

# Manifold Learning for Olfactory Habituation to Strongly Fluctuating Backgrounds

François X. P. Bourassa<sup>1,2,3,\*</sup>, Paul François<sup>3,4</sup>, Gautam Reddy<sup>1,†</sup> and Massimo Vergassola<sup>5,6</sup>

<sup>1</sup>Joseph Henry Laboratories of Physics, [Princeton University](#), Princeton, New Jersey 08544, USA

<sup>2</sup>Department of Physics, [McGill University](#), 3600 rue University, Montréal, Québec H3A 2T8, Canada

<sup>3</sup>Département de biochimie et médecine moléculaire, [Université de Montréal](#),  
5155 Chemin de la Rampe, Montréal, Québec H3T 1J4, Canada

<sup>4</sup>[MILA Québec](#), 6666, rue Saint-Urbain, Montréal, Québec H2S 3H1, Canada

<sup>5</sup>[Laboratoire de Physique de l'Ecole normale supérieure \(ENS\)](#), [Université PSL](#),  
CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France

<sup>6</sup>Department of Physics, [University of California, San Diego](#), 9500 Gillman Drive, La Jolla, California 92093, USA



(Received 15 May 2025; accepted 1 December 2025; published 16 January 2026)

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 the Intrator, Bienenstock, Cooper, and Munro (IBCM) model or an online principal components analysis algorithm 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 toward distinguishing plasticity rules in experiments and could be leveraged by other biological circuits facing fluctuating environments.

DOI: [10.1103/q62z-7j1d](https://doi.org/10.1103/q62z-7j1d)

## I. 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. 1(a)]. 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. 1(b)] [9–11]. These strongly non-Gaussian statistics make it nontrivial 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. 1(c)]. 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 ~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].

\*Contact author: [fb3@princeton.edu](mailto:fb3@princeton.edu)

†Contact author: [greddy@princeton.edu](mailto:greddy@princeton.edu)

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](#) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

