\mathrm{B}}$ surely have a positive factor $(n_\gamma N_{\mathrm{B}} - KB_\gamma)$, and since $h_{\mathrm{ns}} < 0$ in general, they have a negative bias in eq. Eq. (83), i.e. they are suppressed less than other background odor components. Conversely, if some $n_\gamma = 0$ (no neuron specific to that odor), then this odor is still partly subtracted, due to the non-specific response of other neurons, $h_{\mathrm{ns}} < 0$, but at the cost of less efficient inhibition of all other odors, and without overall reduction of fluctuations since the factor $\frac{\beta/\alpha}{B_\gamma} = 1$ if $n_\gamma = 0$.

When all n_γs are equal, the mean PN response to the background, Eq. (83), is minimized for the IBCM network, and it reduces to a simple expression,

$$\langle \mathbf{y} \rangle = \frac{\beta/\alpha}{B + N_{\mathrm{B}}A}\mathbf{s}_{\mathrm{d}} \tag{84}$$

where, recall, A is given by Eq. (79), B by Eq. (81) with all n_γ equal, and $\mathbf{s}_{\mathrm{d}} = \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$ is the average background. We will use this simplified expression of the maximal habituation by IBCM neurons determine the Λ factor needed to match the BioPCA and IBCM performances (section 8B).

5. Analytical fixed point solutions of the BioPCA model

We assume that the average background $\langle \mathbf{s} \rangle = \mathbf{s}_{\mathrm{d}} = \langle c \rangle \sum_{\gamma=1}^{N_{\mathrm{B}}} \hat{\mathbf{s}}_\gamma$ is subtracted from $\mathbf{s}$, from the input to the BioPCA inhibitory neurons, and from $\mathbf{y}$; hence, the effective background to consider here is $\tilde{\mathbf{s}} = \sum_{\gamma=1}^{N_{\mathrm{B}}} \tilde{c}_\gamma \hat{\mathbf{s}}_\gamma$, with $\langle \tilde{c}_\gamma \rangle = 0$ and $\langle \tilde{c}_\gamma^2 \rangle = \sigma^2$ for all components γ. The LN activity is $\overline{\mathbf{h}}(t) = LM\tilde{\mathbf{s}}(t)$ and the PN response is $\mathbf{y}(t) = \tilde{\mathbf{s}}(t) - W\overline{\mathbf{h}}(t)$. The covariance matrix, from which a PCA with N_{B} components can be computed, is

$$C = \langle \tilde{\mathbf{s}}\tilde{\mathbf{s}}^\intercal \rangle = \sigma^2 \sum_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^\intercal = UDU^\intercal \ ,$$

where U is $N_\mathrm{s} \times N_\mathrm{B}$, with its columns containing the N_B principal component vectors with non-zero eigenvalue, and D is $N_\mathrm{B} \times N_\mathrm{B}$ and diagonal with the principal values in it, σ_i^2, $i \in \{1, 2, \dots, N_\mathrm{B}\}$. Since the $\hat{\mathbf{s}}_\gamma$ are not orthogonal in general, these eigenvalues are not all equal to the variance σ^2, but should be on the same scale.

Given a background, U and D are known; we can thus express the steady-state solution of the BioPCA model in terms of these PCA matrices. From Lemma 3 in Minden *et al.*, 2018 [47], we expect the BioPCA model with $N_\mathrm{I} = N_\mathrm{B}$ neurons (one per background component) to have the following stationary solution:

$$L' = D \quad \text{(principal values)} \tag{85}$$

$$LM = \underline{\Lambda} U^\intercal \quad \text{(projection on principal components)} \tag{86}$$

where, recall, $L' = L^{-1}$. The input projections give interneuron activities of

$$\overline{\mathbf{h}}(t) = LM\tilde{\mathbf{s}}(t) = \underline{\Lambda} U^\intercal \tilde{\mathbf{s}}(t) \ .$$

We insert the BioPCA stationary solution into the W equation 72, where we average over fast $\tilde{\mathbf{s}}$ fluctuations, assuming a separation of time scales between these fluctuations and slow M, L, W learning, as in the IBCM case. Writing the W equation in matrix form, this leads to

$$\begin{aligned} \frac{\mathrm{d}W}{\mathrm{d}t} &= \alpha \left\langle \mathbf{y}(t) \overline{\mathbf{h}}^\intercal(t) \right\rangle - \beta W \\ &= \alpha (\mathbb{I} - W \underline{\Lambda} U^\intercal) \left\langle \tilde{\mathbf{s}} \tilde{\mathbf{s}}^\intercal \right\rangle U \underline{\Lambda}^\intercal - \beta W \end{aligned}$$

Setting the W derivative to zero, we can rearrange to isolate W,

$$W (\beta \mathbb{I}/\alpha + \underline{\Lambda} D \underline{\Lambda}^\intercal) = U D \underline{\Lambda}^\intercal \ .$$

Since D åand $\underline{\Lambda}$ are both diagonal, the matrix in parentheses on the left-hand side is diagonal with entries $\beta/\alpha + \Lambda_{ii}^2 \sigma_i^2$. It is full-rank when $N_\mathrm{I} \leq N_\mathrm{B}$, so we can invert the equation explicitly. Since $D\underline{\Lambda}^\intercal$ is also diagonal, we find

$$W = U \mathrm{diag}\left(\frac{\Lambda_{ii} \sigma_i^2}{\beta/\alpha + \Lambda_{ii}^2 \sigma_i^2} \right) \ . \tag{87}$$

Inserting this W back in the expression for the PN response, $\mathbf{y}(t) = \tilde{\mathbf{s}} - W\overline{\mathbf{h}}$, we find

$$\mathbf{y} = \tilde{\mathbf{s}} - W \underline{\Lambda} U^\intercal \tilde{\mathbf{s}} = U \mathrm{diag}\left(\frac{\beta/\alpha}{\beta/\alpha + \Lambda_{ii}^2 \sigma i^2} \right) U^\intercal \tilde{\mathbf{s}} \tag{88}$$

where we have used the fact that $UU^\intercal$ is a projector on the background subspace to write $\tilde{\mathbf{s}} = UU^\intercal \tilde{\mathbf{s}}$, and where $\Lambda_{ii} = \Lambda_\mathrm{PCA}(1 - \lambda_r (i+1)/N_\mathrm{B})$ for $i = 0, 1, \dots, N_\mathrm{I}$. Hence, we see that the BioPCA network projects the inputs on the principal directions ($U^\intercal \tilde{\mathbf{s}}$), reduces the amplitude of each component by a factor $\frac{\beta/\alpha}{\beta/\alpha + \Lambda_{ii}^2 \sigma i^2}$, then reassembles these components (leftmost U). Comparing to equation 84 for the IBCM model, the latter has a better reduction by a factor of approximately $(h_\mathrm{sp} - h_\mathrm{ns})^2$ in the denominator, hence we need to increase the Λ scale in the BioPCA network to match the performance of these models, as explained in section 8B. We check some of these predictions against numerical simulations in a background with log-normal (Fig. S6B-C) and turbulent (Fig. 4C-D) concentration statistics. We also check that these properties are relatively robust against OSN noise (Fig. S8); the first N_B neurons still capture odor directions corresponding to real odors, while additional neurons align with orthogonal OSN noise components (which are part of the full background PCA decomposition).

6. Analytical results for a two-odor simplified background process

To gain further analytical insight into the convergence dynamics of IBCM neurons in particular, we study the simplest non-trivial background, illustrated in Fig. S4A-B. It consists of two odors ($\mathbf{s}_\mathrm{a}$, $\mathbf{s}_\mathrm{b}$) with fluctuating proportion $\overline{\nu}(t)$ following a Ornstein-Uhlenbeck process (section 2B),

$$\mathbf{s}(t) = \left(\frac{1}{2} + \overline{\nu}(t) \right) \mathbf{s}_\mathrm{a} + \left(\frac{1}{2} - \overline{\nu}(t) \right) \mathbf{s}_\mathrm{b} \ . \tag{89}$$

We start by calculating the average fixed points of the IBCM neurons synaptic weights, $\overline{\mathbf{m}}_i$. The fixed point equations 54 are identical for all neurons, so we focus on one neuron and omit index i. As in section 4, we assume that time scales are well separated, replace $\Theta = \langle h^2 \rangle$, and neglect correlations between $\overline{\nu}$, Θ, and $\mathbf{m}$. We work with reduced weights and activities $\overline{h}_i$, averaged over fast fluctuations, so overlines and $\langle \, \rangle$ are implied for the rest of the section. Individual neurons' weights $\mathbf{m}$ can be recovered from equation Eq. (55). Moreover, for this simple background, the learning rate can be chosen constant, $\mu_{\langle \overline{\Theta}_i \rangle} = \mu$. Hence, the fixed point equation to solve here is

$$0 = \mu \left\langle h^2 \mathbf{s}(t) \right\rangle - \left\langle \Theta \right\rangle \left\langle h \mathbf{s}(t) \right\rangle$$

Now, we rewrite

$$\mathbf{s}(t) = \mathbf{s}_{\mathrm{d}} + \overline{\nu}(t)\mathbf{s}_{\mathrm{s}} \ ,$$

where we have defined $\mathbf{s}_{\mathrm{d}} = \frac{1}{2}(\mathbf{s}_{\mathrm{a}} + \mathbf{s}_{\mathrm{b}})$ (deterministic part) and $\mathbf{s}_{\mathrm{s}} = \mathbf{s}_{\mathrm{a}} - \mathbf{s}_{\mathrm{b}}$ (stochastic part). We examine the dot products of synaptic weights with these components, $h_{\mathrm{d}} = \mathbf{m} \cdot \mathbf{s}_{\mathrm{d}}$ and $h_{\mathrm{s}} = \mathbf{m} \cdot \mathbf{s}_{\mathrm{s}}$, such that $h(t) = h_{\mathrm{d}} + \overline{\nu}(t) h_{\mathrm{s}}$. We can solve for the two dot products h_{d} and h_{s} because they specify the fixed points completely for $N_{\mathrm{B}} = 2$ background components. In term of these quantities, the fixed point equation becomes

$$\begin{aligned} 0 &= \left\langle (h_{\mathrm{d}} + \overline{\nu} h_{\mathrm{s}})^2 (\mathbf{s}_{\mathrm{d}} + \overline{\nu}\mathbf{s}_{\mathrm{s}}) \right\rangle - \left\langle (h_{\mathrm{d}} + \overline{\nu} h_{\mathrm{s}})^2 \right\rangle \left\langle (h_{\mathrm{d}} + \overline{\nu} h_{\mathrm{s}})(\mathbf{s}_{\mathrm{d}} + \overline{\nu}\mathbf{s}_{\mathrm{s}}) \right\rangle \\ 0 &= (h_{\mathrm{d}}^2 + \sigma^2 h_{\mathrm{s}}^2 - h_{\mathrm{d}}^3 - h_{\mathrm{d}} h_{\mathrm{s}}^2 \sigma^2)\mathbf{s}_{\mathrm{d}} + (2h_{\mathrm{s}} h_{\mathrm{d}} - h_{\mathrm{d}}^2 h_{\mathrm{s}} - \sigma^2 h_{\mathrm{s}}^3)\sigma^2 \mathbf{s}_{\mathrm{s}} \end{aligned}$$

Since $\mathbf{s}_{\mathrm{d}}$ and $\mathbf{s}_{\mathrm{s}}$ are linearly independent, both coefficients must be zero, leading to a system of two equations for h_{d} and h_{s},

$$\begin{aligned} 0 &= h_{\mathrm{d}}^3 - h_{\mathrm{d}}^2 - \sigma^2 h_{\mathrm{s}}^2 + h_{\mathrm{d}} h_{\mathrm{s}}^2 \sigma^2 \\ 0 &= \sigma^2 h_{\mathrm{s}} (h_{\mathrm{d}}^2 + \sigma^2 h_{\mathrm{s}}^2 - 2h_{\mathrm{d}}) \ . \end{aligned}$$

There is a trivial solution $h_{\mathrm{s}} = h_{\mathrm{d}} = 0$, which is unstable. The other solutions are, by inspection,

$$h_{\mathrm{d}} = \mathbf{m} \cdot \mathbf{s}_{\mathrm{d}} = 1 \text{ and } h_{\mathrm{s}} = \mathbf{m} \cdot \mathbf{s}_{\mathrm{s}} = \pm\frac{1}{\sigma} \ ,$$

or, in terms of the dot products with $\mathbf{s}_{\mathrm{a}}$ and $\mathbf{s}_{\mathrm{b}}$,

$$\mathbf{m}_{\pm} \cdot \mathbf{s}_{\mathrm{a}} = 1 \pm \frac{1}{2\sigma} \text{ and } 1 \mp \frac{1}{2\sigma} \ .$$

Hence, we have two different stable fixed points, which we call $\mathbf{m}_+$ and $\mathbf{m}_-$ to indicate which sign the dot product with $\mathbf{s}_{\mathrm{s}}$ takes. Figure S4C shows the convergence of a two-neuron network to these fixed points. To interpret these expressions, consider the response at the fixed point to some input sample $\mathbf{s}(t)$:

$$h_{\pm}(t) = \mathbf{m}_{\pm} \cdot (\mathbf{s}_{\mathrm{s}} + \overline{\nu}(t)\mathbf{s}_{\mathrm{d}}) = 1 \pm \frac{\overline{\nu}(t)}{\sigma} \tag{90}$$

We notice that $h_{\pm} = 0$ when $\overline{\nu} = \mp\sigma$, that is, the IBCM neuron is non-responsive to an odor component one standard deviation away on one side of the average background, while it responds strongly to odors on the other side of the average. Hence, in this simplified background, the specificity property of IBCM neurons translates into selecting inputs one standard deviation away from the average.

A. Analytical results: PN inhibition. From the steady-state solution for $\mathbf{m}$, we can also compute the steady-state inhibitory weights $\mathbf{w}$. We assume there are $N_{\mathrm{I}} = 2$ neurons, one at each fixed point ±.

Averaging the W equation 72 over background fluctuations, writing out $\mathbf{y} = \mathbf{s} - WLM\mathbf{s}$, and focusing on the column for one neuron j, we have

$$\frac{\mathrm{d} \left\langle \mathbf{w}_j \right\rangle}{\mathrm{d}t} = \alpha \left\langle \overline{h}_j (\mathbf{s} - WLM\mathbf{s}) \right\rangle - \beta \left\langle \mathbf{w}_j \right\rangle \ \ \forall j \ . \tag{91}$$

To solve for $\mathbf{w}_j$, we set the derivative to zero, and we assume there are $N_{\mathrm{I}} = 2$ IBCM neurons, one at each fixed point ±. We still assume a separation of time scales, and assume the $\overline{\mathbf{m}}$ are equal to their average fixed point values, so at any time t, the IBCM neuron activity $\overline{h}(t)$ is given by Eq. (90). We assume the two neurons converge to fixed points + and −, respectively. We thus establish equations to solve for the $\mathbf{w}_j$ weights fixed point values, $\mathbf{w}_+$ and $\mathbf{w}_-$. We have

$$\frac{\mathrm{d}\mathbf{w}_{\pm}}{\mathrm{d}t} = 0 = \alpha \Big\langle \left(1 \pm \frac{\overline{\nu}(t)}{\sigma}\right) \left[\mathbf{s}_{\mathrm{d}} + \overline{\nu}(t)\mathbf{s}_{\mathrm{s}} - \mathbf{w}_+ \left(1 + \frac{\overline{\nu}(t)}{\sigma}\right) - \mathbf{w}_- \left(1 - \frac{\overline{\nu}(t)}{\sigma}\right) \right] \Big\rangle - \beta \mathbf{w}_{\pm} \ .$$

Solving for $\mathbf{w}_+$ and $\mathbf{w}_-$, we find answers summarized as

$$\mathbf{w}_{\pm} = \frac{\alpha}{2\alpha + \beta} (\mathbf{s}_{\mathrm{d}} \pm \sigma \mathbf{s}_{\mathrm{s}}) \tag{92}$$

Hence, each IBCM neuron inhibits the off-average component for which it is selective, $\mathbf{s}(\overline{\nu} = \pm\sigma)$. Combining the two IBCM neurons, the instantaneous PN activity is reduced to

$$\begin{aligned}\mathbf{s}(t) &= \mathbf{s}_{\mathrm{d}} + \overline{\nu}(t)\mathbf{s}_{\mathrm{s}} - \overline{h}_{+}\mathbf{w}_{+} - \overline{h}_{-}\mathbf{w}_{-} \\ &= \mathbf{s}_{\mathrm{d}} - \frac{\alpha}{2\alpha+\beta}\left(1 + \frac{\overline{\nu}}{\sigma}\right)(\mathbf{s}_{\mathrm{d}} + \sigma\mathbf{s}_{\mathrm{s}}) - \frac{\alpha}{2\alpha+\beta}\left(1 - \frac{\overline{\nu}}{\sigma}\right)(\mathbf{s}_{\mathrm{d}} - \sigma\mathbf{s}_{\mathrm{s}}) \\ &= \frac{\beta/\alpha}{2+\beta/\alpha}\mathbf{s}(t) \ . \end{aligned} \tag{93}$$

Figure S4F-G show close agreement at steady-state between numerical simulations and equations 93 and 92. Hence, by learning $N_{\mathrm{B}} = 2$ linearly independent components $\mathbf{w}_{\pm}$ that are one standard deviation away from the average background, the network is able to suppress any $\mathbf{s}(t)$ from that background, in real time, to a fraction $\frac{\beta}{2\alpha+\beta}$ of its original amplitude. Therefore, not only the average, but also the variance of the background is reduced: background fluctuations are actively suppressed by the IBCM-inhibitory neuron pairs. However, new odors would not be suppressed in the same way, because they have a component orthogonal to the vector space of learnt background.

B. Analytical results: convergence time. The convergence time of the $\mathbf{m}$ weights of an IBCM neuron can be estimated analytically in the two-odor simplified background; this analysis reveals the main parameters influencing how long it takes to habituate to a fluctuating background. Numerically, we observe that with the α, β rates chosen, W weights converge at a similar pace.

We again make a quasi-static approximation on the threshold Θ, assuming it averages over the fast background fluctuations but also converges fast enough to track the slow variations of $\mathbf{m}$,

$$\Theta = \langle(\mathbf{m}\cdot\mathbf{s}_{\mathrm{d}} + \overline{\nu}\mathbf{m}\cdot\mathbf{s}_{\mathrm{s}})^2\rangle = h_{\mathrm{d}}^2 + \sigma^2 h_{\mathrm{s}}^2$$

where we made use of $\langle\overline{\nu}\rangle = 0$ and $h(t) = h_{\mathrm{d}} + \overline{\nu}(t)h_{\mathrm{s}}$. Then, we derive dynamical equations for the slow variables h_{d} and h_{d}, by taking the dot product of $\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}$ with $\mathbf{s}_{\mathrm{s}}$ and $\mathbf{s}_{\mathrm{d}}$, averaging over fast time scales of $\overline{\nu}(t)$, and using the quasi-static Θ above. To simplify calculations, we assume that the two odor vectors, $\mathbf{s}_{\mathrm{a}}$ and $\mathbf{s}_{\mathrm{b}}$, have the same norm (like the $\hat{\mathbf{s}}_{\gamma}$ in the general case are unit normed); in this case, the vectors $\mathbf{s}_{\mathrm{d}}$ and $\mathbf{s}_{\mathrm{s}}$ are orthogonal. Making use of these properties, we calculate for instance, for h_{d},

$$\begin{aligned}\frac{\mathrm{d}h_{\mathrm{d}}}{\mathrm{d}t} &= \left\langle \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}\cdot\mathbf{s}_{\mathrm{d}}\right\rangle = \mu\left\langle(h_{\mathrm{d}} + \overline{\nu}h_{\mathrm{s}})(h_{\mathrm{d}} + \overline{\nu}h_{\mathrm{s}} - \Theta)\right\rangle\mathbf{s}_{\mathrm{d}}^2 \\ &= \mu\left(h_{\mathrm{d}}^2 + \sigma^2 h_{\mathrm{s}}^2 - h_{\mathrm{d}}\left(h_{\mathrm{d}}^2 + \sigma^2 h_{\mathrm{s}}^2\right)\right)\mathbf{s}_{\mathrm{d}}^2 \\ &= \mu\mathbf{s}_{\mathrm{d}}^2(1 - h_{\mathrm{d}})\left(h_{\mathrm{d}}^2 + \sigma^2 h_{\mathrm{s}}^2\right) \quad . \end{aligned} \tag{94}$$

By a similar calculation, we find for h_{s}

$$\frac{\mathrm{d}h_{\mathrm{s}}}{\mathrm{d}t} = \mu\sigma^2\mathbf{s}_{\mathrm{s}}^2 h_{\mathrm{s}}\left(2h_{\mathrm{d}} - \left(h_{\mathrm{d}}^2 + \sigma^2 h_{\mathrm{s}}^2\right)\right) \quad . \tag{95}$$

From equations Eq. (94) and Eq. (95), we can conclude there will be two phases to the dynamics if the initial values of $h_{\mathrm{s}}(0) = \epsilon_{\mathrm{s}}$ and $h_{\mathrm{d}}(0) = \epsilon_{\mathrm{d}}$ are small, and σ^2 is small also. The only positive term in $\frac{\mathrm{d}h_{\mathrm{s}}}{\mathrm{d}t}$ contains h_{d}; hence, as long as h_{d} is small, h_{s} will remain close to zero. The first phase therefore consists in the growth of h_{d} to its steady-state value of 1, while h_{s} remains approximately equal to its initial value, ϵ_{s}. We call t_{d} its duration. After h_{d} has converged, the second phase consists in the growth of h_{s}. We call t_{s} the duration of that phase. Hence, h_{s} reaches steady-state after a total time of $t_{\mathrm{d}} + t_{\mathrm{s}}$.

We compute t_{d} (first phase duration) by integrating equation Eq. (94) from 0 to some fraction ξ (close to unity; we use $\xi = 0.9$ in practice) of the steady-state $h_{\mathrm{d}} = 1$, with the assumption that h_{s}^2 is approximately constant and sub-dominant in that phase, i.e. $h_{\mathrm{s}}^2 \approx \epsilon_{\mathrm{s}}^2 \approx 0$. We find

$$\begin{aligned}\int_{\epsilon_{\mathrm{d}}}^{\xi}\frac{\mathrm{d}h_{\mathrm{d}}}{h_{\mathrm{d}}^2(1 - h_{\mathrm{d}})} &= \int_0^{t_{\mathrm{d}}}\mu s_{\mathrm{d}}^2\mathrm{d}t \\ \Rightarrow t_{\mathrm{d}} &= \frac{1}{\mu s_{\mathrm{d}}^2}\left[\frac{1}{\epsilon_{\mathrm{d}}} - \frac{1}{\xi} + \ln\left(\frac{\xi(1-\epsilon_{\mathrm{d}})}{\epsilon_{\mathrm{d}}(1-\xi)}\right)\right] \end{aligned} \tag{96}$$

where $s_{\mathrm{d}}^2 = \|\mathbf{s}_{\mathrm{d}}\|^2$.