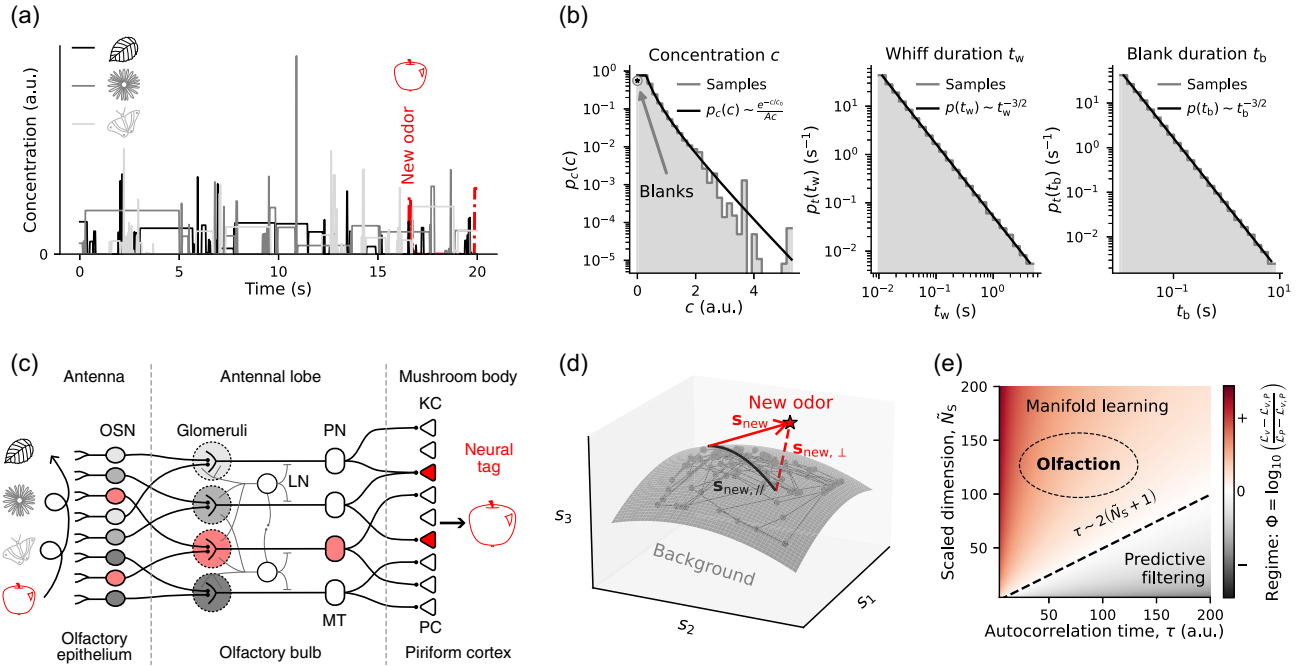


FIG. 1. Olfactory systems face stochastic, turbulent odor mixture inputs, for which manifold learning may be an optimal habituation strategy. (a) Illustration of background and new odor concentrations time series, strongly fluctuating in a series of whiffs and blanks, according to the turbulent atmosphere statistics derived in Ref. [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. MT: Mitral/tufted cell (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,  $s_{\text{new}}$ , generally has a component,  $s_{\text{new},\perp}$ , 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 ( $\mathcal{L}_{v,p}$ ), in the special case of independent, orthogonal background odors  $s_j$  and new odors uniformly distributed on the unit hypersphere (Appendix A). 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:  $\tilde{N}_S = N_S \sigma^2 / \sigma_{\text{new}}^2$ .

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 local interneurons (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 (Appendix C), 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 fluctuating 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. 1(d)]. 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 used by other biological circuits for background subtraction or novelty detection in fluctuating, 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—the Intrator, Bienenstock, Cooper and Munro (IBCM) model [37] and a biologically plausible online principal components analysis (BioPCA) model [38]—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.

## II. RESULTS

### A. 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  $\hat{\mathbf{s}}_\gamma$ ,

$$\mathbf{b}(t) = \sum_{\gamma=1}^{N_B} c_\gamma(t) \hat{\mathbf{s}}_\gamma, \quad (1)$$

where the concentration of the  $\gamma$ th odor at time  $t$  is  $c_\gamma(t)$ . Each vector  $\hat{\mathbf{s}}_\gamma$  encodes how strongly each OSN type is activated by odor  $\gamma$  and could represent a single odorant molecule or a blend coming from the same source with fixed proportions. Here, we have assumed additive odor mixtures to keep our analysis tractable; our conceptual argument should also extend to nonadditive odor mixture coding [17,39–41] 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  $\mathbf{s}_{\text{new}}$  which the target aims to recognize; the total input is  $\mathbf{b}(T) + \mathbf{s}_{\text{new}}$ . An idealized circuit subtracts a vector  $\mathbf{u}_T$  from the target-background mixture, where

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

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  $\mathbf{s}_{\text{new}}$  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  $\mathbf{s}_{\text{new}}$  from the mixture with  $\mathbf{b}(T)$ , which depends on background and new odor statistics (Appendix A). Figure 1(e) illustrates the optimal reconstruction error for different autocorrelation timescales of background fluctuations ( $\tau$ ), and dimensionalities of the olfactory coding space [ $N_S$ , Figs. 7(a) and 7(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  $\tau$ , large  $N_S$ ).

The spectrum of turbulent fluctuations is dominated by brief whiffs and blanks that can reach down to  $\sim 10$  ms [Fig. 1(b)]; 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) [42,43], we estimate that  $\tau < 100$  steps for typical turbulent backgrounds [Fig. 7(c)]. 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 independent 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.