Then, once $h_{\mathrm{d}} \approx 1$, the second phase starts. We neglect sub-dominant terms and we integrate from t_{d} (time at which $h_{\mathrm{d}} \approx 1$ but $h_{\mathrm{s}} \approx \epsilon_{\mathrm{s}}$ still) to $t_{\mathrm{d}} + t_{\mathrm{s}}$. We integrate h_{s} from ϵ_{s} to $\pm\xi/\sigma$: depending on the sign of the initial value ϵ_{s}, the system goes to either fixed point $\pm 1/\sigma$ (same sign as the initial value). Hence,

$$\begin{aligned}\frac{\mathrm{d}h_\mathrm{s}}{\mathrm{d}t} &\approx \mu \mathbf{s}_\mathrm{s}^2 \sigma^2 h_\mathrm{s} \\ \Rightarrow \int_{\epsilon_\mathrm{s}}^{\pm\xi/\sigma} \frac{\mathrm{d}h_\mathrm{s}}{h_\mathrm{s}} &= \mu \mathbf{s}_\mathrm{s}^2 \sigma^2 h_\mathrm{s} \int_{t_\mathrm{d}}^{t_\mathrm{s}+t_\mathrm{d}} \mathrm{d}t \\ \Rightarrow t_\mathrm{s} = \frac{1}{\mu \mathbf{s}_\mathrm{s}^2 \sigma^2} &\ln \left| \frac{\xi}{\sigma \epsilon_\mathrm{s}} \right| \ . \end{aligned} \tag{97}$$

Fig. S4D and E show that the approximations Eq. (96) and Eq. (97) hold well in a range of initial values ϵ_s between 0.005 and 0.05, and ϵ_d between 0.01 and 0.1. Importantly, these analytical expressions show that convergence time is faster if initial conditions are larger (note inverse ϵ terms) and the noise σ^2 is larger. It means that initial conditions must be chosen to avoid very small dot products with inputs initially. Then, larger background fluctuations trigger faster convergence too, at least in this simple background model. This is a surprising property of the IBCM model: fluctuations drive the dynamics.

C. BioPCA neuron on the two-odor simplified background. We also characterized the behavior of the BioPCA model in this simplified background. Since the background effectively has one principal direction, $\mathbf{s}_\mathrm{s}$ (Fig. S7A), a single BioPCA neuron is needed to capture it. Then, the matrix M is only a row vector $\mathbf{m}^\intercal$ and the matrix L is only a scalar ℓ, so $L' = \ell' = 1/\ell$. Likewise, W is a column vector $\mathbf{w}$ and the matrix $\underline{\Lambda}$ is just the scalar parameter Λ. Since we assume BioPCA receives the input with the average subtracted, the background we consider is $\tilde{\mathbf{s}} = \overline{\nu}(t)\mathbf{s}_\mathrm{s}$. We illustrate how to solve the BioPCA steady-state equations in this simple case. With $\overline{h} = \overline{\nu}(t)\mathbf{m}^\intercal\mathbf{s}_\mathrm{s}/\ell'$, the dynamical equations simplify to

$$\begin{aligned}\frac{1}{\mu_M}\frac{\mathrm{d}\mathbf{m}^\intercal}{\mathrm{d}t} &= (\overline{\nu}(t)\mathbf{m}^\intercal\mathbf{s}_\mathrm{s}/\ell')\overline{\nu}(t)\mathbf{s}_\mathrm{s}^\intercal - \mathbf{m}^\intercal \\ \frac{1}{\mu_L}\frac{\mathrm{d}\ell'}{\mathrm{d}t} &= \overline{\nu}(t)\mathbf{m}^\intercal\mathbf{s}_\mathrm{s}/\ell')^2 - \Lambda^2\ell'\end{aligned}$$

Averaging over $\overline{\nu}(t)$ and setting the derivatives equal to zero, we have two fixed point equations,

$$\begin{aligned}0 &= \sigma^2 h_\mathrm{s}/\ell'\mathbf{s}_\mathrm{s} - \mathbf{m} \\ 0 &= \sigma^2 h_\mathrm{s}^2/\ell'^2 - \Lambda^2\ell'\end{aligned}$$

where we have defined $h_\mathrm{s} = \mathbf{m}^\intercal\mathbf{s}_\mathrm{s}$. The first equation shows that $\mathbf{m}$ is parallel to the first principal component: $\mathbf{m} = \|\mathbf{m}\|\mathbf{s}_\mathrm{s}/\|\mathbf{s}_\mathrm{s}\|$. To find its magnitude, we take the dot product of that equation with $\mathbf{s}_\mathrm{s}$, which allows to factor out h_s and solve for ℓ'. We find that ℓ' does converge to the first principal eigenvalue, which is $\sigma^2\|\mathbf{s}_\mathrm{s}\|^2$ in this simplified background,

$$L_{11} = 1/L'_{11} = \frac{1}{\sigma^2\|\mathbf{s}_\mathrm{s}\|^2} \ . \tag{98}$$

From the second equation, we then find $h_\mathrm{s} = \Lambda\sigma^2\|\mathbf{s}_\mathrm{s}\|^3$, which means that $\mathbf{m}$ has a norm $\|\mathbf{m}\| = \Lambda\sigma^2\|\mathbf{s}_\mathrm{s}\|^2$. Hence, we have $\mathbf{m}$ parallel to the first principal component,

$$\mathbf{m} = \sigma^2\|\mathbf{s}_\mathrm{s}\|\mathbf{s}_\mathrm{s} \ . \tag{99}$$

Also, the instantaneous response of this neuron to $\tilde{\mathbf{s}}$ is $\overline{h}(t) = \ell\mathbf{m}\cdot\tilde{\mathbf{s}}(t) = \Lambda\|\mathbf{s}_\mathrm{s}\|\overline{\nu}(t)$. Inserting these expressions in the $\mathbf{w}$ equation for this single neuron,

$$\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \alpha(\Lambda\|\mathbf{s}_\mathrm{s}\|\overline{\nu}(t))(\overline{\nu}(t)\mathbf{s}_\mathrm{s} - \mathbf{w}(\Lambda\|\mathbf{s}_\mathrm{s}\|\overline{\nu}(t))) - \beta\mathbf{w} \ .$$

Averaging over $\overline{\nu}(t)$ and solving for $\mathbf{w}$, we find that it is also a vector parallel to $\mathbf{s}_\mathrm{s}$,

$$\mathbf{w} = \frac{\sigma^2\Lambda\|\mathbf{s}_\mathrm{s}\|}{\sigma^2\|\mathbf{s}_\mathrm{s}\|^2\Lambda^2 + \beta/\alpha}\mathbf{s}_\mathrm{s} \ . \tag{100}$$

Lastly, computing the PN instantaneous activity once the BioPCA neuron has reached its learning fixed point, we find

$$\mathbf{y}(t) = \tilde{\mathbf{s}}(t) - WLM\tilde{\mathbf{s}}(t) = \frac{\beta/\alpha}{\beta/\alpha + \sigma^2\Lambda^2\|\mathbf{s}_\mathrm{s}\|^2}\overline{\nu}(t)\mathbf{s}_\mathrm{s} \ . \tag{101}$$

Fig. S7 shows that these analytical predictions for $\mathbf{m}$, L, $\mathbf{w}$, and $\mathbf{y}(t)$ match numerical simulations very well. Hence, by learning the direction of fluctuations along the first principal component, the BioPCA neuron reduces the mean and standard deviation of fluctuations by a factor $\frac{\beta/\alpha}{\beta/\alpha+\sigma^2\Lambda^2\|\mathbf{s}_\mathrm{s}\|^2}$, to be compared with the reduction achieved by the IBCM network, in equation 93.

7. Testing different L-norms for the W Hebbian learning rule

The Hebbian learning rule for W used in the main text (derived in *Methods*) causes our proposed models to subtract the entire component of the input lying in the background subspace. It gives them a sub-optimal performance, limited to the similarity between the new odor and its orthogonal component (Figures 5, S10). To start exploring alternative W rules that could better exploit the odor-specific projections learnt by IBCM neurons, we considered the effect of using different L^p-norms in the cost function from which the W dynamics are derived. We write a cost function for W weights based on the L^p-norm of PN activity and the entry-wise L^q-norm of W (generalization of the Frobenius norm, which corresponds to $q = 2$), defined as

$$\|\mathbf{y}\|_p = \left(\sum_{i=1}^{N_\mathrm{S}} |y_i|^p \right)^{1/p}$$

$$\|W\|_q = \left(\sum_{i=1}^{N_\mathrm{S}} \sum_{j=1}^{N_\mathrm{I}} \|W_{ij}\|^q \right)^{1/q} \quad .$$

In the cost function, we square terms to preserve direct comparisons with the default L^2-norm cost function:

$$\mathcal{L}_W = \frac{1}{2}\|\mathbf{y}\|_p^2 + \frac{\beta}{2\alpha}\|W\|_q^2 \; .$$

Taking gradient descent dynamics on this loss function gives a generalized Hebbian learning rule,

$$\frac{\mathrm{d}W_{ij}}{\mathrm{d}t} = \alpha \|\mathbf{y}\|_p^{2-p} |y_i|^{p-1} \mathrm{sgn}(y_i) \overline{h}_j - \beta \|W\|_q^{2-q} |W_{ij}|^{q-1} \mathrm{sgn}(W_{ij}) \; . \tag{102}$$

Notice that it reduces to the main text rule when $p = q = 2$. We performed numerical experiments of habituation and new odor recognition, analogous to Fig. 2, for various p, q choices in this generalized Hebbian rule. For each (p, q) choice, we optimized the performance with a grid search over a few α and β learning rate values, centered on relevant windows (*e.g.*, a smaller p requires a smaller α to prevent numerical instabilities). Unfortunately, different p, q choices did not fundamentally alter the model performance for habituation (Fig. S9A-C) or new odor recognition (Fig. S9D-F); in fact, the default L^2-norm provided the best results. Hence, more strongly nonlinear versions of manifold learning, such as online manifold tiling [57], or learning rules with positive feedbacks to learn the optimal matrix P of section 1, would be needed to further improve new odor recognition performance.

8. Projection weights scale factor, Λ

As explained in *Methods*, we defined a parameter Λ controlling the scale of M weights and compensating for the regularization on W. Here, we explain how to introduce Λ in the IBCM model, and how to set this parameter to make BioPCA and IBCM perform equivalently.

A. Introducing a scale factor Λ in the IBCM model. We start with the equations for a single neuron. We seek to introduce Λ where appropriate in the equations to maintain the exact same dynamics, only with the numerical values of $\mathbf{m}$ weights (including their initial values) scaled by some factor Λ. By definition, as we scale $\mathbf{m} \sim \Lambda$, then $h = \mathbf{m} \cdot \mathbf{s} \sim \Lambda$ as well. The IBCM equation contains terms of the form $h - \Theta$, with $\Theta \sim h^2$; to keep these terms matched, we need $\Theta \sim \Lambda$, which we achieve by letting $\Theta \to h^2/\Lambda$. Hence, we start by modifying the threshold equation to

$$\frac{\mathrm{d}\Theta}{\mathrm{d}t} = \frac{1}{\tau_\Theta}(h^2/\Lambda - \Theta) \; . \tag{103}$$

We do not need to rescale the learning rate here, because both sides have a homogeneous scaling $\sim \Lambda$. However, in the $\mathbf{m}$ equation, we need to rescale the learning rate to preserve the dynamics, because the right-hand side has terms $\sim h^2$. To keep both sides scaling as $\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} \sim \Lambda$, we modify the $\mathbf{m}$ equation to

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \frac{\mu_\Theta}{\Lambda} h(h - \Theta)\mathbf{s}(t) - \varepsilon \frac{\mu_\Theta}{\Lambda}\mathbf{m}, \; . \tag{104}$$

Moreover, since the scale of $\Theta \sim \Lambda$, we need to rescale it in the Θ-dependent learning rate (from our variant of the Law and Cooper version of IBCM),

$$\mu_\Theta = \frac{\mu_0}{\Theta/\Lambda + k_\Theta} \; . \tag{105}$$

Bourassa *et al.* | Manifold learning for olfactory habituation to strongly fluctuating backgrounds bioRχiv | 33

The generalization to a network of IBCM neurons is straightforward, since Λ is a unique scale parameter for all neurons. Each $\overline{\Theta}_i$ equation has the rescaled term $\overline{h}_i^2/\Lambda$ as in 103. All terms in neuron i's $\mathbf{m}_i$ equation, including those from coupling with other neurons j, have their learning rates $\mu_{\overline{\Theta}_j}$ divided by Λ, as in 104. Also, $\overline{\Theta}_i$ is divided by Λ in the denominator of each learning rate $\mu_{\overline{\Theta}_j}$ as in eq. 105.

These are the IBCM equations we use for general Λ values. In Fig. S11, we characterize the performance of the network for habituation and new odor recognition as as function of the scale Λ, and we observe that $\Lambda_{\mathrm{IBCM}} = 1$ is large enough to maximize the performance.

B. Scaling the BioPCA model for performance equivalent to IBCM. As explained in *Methods*, the scale Λ is already built into the BioPCA model, in the matrix $\underline{\Lambda}$ intervening in the L equation. Λ_{PCA} is set to 1 by default [47], but this leads to a smaller M weights magnitude than by default in the IBCM model. For this reason, Fig. S11 shows that the BioPCA model requires a $\Lambda_{\mathrm{PCA}} > 1$ to achieve the same performance.

For other simulations where we compare the two models, we use our analytical results on the IBCM and BioPCA models to estimate beforehand what Λ_{PCA} value should yield a comparable performance from both models. As derived in Eq. (84), the PN response is reduced, in an IBCM network, by a factor

$$f_{\mathrm{IBCM}} = \frac{\beta/\alpha}{\beta/\alpha + \sigma^2 (h_{\mathrm{sp}} - h_{\mathrm{ns}})^2 \Lambda_{\mathrm{IBCM}}^2 + N_{\mathrm{B}} \left(\overline{h}_{\mathrm{d}}^2 + \sigma^2 \overline{h}_{\mathrm{d}} h_{\mathrm{ns}} / \langle c \rangle + \sigma^2 h_{\mathrm{ns}} (h_{\mathrm{sp}} - h_{\mathrm{ns}}) \right) \Lambda_{\mathrm{IBCM}}^2}$$

where we have explicited how Λ_{IBCM} controls this factor by multiplying the default-scale LN activities, $h_{\mathrm{sp}}, h_{\mathrm{ns}}, \overline{h}_{\mathrm{d}}$ – recall that these are the specific, non-specific, and average alignments of the IBCM fixed points, derived in section 4.

In comparison, Eq. (88) shows that the PN activity in a BioPCA network is reduced along each principal direction by a factor of approximately

$$f_{\mathrm{PCA}} = \frac{\beta/\alpha}{\beta/\alpha + \sigma^2 \Lambda_{\mathrm{PCA}}^2} \ .$$

We set Λ_{PCA} to make these two factors equal, which occurs at

$$f_{\mathrm{IBCM}} = f_{\mathrm{PCA}} \ \Rightarrow \ \Lambda_{\mathrm{PCA}} = \frac{\beta}{\alpha\sigma^2} \frac{1 - f_{\mathrm{IBCM}}}{f_{\mathrm{IBCM}}} \ . \tag{106}$$

So, to set up parameters for a numerical simulation, we compute the statistics of the chosen background ($\langle c \rangle$, σ^2, m_3), the analytical predictions for the IBCM fixed points (h_{sp}, h_{ns}, $\overline{h}_{\mathrm{d}}$), and we set Λ_{PCA} to these parameter values inserted in Eq. (106).

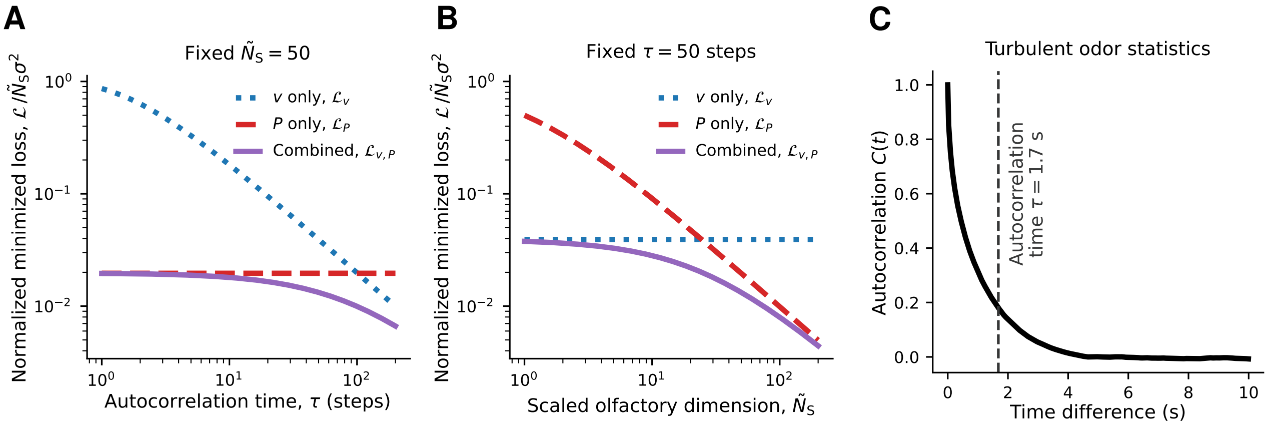


Fig. S1. Comparing manifold learning and predictive filtering. Supplement to Fig. 1E. (**A**) Minimized loss function for new odor recognition in a simple background (Ornstein-Uhlenbeck), for the combined strategies (purple) and either single strategy (blue, red), as a function of the autocorrelation time τ, for fixed olfactory space dimensionality. The combined strategy is always better, but manifold learning explains essentially all the loss reduction at low autocorrelation times ($\mathcal{L}_{P,v} \approx \mathcal{L}_P$). (**B**) Same, as a function of the scaled olfactory space dimension, $\tilde{N}_S = N_S\sigma^2/\sigma^2_{\text{new}}$, for fixed autocorrelation time $\tau = 50$ steps. Most of the loss reduction comes from one strategy or the other on either side of the crossover region; manifold learning dominates in high dimensions. (**C**) Autocorrelation function of the concentration fluctuations in the turbulent background of Fig. 1B-C, showing an autocorrelation time of ~ 1.7 s.

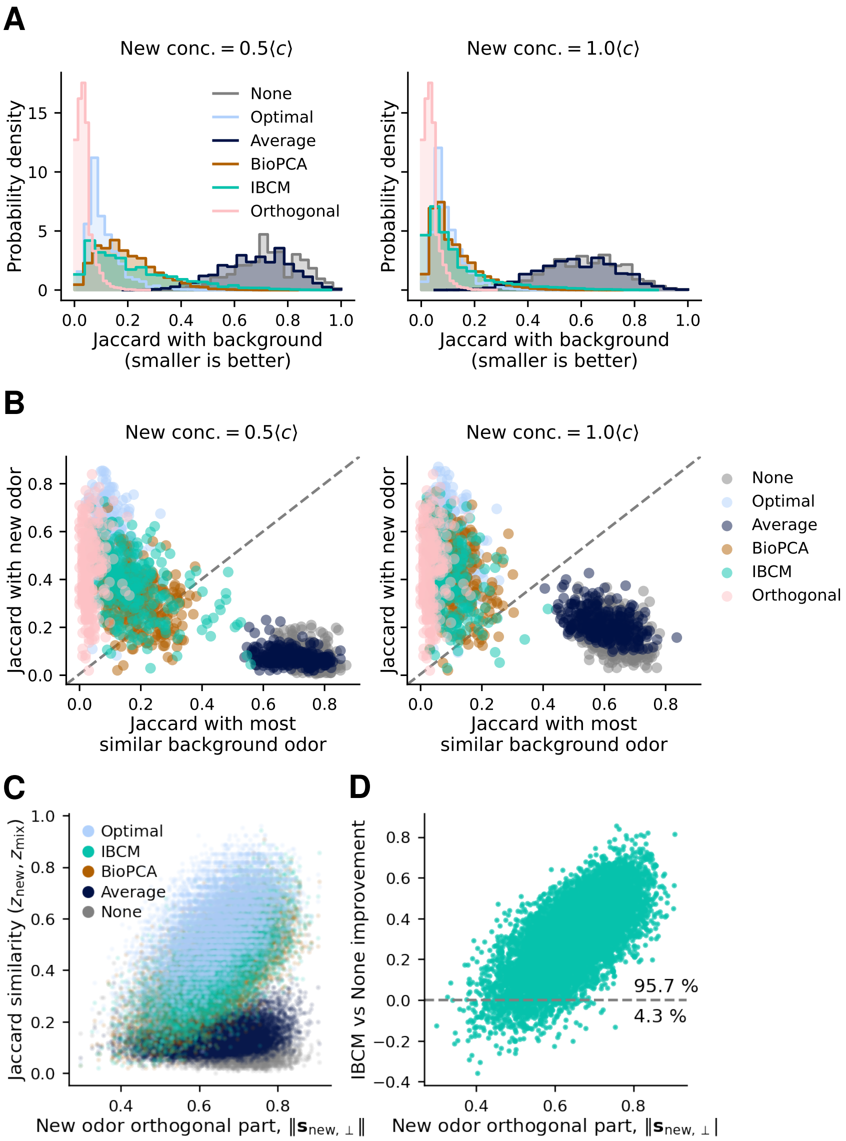


Fig. S2. New odor recognition compared to background recognition after habituation. Supplement to Fig. 2. (**A**) Histograms of the Jaccard similarity between the response to the mixture (background and new odor) and the most similar background odor, across all repeats (new odors, background samples, etc.) described in Fig. 2A, for two tested new odor concentrations. Low similarity is better: it means that background odors are suppressed and do not dominate the response to the new odor, due to habituation. In the absence of habituation, the similarity to background odors remains high. (**B**) Scatter plots of the similarity to the new odor (y axis, larger is better) versus the similarity to the background (x axis). Manifold learning models are generally above the diagonal, meaning their response is more similar to the new odor than to the background, while habituation by average subtraction (as well as no habituation) produce responses still dominated by the background fluctuations (below diagonal). (**C**) Jaccard similarity between $z_{\rm mix}$ and $z_{\rm new}$ plotted as a function of the magnitude of the new odor component orthogonal to the background (or equivalently, of the distance between the background odors and the new odor), in each trial. For all models, the performance improves as the new odor is less similar to the background. (D) For the IBCM model, improvement in Jaccard similarity compared to no habituation ($J_{\rm IBCM} - J_{\rm None}$). Habituation by manifold learning almost always (> 95 % of trials) provides an improvement, except for rare background samples (< 5 %) where background odors are all in blanks or in very dilute whiffs (then the new odor, by chance, is not masked by the background). In (B), (C), and (D), each point is the median across test times and background samples in each individual background simulation (of which there are 100 repeats).