### B. 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 [36,44], we formulate a mathematical description [Fig. 2(a)] of the early olfactory circuit [Fig. 1(c)]. We use *Drosophila* cell types for conciseness, but the model generalizes to other organisms. The key component for habituation is a layer of  $N_I$  LNs 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 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  $WLM$ . Last, PN activities are projected on a large layer of  $N_K$  Kenyon cells (KCs) by sparse random connections (fixed, not learned). The condition  $N_K \gg N_S$  ensures that the 5% most active KCs represent a distinct neural tag  $\mathbf{z}$  for each

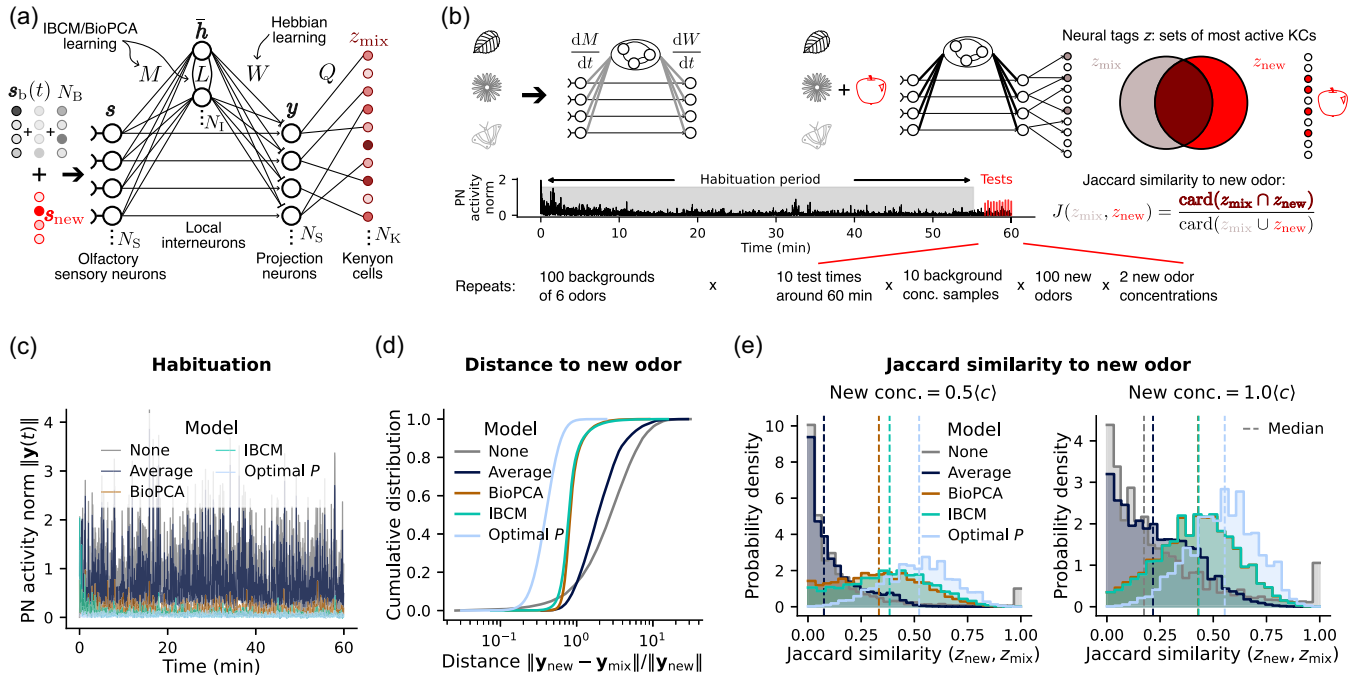


FIG. 2. Recognition of new odors after habituation to a background with different learning models. (a) Mathematical description of the olfactory network. (b) Schematic of the numerical experiments performed to assess habituation performance. We generate a set of background odors randomly, then integrate the network's synaptic plasticity equations for 60 min of habituation to a simulated background time series, where odor concentrations fluctuate according to the turbulent stochastic process illustrated in Figs. 1(a) and 1(b) (see Appendix B for simulation methods). We then compute the network's response to a new odor  $s_{\text{new}}$  mixed with the background. Odor vectors  $s_{\gamma}$ ,  $s_{\text{new}}$  are randomly generated with exponentially distributed elements (Methods). The recognition performance is quantified by the Jaccard similarity between the neural tag of the mixture,  $z_{\text{mix}}$ , and the neural tag (prehabitation) of the new odor alone,  $z_{\text{new}}$ . 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 P": response with the optimal manifold learning matrix  $P$  (no predictive filtering) as in Fig. 1(e) but generalized to encompass nonorthogonal and correlated background odors (Methods). (d) Cumulative distribution of the Euclidean distance between new odors  $s_{\text{new}}$  and each model's PN response to new odors mixed with the background,  $y_{\text{mix}}$ , after habituation, across all background, odors, and new odor concentrations tested. (e) Distribution of Jaccard similarities  $J(z_n, z_{\text{mix}})$  of the various models, across all backgrounds and odor samples tested.

possible odor. This dimensional expansion from PNs to KCs implements locality-sensitive hashing of input identity [44].

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. 1(e) involves nonlocal 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  $y$  by using interneuron activities  $\bar{h}$ . This optimization principle results in simple Hebbian dynamics (Methods),

$$\frac{dW_{ij}}{dt} = \alpha y_i \bar{h}_j - \beta W_{ij}, \quad (3)$$

with learning rate  $\alpha$  and a regularization rate  $\beta$ . This simple update rule allows us to compare different models for the projection weights  $M$  and  $L$ . We consider two models: (a) the IBCM model of synaptic plasticity [37,45–47], and (b) the BioPCA model [38] (for full model definitions, see Methods). The IBCM model was proposed to explain neuronal selectivity to specific stimulus components [45]. Its connections with

independent component analysis (ICA) [48,49] suggest that it could provide a biologically meaningful basis for learning the manifold of non-Gaussian backgrounds.

We perform initial numerical simulations, outlined in Fig. 2(b), to assess the performance of different habituation schemes. We compare these rules to the average subtraction mechanism proposed in Ref. [36] ("Average"), as well as with the absence of habituation ("None") and the optimal manifold learning matrix  $P$  derived in the previous section (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 Figs. 1(a) and 1(b). After this habituation period, we present the network with mixtures of the background and new odors,  $s_{\text{mix}} = s_b(t) + s_{\text{new}}$ . To assess how well that odor is decoded from the mixture, we compare its output with the neural tag  $z_{\text{new}}$  of the new odor alone. The background and new odor vectors are randomly generated for each simulation with exponentially distributed elements and are thus nonnegative and generically nonorthogonal (Methods).

We find that while average subtraction cannot inhibit the strong fluctuations of turbulent backgrounds (Appendix C),