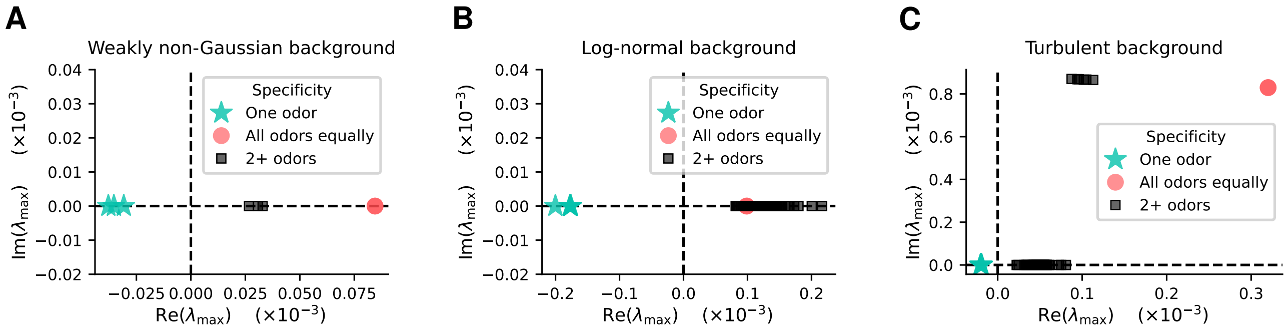


Fig. S3. Selective states are the only stable fixed points in the IBCM model. Eigenvalue with the largest negative real part in the Jacobian matrix obtained in the linear stability analysis of the IBCM model (section 4F), evaluated for each possible fixed point in (**A**) the weakly non-Gaussian background of Fig. 3, (**B**) the log-normal background of Fig. S6, and (**C**) the turbulent background of Figs. 1B-C, 2, 5. The Jacobian matrix derived in eq. 71 is diagonalized numerically for each possible choice of specific and non-specific odors. In the three background examples considered, the only stable fixed points, *i.e.*, fixed points where even the largest eigenvalue has a negative real part, are those where the neuron is specific to only one odor, *i.e.*, selective states. The saddle points where the neuron is equally sensitive to all odors, in red, have some eigenvalues with positive real parts (largest is shown), and some with negative real parts.

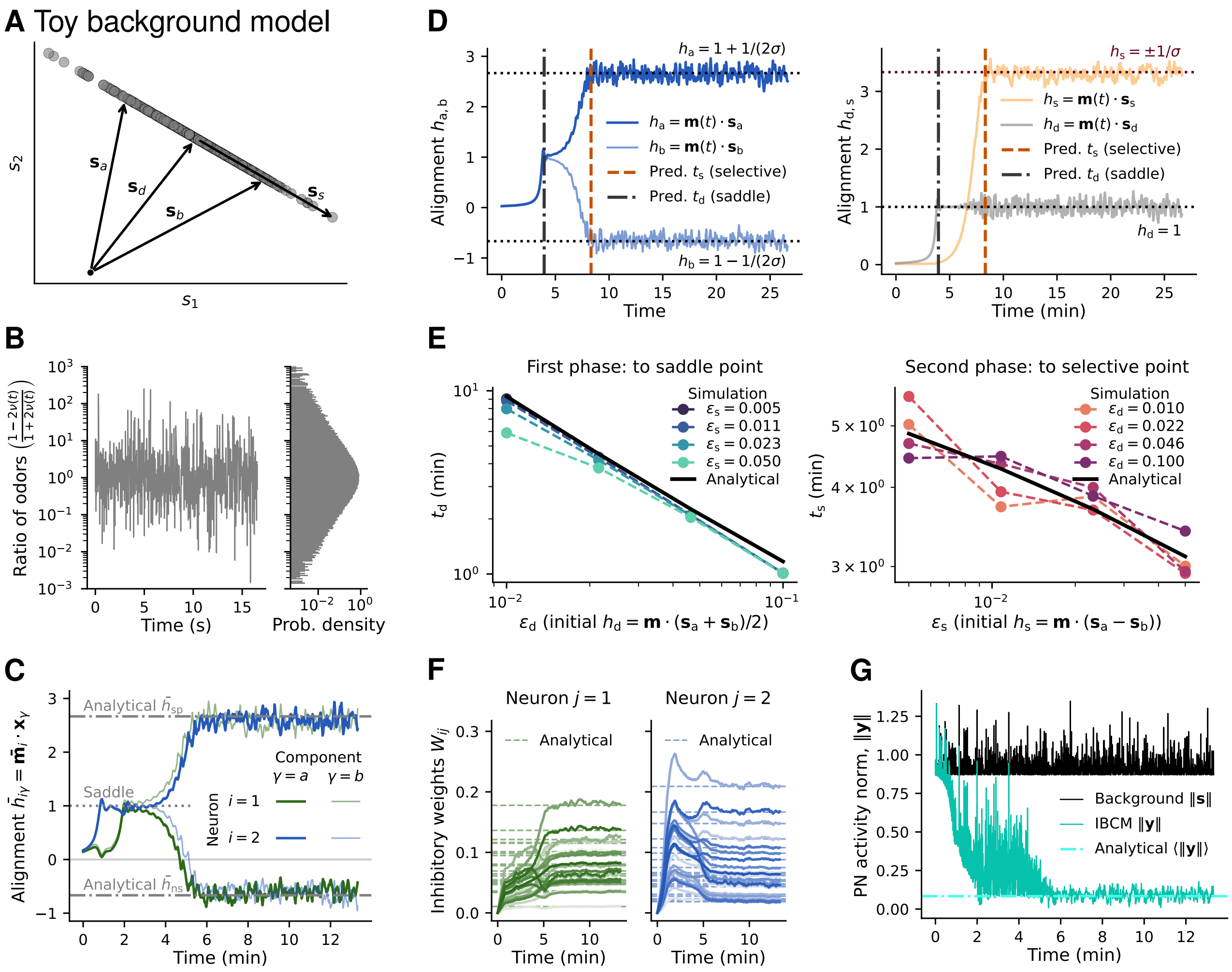
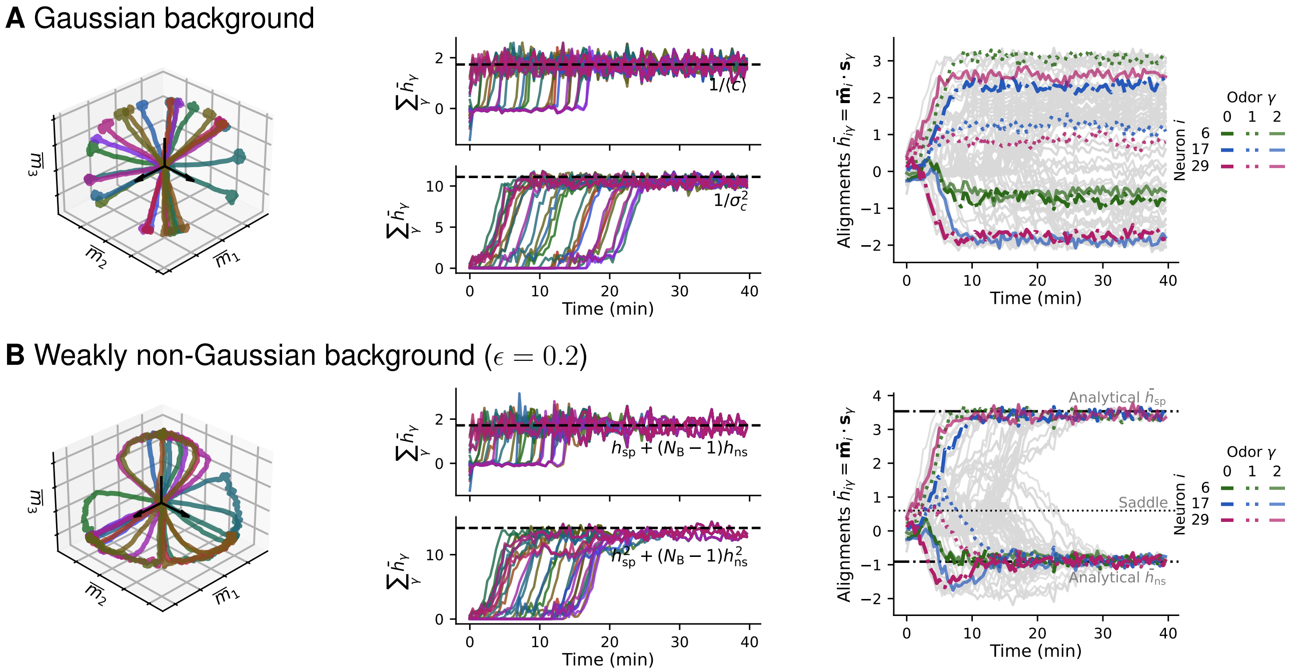


Fig. S4. IBCM model learning dynamics on a simplified background model. (**A**) Schematic of the simplified background model described in section 6: a 1D manifold generated by two odors $\mathbf{s}_\mathrm{a}$, $\mathbf{s}_\mathrm{b}$ with fluctuating proportion $\overline{\nu}(t)$ following a Ornstein-Uhlenbeck process with $\langle\overline{\nu}\rangle = 0$, $\sigma^2 = 0.09$, autocorrelation time $\tau_\mathrm{b} = 20$ ms. The deterministic part is $\mathbf{s}_\mathrm{d} = \frac{1}{2}(\mathbf{s}_\mathrm{a} + \mathbf{s}_\mathrm{b})$ and the stochastic part, $\mathbf{s}_\mathrm{s} = \mathbf{s}_\mathrm{a} - \mathbf{s}_\mathrm{b}$. The two odors are randomly sampled from the default distribution (exponential iid elements, then normalized). (**B**) Time series and stationary distribution of the ratio of odors a and b. (**C**) Alignment of the two IBCM neurons with the background odors. One neuron becomes specific to $\mathbf{s}_\mathrm{a}$, the other, to $\mathbf{s}_\mathrm{b}$. The average dot products at steady-state match the analytical fixed points calculated in 90, $\overline{h}_\pm = 1 \pm 1/(2\sigma)$. (**D**) Analysis of the two-phase dynamics of an IBCM neuron. Left: visualized in terms of the alignments (dot products) with odors a and b, $\mathbf{m}$ first reaches a saddle point, then one of the two selective stable fixed points, respectively at times t_d (eq. 96) and t_s closely matching their analytical values (equations 97 and 96). Right: visualized in terms of the alignments h_d and h_s with the deterministic and stochastic components, $\mathbf{s}_\mathrm{d}$ and $\mathbf{s}_\mathrm{s}$, the saddle at time t_d corresponds to the time at which the deterministic part h_d reaches its fixed point average value 1, and the final steady-state, when the stochastic part reach its fixed point value $\pm 1/\sigma$. (**E**) Scaling of the convergence times t_d (left) and t_s (right) as a function of the initial dot product magnitudes, ϵ_d and ϵ_s. The numerical simulations match the analytical expressions, valid for small ϵs. (**F**) Time series of the inhibitory weights (elements of W), following the Hebbian dynamics of eq. 91, for the two neurons in the network (weights of LN j are in column j of W). The steady-state values match closely the analytical values derived in eq. 92. (**G**) Reduction in PN activity after habituation, compared to the background norm. The steady-state reduction matches the analytical prediction (eq. 93) and is reached once the two neurons select one odor each.