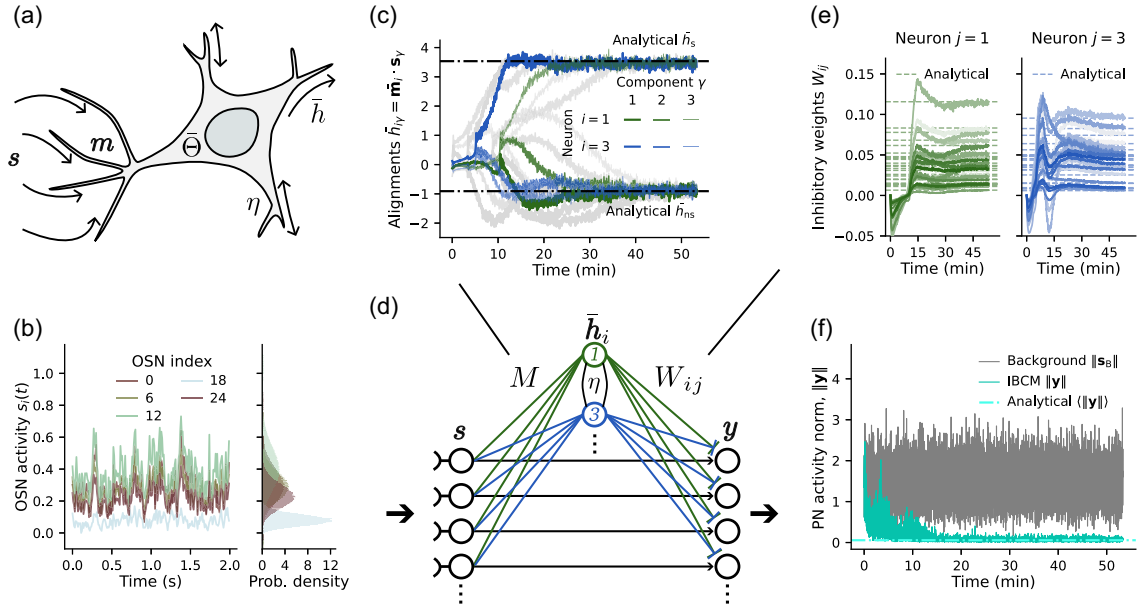


FIG. 3. The IBCM model of synaptic plasticity learns olfactory background components. (a) Illustration of an IBCM neuron's inputs from OSNs,  $s$ , synaptic weights  $m$ , activity  $\bar{h}$ , and internal threshold  $\bar{\Theta}$ . (b) OSN activities for a simple background process with 3 odors and 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  $\gamma$ ). Each neuron predominantly aligns with one odor (large dot product value  $\bar{h}_s$  for one odor, small value  $\bar{h}_{ns}$  for the others). This steady-state alignment matches our analytical predictions (dashed lines, Appendix D). (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 panel (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  $W$ .

both the IBCM and BioPCA networks significantly reduce PN activity in response to the background [Fig. 2(c)], 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  $W$  (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. 2(d)] and neural tags [Fig. 2(e)]. 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. With respect to the distance in PN activity between the new odor alone and mixed with the background [Fig. 2(d)], 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. 2(b), right], the IBCM and BioPCA networks perform within 15% similarity of the optimum [Fig. 2(e)], producing responses much closer to the new odor than to background odors [Figs. 8(a) and 8(b)]. This performance increases with the distance between the new odor and the background manifold, with close pairs ( $\|s_{\text{new},\perp}\| \approx 0.4\|s_{\text{new}}\|$ ) benefiting the least from manifold

projection [Fig. 8(c)]; the neural tag is altered in these cases because the subtracted component of the new odor is significant. Finer metrics than Jaccard similarity would be needed to recover the full odor identity from the remaining component. Still, manifold learning models improve recognition in  $>95\%$  of cases [Fig. 8(d)] and recover, on average, roughly 50% of the KCs which are most activated by the new odor alone and define its identity, even when it is present in the mixture at just half the average whiff concentration [Fig. 2(e), left]. These results prompt us to investigate in more detail how the IBCM and BioPCA models learn the background manifold.

### C. 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 [37,50]. In its simplest form [Fig. 3(a)], the IBCM model describes the slow variation of a neuron's synaptic input weights  $m$  (a row in matrix  $M$ ) as

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

where  $h$  is the activity in response to input  $s(t)$ ,  $\mu$  is the learning rate, and  $\Theta$  is an internal threshold converging to a temporal average  $\Theta = \langle h^2 \rangle$  with an averaging timescale  $\tau_\Theta$

( $\langle \cdot \rangle$ ) denotes averages over the input fluctuations). In practice, we include lateral mean-field inhibition between interneurons with coupling parameter  $\eta$ , a mild nonlinearity to  $h$  preventing excessively large activations, and a small ( $\varepsilon \ll 1$ ) decay term  $-\varepsilon \mu \mathbf{m}$ ; see Methods for the full model equations. For simulations with turbulent backgrounds, we use a variant of the model from Ref. [46] where the learning rate is divided by  $\Theta$  to speed up convergence (Methods).