A Gaussian background

B Weakly non-Gaussian background ($\epsilon = 0.2$)

Fig. S5. IBCM learning depends on higher-oder moments of the background distribution. (**A**) Three-odor background with Gaussian concentration fluctuations: each odor has $c_\gamma = g_\gamma$ where g_γ follows a Ornstein-Uhlenbeck process (with $\langle g \rangle = 1/\sqrt{N_{\mathrm{B}}}$, $\sigma_c^2 = 0.09$, correlation time $\tau = 20$ ms). Left: when the third moment is zero, synaptic weights $\overline{\mathbf{m}}_i = \mathbf{m}_i - \eta \sum_{j \neq i} \mathbf{m}_j$ have non-isolated fixed points on the codimension-two ring defined by constraints Eq. (59) (hyperplane) and Eq. (60) (hypersphere). The figure is showing the first three dimensions. Center: All neurons converge to these constraints. Right: alignments with individual odors can take a continuous range of values and do not split into specific and non-specific odors for each neuron. (**B**) Same as (A), but with a small non-zero third moment added to the background concentrations statistics, by taking $c = g + \epsilon g^2$ for $\epsilon = 0.2$, leading to $\langle (c - \langle c \rangle)^3 \rangle \approx 0.01$. Left: the fixed points become isolated, individual neurons first converge to the codimension-two ring of the Gaussian case, then slowly approach one of three selective fixed points near that ring (driven by the small third moment). Center: for each neurons, the sums of $\overline{h}_\gamma$s and $\overline{h}_\gamma^2$s have similar values to the Gaussian case, perturbed by the small third moment. Right: alignments converging to selective fixed points with dot product values h_{sp} (with one odor), h_{ns} (with $N_{\mathrm{B}} - 1$ odors); three neurons highlighted. Odor vectors were chosen to be symmetric around the origin in the first three dimensions, to clarify the geometric picture.

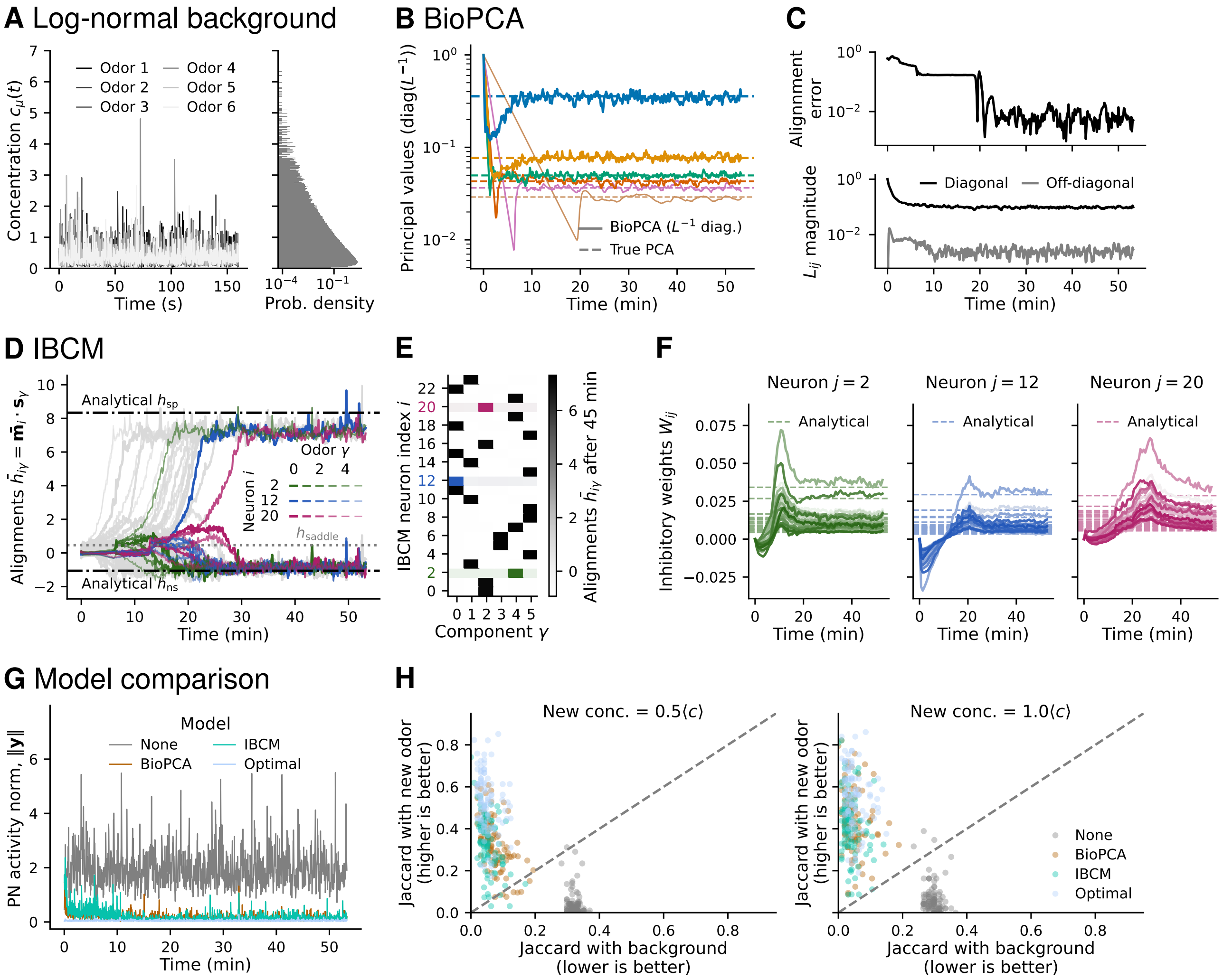


Fig. S6. Habituation to log-normal background statistics. (**A**) (Left) Excerpt from the concentration time series of $N_{\mathrm{B}} = 6$ odors, and (right) histogram of the concentrations showing they follow a log-normal distribution. (**B**) In a six-neuron BioPCA network, learning dynamics of the L matrix diagonal entries, which converge to the background's principal values, as predicted in section 5 (horizontal dashed lines). (**C**) (Top) Alignment error (defined in *Methods*) of the BioPCA M weights with the background subspace, showing that the synaptic weights converge to the principal components vectors, and (bottom) time series of the average diagonal and off-diagonal elements magnitude in L, showing the matrix becomes nearly diagonal, as predicted. (**D**) Time series of $N_{\mathrm{I}} = 24$ IBCM neurons' alignment with the background odors $\hat{\mathbf{s}}_\gamma$, with three neurons highlighted ($N_{\mathrm{B}} = 6$ dot products each). Each neuron aligns with one odor and reaches dot product magnitudes close to the analytical expressions for h_{sp}, h_{ns} (section 4E). Different neurons select different odors. (**E**) Table summarizing the alignment of each IBCM neuron (with three highlighted). Plotted values are the dot products $\overline{h}_\gamma$ averaged over the last 15 minutes of the simulation example. (**F**) Time series of the W weights in the IBCM network, showing their convergence to the analytical predictions (equation 82). (**G**) Example PN time series during habituation by IBCM, BioPCA, and optimal manifold learning networks, compared to the absence of habituation (OSN input). (**H**) Performance of the different models for new odor recognition in log-normal backgrounds, tested in a sample background, across 100 new odors, 10 background samples and 10 test times in that simulation. The plot compares the Jaccard similarity between mixture and new odor responses to the similarity between mixture and background odor responses. Individual dots are medians across background samples and test times. The plot shows that after habituation by the manifold learning models, the response is more similar to the new odor (above the diagonal), while that is not the case in the absence of habituation (below diagonal).

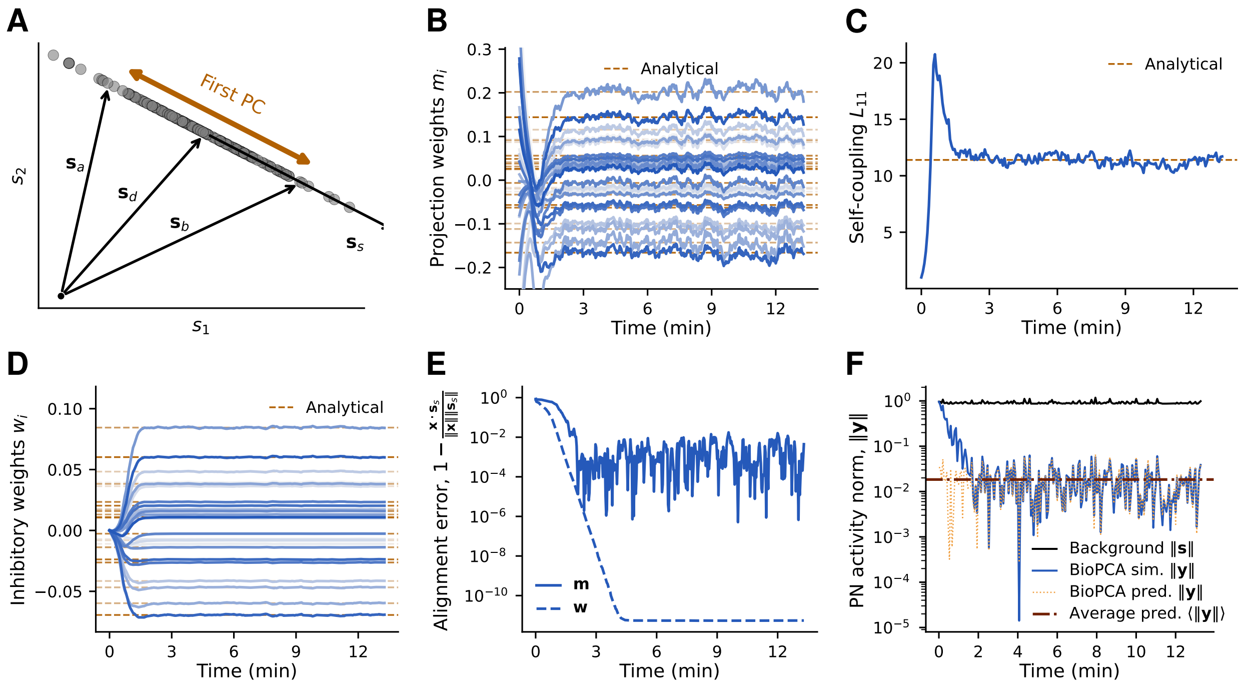


Fig. S7. BioPCA model learning dynamics on a simplified background model. We simulate a network with one BioPCA interneuron habituating to the the same two-odor toy model used in Fig. S4. (**A**) The first principal component (PC) corresponds to the stochastic component $\mathbf{s}_{\mathrm{s}}$ along which the background fluctuates. (**B**) Time series of the synaptic weights $\mathbf{m}$ (single-row matrix M) of the BioPCA neuron, which converge to the elements of the first PC vector multiplied by $\sigma^2\|\mathbf{s}_{\mathrm{s}}\|$ (dashed lines), as predicted in eq. 99. (**C**) Time series of the self-coupling L_{11} of this neuron, which converges to the inverse of the eigenvalue (dashed), eq. 98. (**D**) Time series of the inhibitory weights $\mathbf{w}$ (single-column matrix W), which converge to the first PC vector with a scale given in eq. 100. (**E**) Alignment error of the $\mathbf{m}$ (solid line) and $\mathbf{w}$ (dashed) vectors with the first PC. Both alignment errors are below 0.1 %; $\mathbf{w}$ aligns especially well due to its slow Hebbian dynamics (fluctuations are amplified by nonlinearities in the BioPCA equations). (**F**) PN activity norm, $\|\mathbf{y}(t)\|$ (solid blue line), closely matching the analytical predictions for its average (dark orange) and instantaneous (dashed orange) values given the background trajectory, derived in eq. 101. The response to the background is reduced to ~ 2 % of its original amplitude.