The IBCM theory rests on the intuition that the  $\mathbf{m}$  equation introduces a competition between input patterns: inputs that cause  $h > \Theta$  are further reinforced, while subthreshold ones are further depressed; consequently, a neuron responds specifically to some inputs and does not respond to others [45]. Based on previously studied settings (e.g., visual cells [47]), this mechanism should work when the threshold timescale  $\tau_\Theta$  is slow enough to average over fast input fluctuations  $\mathbf{s}(t)$ , yet still fast compared to the learning rate  $\mu$ . Remarkably, in our olfactory manifold learning network, the IBCM neurons converged to odor-specific states, analogous to the model's properties in nonolfactory scenarios. The required separation of timescales can be achieved, despite turbulent plumes lasting seconds, to reach habituation within tens of 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 (Appendix D), finding exact expressions for these solutions in terms of the background concentration moments. From a linear stability analysis (Appendix D6 and Fig. 9), 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  $\gamma$  and a small, possibly negative value  $h_{ns}$  ("nonspecific") for all other background components. Hence, an IBCM neuron learns to selectively respond to one background odor: the classical specificity property of this model [37] 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 (Appendix D7). 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  $L$ ) helps to push each neuron toward a different odor component [51].

To confirm our analysis, we perform numerical simulations with a simpler, weakly non-Gaussian background [Fig. 3(b), Methods]. As expected, each IBCM neuron evolves over time to align with one background component [Figs. 3(c) and 3(d)], with steady-state average values of the dot products  $\bar{h}_\gamma$  closely matching our analytical predictions  $h_{sp}, h_{ns}$  (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. 3(e)]. 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

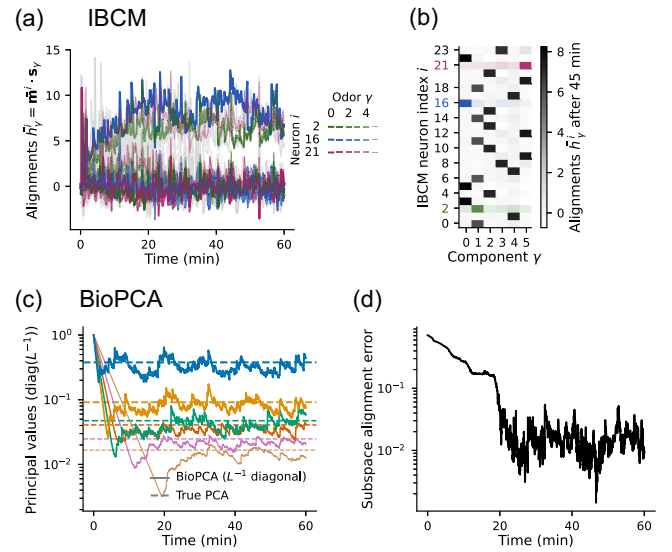


FIG. 4. Habituation of IBCM and BioPCA neurons to turbulent olfactory backgrounds. (a) Time series of the IBCM neurons' synaptic weight alignment with background odor during habituation to a six-odor background with the turbulent concentration stochastic process illustrated in Figs. 1(a) and 1(b). Three neurons are highlighted with colors. (b) Table of each neuron's alignment after habituation, showing that IBCM neurons become selective for one odor even in this strongly fluctuating background. (c) Time series of the principal values learned by local interneurons (LNs) 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  $N_B$  nonzero 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.

before converging to a selective fixed point (Appendices F and G, Figs. 9 and 14). This selectivity is driven by skewness (nonzero third moment) in the background statistics [37,52] (Fig. 11).

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. 12) and in the simulations of Fig. 2 with the full turbulent statistics [Figs. 4(a) and 4(b)]. In the latter case, averaging over long whiffs and blanks requires a slower averaging time  $\tau_\Theta$  and learning rate  $\mu$  (Methods and Table II); despite stronger fluctuations, IBCM neurons align with a single background odor each. We investigate the convergence of the IBCM network as a function of these rates and of the turbulent background parameters in Fig. 15. We find robust convergence across a 10-fold range of rates centered around  $\mu = 0.1 \text{ s}^{-1}$ ,  $\tau_\Theta = 16 \text{ s}$  [Figs. 15(b) and 15(c)]. The model can learn backgrounds up to at least 8 odors [Fig. 15(f)] with whiffs and blanks lasting up to the averaging time  $\tau_\Theta$  [20 s in Fig. 15(g)]. The optimal learning rate, however, depends on the magnitude of OSN



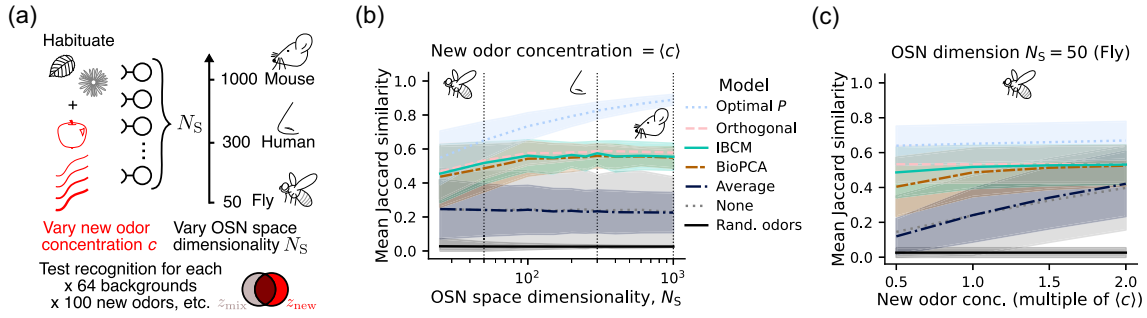


FIG. 5. Recognition performance as a function of olfactory space dimensionality and new odor concentration. See also Fig. 18. (a) For various olfactory space dimensionalities  $N_S$  (i.e., number of OSN types) and new odor concentrations  $c_n$  at test times, we perform simulations like those in Fig. 2: the model habituates to a six-odor turbulent background for  $\sim 60$  min before a new odor is introduced in the mixture. Each  $(N_S, c_{\text{new}})$  condition is tested across 64 backgrounds, 100 new odors, 5 test times post-habitation, and four 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  $N_S$ , for different manifold learning models. Results shown here are for the new odor  $c_{\text{new}}$  equal to the average concentration of background odors,  $\langle c \rangle$  (see Fig. 18 for all concentrations). “Optimal”: manifold learning matrix  $W$  derived in Fig. 1(e). “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 ( $N_S = 50$ ; see Fig. 18(a) for all dimensions). Same legend as panel (b).

activity; we provide an analytical understanding of the IBCM convergence time in Appendix G and Fig. 14. When  $\mu$  is too small, the model does not converge in time, and when  $\mu$  is too large to ensure timescale separation  $\tau_\Theta \ll 1/\mu_\Theta$ , stochastic oscillations arise, as reported in previous (nonolfactory) settings [53] [Figs. 15(d) and 15(e)].

#### D. 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 (Methods, Appendixes E and F3). The matrix  $L$  becomes nearly diagonal, containing the principal values [Figs. 4(c) and 12(b)], while the rows of  $M$  converge to (scaled versions of) the principal component vectors, as evidenced by the error on their alignment decreasing to  $\sim 1\%$  [Figs. 4(d), 12(c), and 13].

On the whole, both models achieve similar habituation levels within  $\sim 30$  min. 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,  $N_I = N_B$ , 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.

Both models produce very similar habituation and odor recognition performances in Fig. 2. They are also similar in their robustness to OSN noise (Fig. 16), and in their performance when combined with alternate Hebbian rules for the

$W$  weights based on different  $L^P$  norms (Appendix H and Fig. 17). The BioPCA model mainly differs from IBCM in its lower sensitivity to the magnitude of background OSN activity but its higher alignment error when the number of background odors increases [Figs. 15(h) and 15(i)]. We provide an analytical understanding of BioPCA convergence dynamics, compared to IBCM, in Appendix G and Fig. 14.