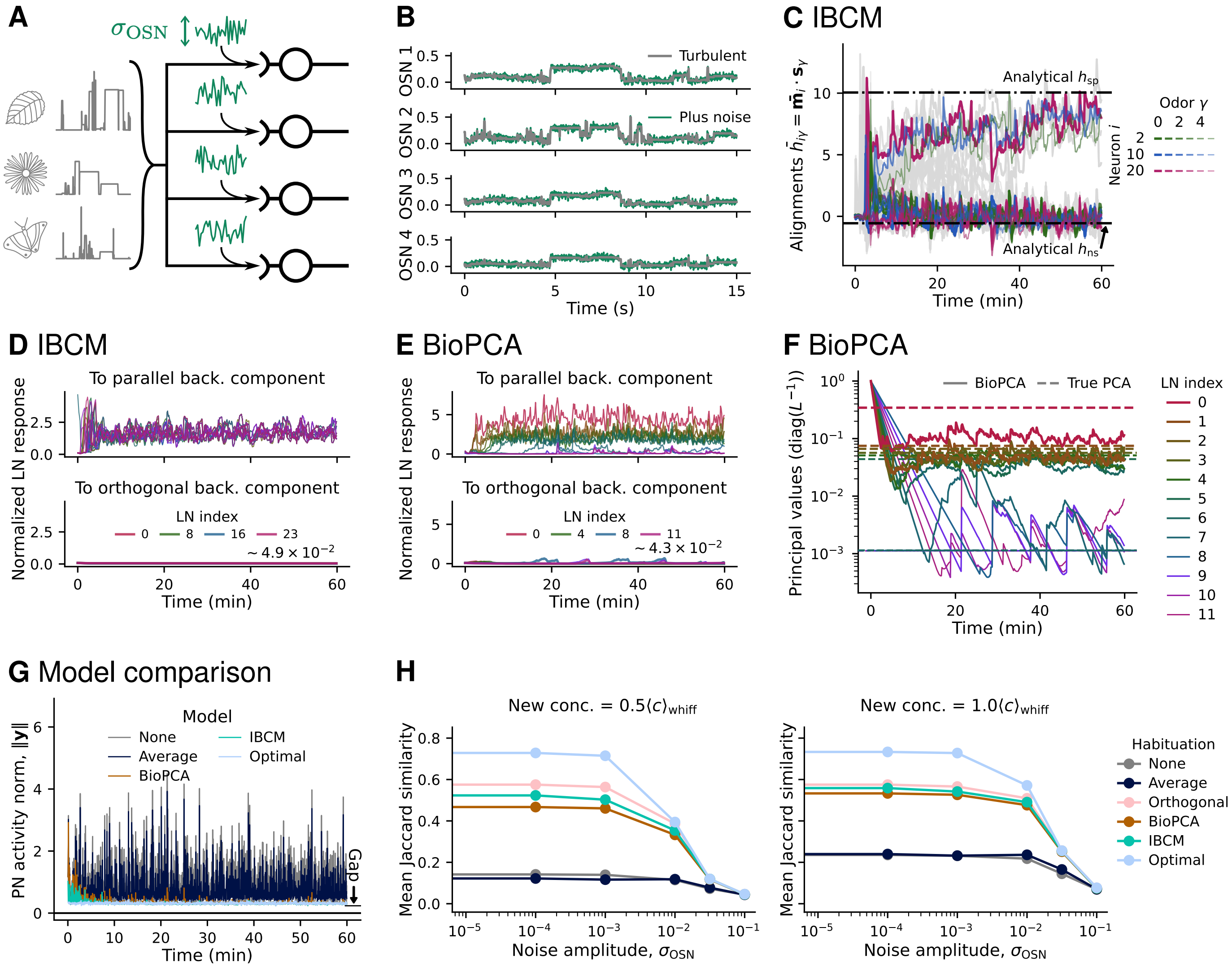


Fig. S8. Robustness against OSN noise in the IBCM and BioPCA manifold learning models. (**A**) Illustration of the OSN noise considered. Turbulent backgrounds with six odors are simulated as usual, then independent Gaussian white noise is added to each OSN input at each time point. We consider simulations in $N_{\mathrm{S}} = 100$ dimensions (100 OSNs). (**B**) Excerpt from the input time series of four OSNs, showing the turbulent background contribution to the input (grey) and the total input with noise added (green). We illustrate the process with noise amplitude $\sigma_{\mathrm{OSN}} = 0.01$; we will consider different amplitudes in panel (H). (**C**) Time series of the alignment of IBCM neurons with background odors, showing that neurons still become selective for true odors while ignoring the added Gaussian noise components. (**D**) Response of each IBCM neuron, $\overline{h}_j(t)$, to the parallel (top) and orthogonal (bottom) components of the background, normalized by the average norm of these components, $\|\mathbf{s}_{//}\|$ or $\|\mathbf{s}_{\perp}\|$. The time series are smoothed with a moving average filter over a window of 3 s, to visualize the average response amplitude. The parallel background component includes the actual turbulent background odors and the component of the Gaussian noise in that subspace; the orthogonal part contains the noise components orthogonal to background odors. We see that IBCM neurons are only responsive to the true olfactory background, not to noise. (**E**) Similar to (D), but for BioPCA neurons, which also respond only to the true background subspace. (**F**) Learning dynamics of the L matrix diagonal elements in the BioPCA network, showing imperfect convergence to the principal values of the background (dashed lines). The first dominant principal component (PC) is missed, but there are $N_{\mathrm{B}} = 7$ neurons roughly capturing the background subspace, while remaining neurons capture the small PCs along the pure Gaussian noise directions (eigenvalues equal to σ_{OSN}). (**G**) Norm of the PN response to the noisy turbulent background over time, showing habituation despite the OSN noise, in all manifold learning models. (**H**) Performance of the various models for new odor recognition as a function of the OSN noise amplitude σ_{OSN}. Numerical experiments similar to those of Fig. 2, repeated for each σ_{OSN} value across 64 backgrounds, 100 new odors, 5 test times and 4 background samples at each time. Lines indicate the average Jaccard similarity, and shaded areas, the standard deviation across repeats. For small noise, the performance remains unchanged, then it rapidly degrades for all models as the OSN noise becomes comparable in magnitude to the new odors. We used $N_{\mathrm{I}} = 24$ IBCM neurons, and $N_{\mathrm{I}} = 12$ BioPCA neurons.

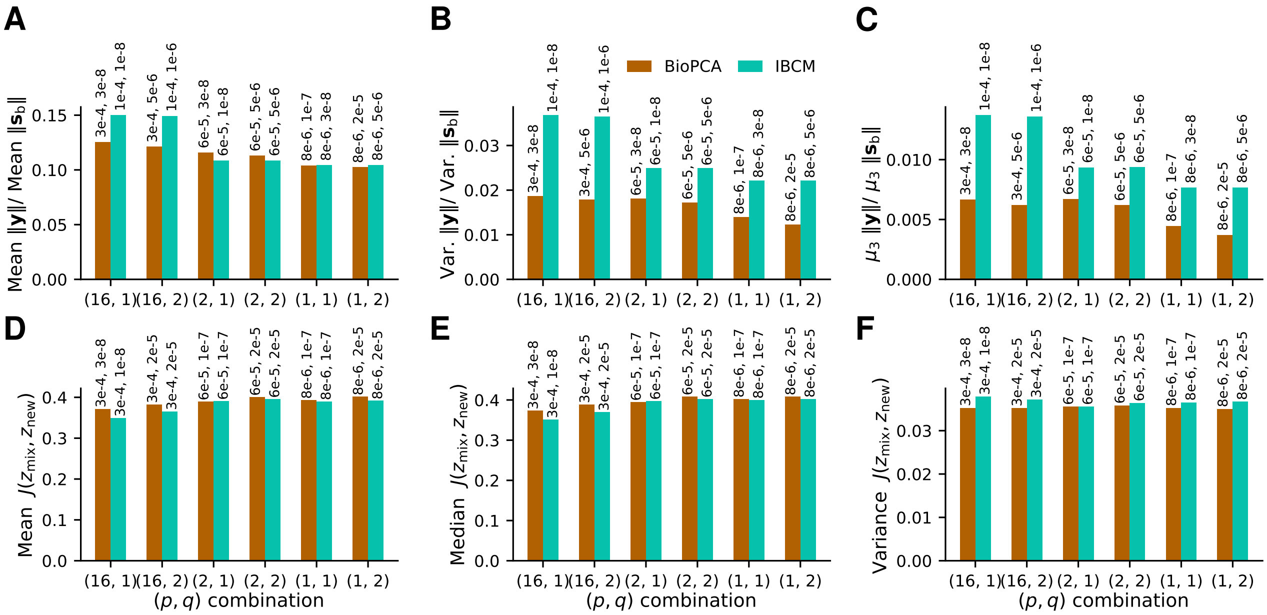


Fig. S9. Performance for various L^p norms in W's cost function learning rule. (**A**) Reduction of the mean, (**B**) variance, (**C**) and third central moment of the norm of the PN response after habituation by IBCM of BioPCA networks. We consider various p, q norm choices in the generalized Hebbian learning rules derived in section 7. For each p, q combination, we performed a grid search over learning rates α and β; the bar graph reports the best performance for each p, q; the α, β rates producing this performance are annotated above the bars. (**D**) Mean and (**E**) median Jaccard similarity between the response to mixtures and the response to the new odor alone, after habituation. We test for various p, q combinations as in (A)-(C). (**F**) Variance in the Jaccard similarity; lower variance is better when the mean or median similarity is high – it signifies that the network is more consistent across backgrounds and trials.

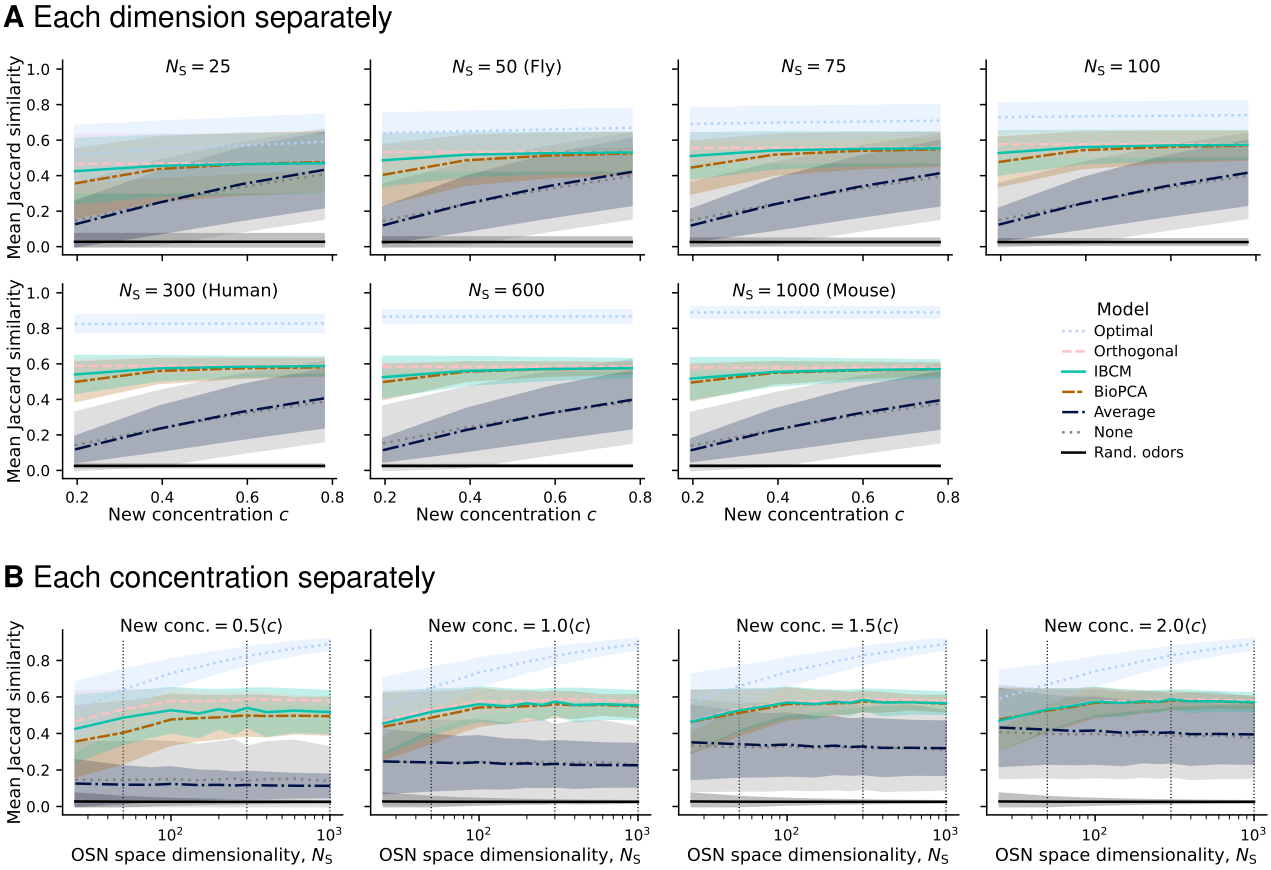


Fig. S10. Odor recognition performance as a function of dimensionality and new odor concentration. Supplement to main Figure 5. (**A**) Jaccard similarity as a function of new odor concentration, for each tested dimensionality separately. (**B**) Jaccard similarity as a function of dimensionality, for each tested new odor concentration. Lines indicate the mean Jaccard similarity, and shaded areas, the standard deviation across repeats. "Rand. odors" indicates the similarity between two odors drawn at random, to show that the similarity of the response to the new odor is significantly higher than the similarity that would occur by chance.

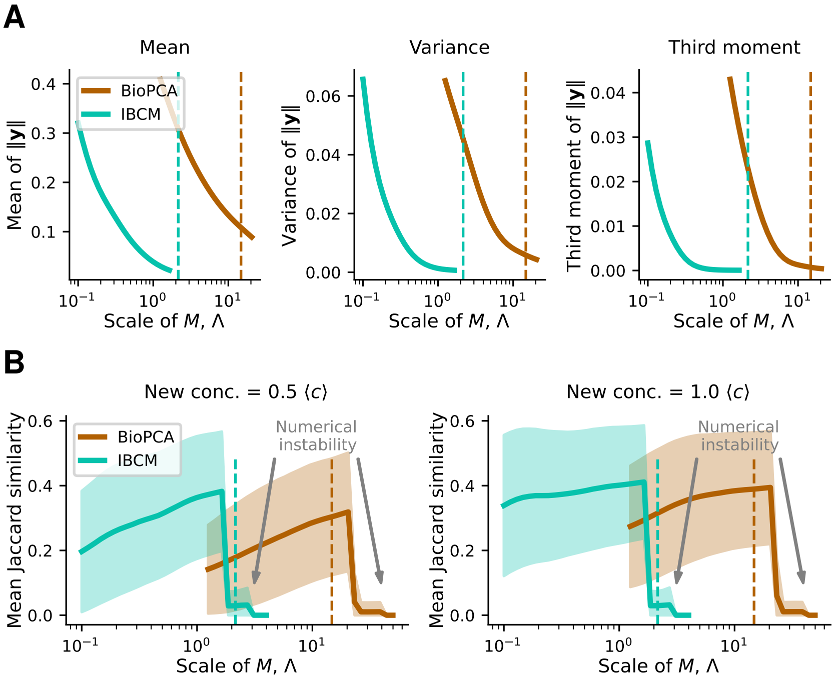


Fig. S11. Effect of the M weights scaling parameter Λ, to match the IBCM and BioPCA models. (**A**) Habituation performance, quantified by reduction in the mean, variance, and third moment of the PN response to the background, $\mathbf{y}$. Habituation runs are performed on the same background time series, with different Λ values for each model. (**B**) New odor recognition performance, quantified by the increase in mean Jaccard similarity between the response to mixtures and the new odor tag, for two new odor concentrations. Same background example as (A), testing 100 new odors at the end (shaded area: standard deviation across tested odors). BioPCA and IBCM can be made to perform equally well by setting Λ appropriately ($\Lambda \sim 10$ times larger for BioPCA). Vertical dashed lines indicate the scale at which numerical instabilities are expected to arise according to a nonlinear stability analysis of the Euler integrator (section 2F). Simulations were run in $N_{\mathrm{S}} = 25$ dimensions and with default parameter values.

Table S1. Definition of olfactory network model parameters

Type	Symbol	Definition
Abbreviations	OSN	Olfactory sensory neuron (input layer)
	PN	Projection neuron (second layer, fly)
	M/T	Mitral/tufted cells (second layer, mouse)
	KC	Kenyon cells (third layer, fly)
	PC	Pyramidal cells (third layer, mouse)
	LN	Lateral inhibitory interneuron (inhibitory layer)
	IBCM	Intrator, Bienenstock, Cooper, Munro (inhibitory layer)
	O-U	Ornstein-Uhlenbeck process
	PCA	Principal components analysis
Dimensions	N_{S}	Number of OSN types, dimension of input and PN layers
	N_{I}	Number of inhibitory neurons
	N_{B}	Number of background odor components
	N_{K}	Number of Kenyon cells in the tag layer
Weight matrices	M	$N_{\mathrm{I}} \times N_{\mathrm{S}}$ projection weights matrix, rows are $\mathbf{m}_j$
	W	$N_{\mathrm{S}} \times N_{\mathrm{I}}$ inhibitory weights matrix, columns are $\mathbf{w}_j$
	L	$N_{\mathrm{I}} \times N_{\mathrm{I}}$ coupling weights matrix
Other dynamical variables	$\mathbf{s}$	N_{S}-dimensional input vector
	$\mathbf{y}$	N_{S}-dimensional vector of PN activities
	$\mathbf{m}_j$	Projection weight vector of the jth LN (jth column of M)
	h_j	Uncoupled activity of the jth LN neuron, given by $\mathbf{m}_j \cdot \mathbf{s}$
	Θ_j	Uncoupled activity threshold of the jth IBCM neuron
	$\overline{h}_j$	Activity of the jth LN neuron after feedforward coupling
	$\overline{\Theta}_j$	Activity threshold of the jth IBCM neuron after coupling
	$\mathbf{w}_j$	Inhibitory weights out of the jth LN neuron (jth row of W)
	z	Neural tag of an odor projected on Kenyon cells
W rates	α	Learning rate of inhibitory weights W
	β	Decay (regularization) rate of W
IBCM parameters	μ_{IBCM}	Learning rate of projection weights M
	η	Coupling strength of IBCM neurons (off-diagonal L_{ij})
	τ_Θ	Averaging time scale of thresholds Θ_j
	Λ_{IBCM}	Scale of M weights
	A_{sat}	Saturating amplitude of IBCM neuron activity
	k_Θ	Minimum denominator in the adaptive learning rate (Law variant)
	ε	Decay rate of M weights
BioPCA parameters	μ	Learning rate of projection weights M
	μ_L	Learning rate of coupling weights L
	Λ_{PCA}	Scale of weights M
	λ_{range}	Range over which $\underline{\Lambda}$'s diagonal decreases
	μ_{avg}	Rate of averaging for the mean subtracted from BioPCA inputs
Background parameters	$\langle c \rangle$	Average odor concentration (from any distribution)
	σ^2	Variance of odor concentrations (from any distribution)
	m_3	Third moment of odor concentrations (from any distribution)
	$t_{\mathrm{w,min}}$	Lower cutoff of whiff durations (turbulent)
	$t_{\mathrm{w,max}}$	Upper cutoff of whiff durations (turbulent)
	$t_{\mathrm{b,min}}$	Lower cutoff of blank durations (turbulent)
	$t_{\mathrm{b,max}}$	Upper cutoff of blank durations (turbulent)
	c_0	Concentration scale in turbulent $p_c(c)$ (turbulent)
	α_c	Lower concentration plateau in $p_c(c)$ (turbulent)
	$\langle g \rangle$	Average of O-U process $g(t)$
	σ_g^2	Variance of O-U process $g(t)$
	τ_{b}	Autocorrelation time (O-U-based backgrounds)
	ϵ	Third moment factor in $c = g + \epsilon g^2$ (weakly non-Gaussian)

Table S2. Value of model and simulation parameters. Fig. S2 uses the same parameters as 2. Fig. S3 uses the same parameters as the simulations related to each panel (A: Fig. 3, B: Fig. S6, C: Fig. 2). Fig. S10 uses the same parameters as 5. Time steps in simulations correspond to 10 ms (such that $360,000$ steps make 60 minutes). Λ^* is the Λ_{PCA} value predicted to make it equivalent to IBCM, given in eq. 106.

Type	Param.	Figure										
		2	3	4	5	S4	S5	S6	S7	S8	S9	S11
Dims.	N_{S}	25	25	25	50–10^3	25	25	25	25	100	25	25
	N_{B}	6	3	6	6	2	3	6	2	6	6	6
	$N_{\mathrm{K}}/1000$	1	-	-	1–40	-	-	1	-	4	1	1
Rates	Duration (min)	60	53	53	60	13	40	53	13	60	60	60
	$\alpha/10^{-4}$	1	2.5	1	1	2.5	2.5	1	2.5	1	Vary	1
	$\beta/10^{-4}$	0.2	0.5	0.2	0.2	0.5	0.5	0.2	0.5	0.2	Vary	0.2
IBCM	N_{I}	24	6	24	24	2	32	24	-	24	24	24
	$\mu_{\mathrm{IBCM}}/10^{-3}$	1.25	1.5	1.25	0.75	2.5	2.5	0.75	-	0.75	1.25	1.25
	τ_Θ (steps)	1600	200	1600	2000	300	150	200	-	2000	1600	1600
	$\eta/10^{-2}$	2.5	25/3	2.5	2.5	20	1.56	2.1	-	2.5	2.5	2.5
	Λ_{IBCM}	1	1	1	1	1	1	1	-	1	1	Vary
	A_{sat}	50	∞	50	50	∞	∞	50	-	50	50	50
	k_Θ	0.1	-	0.1	0.1	-	-	-	-	0.1	0.1	0.1
	$\varepsilon/10^{-2}$	0.5	0	0.5	0.5	0.5	0	0	-	0.5	0.5	0.5
BioPCA	N_{I}	6	-	-	6	-	-	6	1	12	10	6
	$\mu_{\mathrm{PCA}}/10^{-4}$	1	-	-	1	-	-	5	5	1	1	1
	$\frac{\Lambda_{\mathrm{PCA}}^2}{\mu_{\mathrm{PCA}}}\mu_L$	2	-	-	2	-	-	2	2	2	2	2
	Λ_{PCA}	Λ^*	-	-	Λ^*	-	-	Λ^*	5	Λ^*	Λ^*	Vary
	λ_{range}	0.5	-	-	0.5	-	-	0.8	0.5	0.5	0.5	0.5
	$\mu_{\mathrm{avg}}/10^{-4}$	1	-	-	1	-	-	5	5	1	1	1
Turbulent back.	τ_{w} (step)	1	-	1	1	-	-	-	-	1	1	1
	$T_{\mathrm{max,w}}$ (step)	500	-	500	500	-	-	-	-	500	500	500
	τ_{b} (step)	1	-	1	1	-	-	-	-	1	1	1
	$T_{\mathrm{max,b}}$ (step)	800	-	800	800	-	-	-	-	800	800	800
	c_0	0.6	-	0.6	0.6	-	-	-	-	0.6	0.6	0.6
	α_c	0.5	-	0.5	0.5	-	-	-	-	0.5	0.5	0.5
O-U-based	$\langle g \rangle$	-	$1/\sqrt{3}$	-	-	0	$1/\sqrt{3}$	-0.5	0	-	-	-
	σ_g^2	-	0.09	-	-	0.09	0.09	0.09	0.09	-	-	-
	τ_{b} (steps)	-	2	-	-	2	2	2	2	-	-	-
	ϵ	-	0.2	-	-	0	0, 0.2	0	0	-	-	-