#### E. Performance in various olfactory space dimensions

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. 2(b) for increasingly large olfactory space dimensions ( $N_S$ ) and for a wider range of new odor concentrations [Fig. 5(a)]. 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  $N_S \sim 100$  [Fig. 5(b)]. Remarkably, this plateau corresponds to the similarity between the new odor tag,  $z_{\text{new}}$ , and the tag  $z_{\text{new},\perp}$  of the new odor component orthogonal to the background,  $y_{\text{new},\perp}$  [“orthogonal” pink line, Fig. 5(b)]. This observation clarifies why IBCM and BioPCA networks perform similarly: the local rules for  $W$  (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  $y_{\text{mix}} \approx s_{\text{new},\perp}$ . In comparison, the optimal matrix  $P$  preserves some of the new odor’s parallel component, thus maximizing the recognition of  $s_{\text{new}}$ .

Still, the levels of odor recognition reached by the IBCM and BioPCA models are significant, several standard deviations above chance similarity [black line, Fig. 5(b)]. They also perform well above the average subtraction model, which was

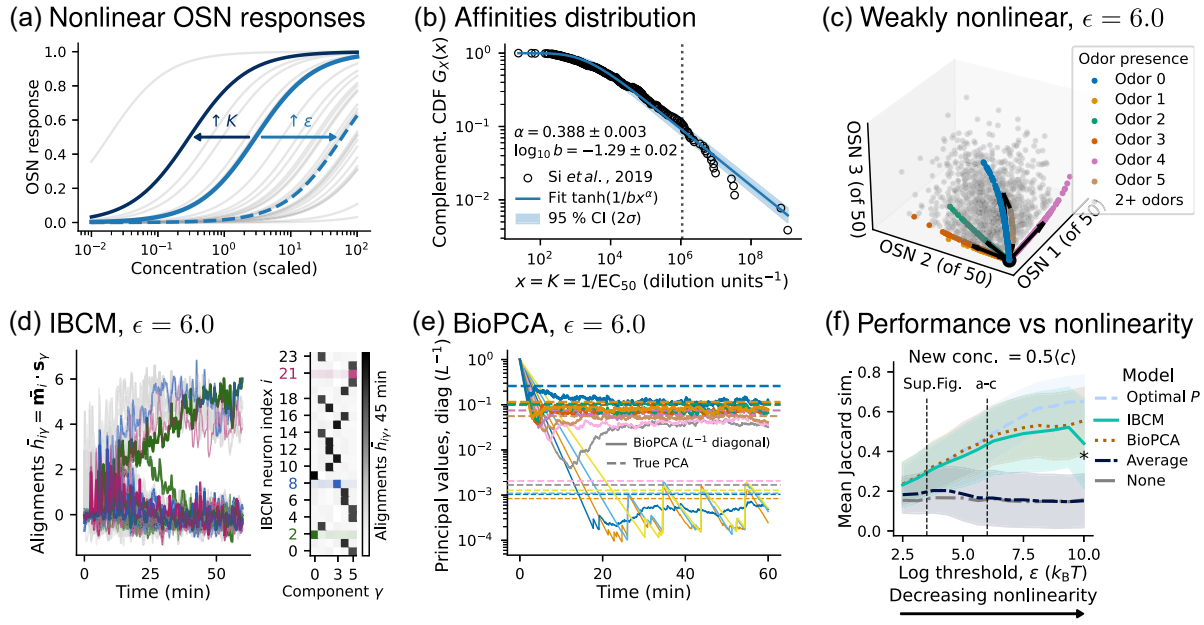


FIG. 6. Habituation and new odor recognition with nonlinear OSN responses. See also Fig. 20. (a) Nonlinear OSN response functions [Eq. (23)] for various affinities  $K$  (grey lines), with highlighted curves showing the effect of changing  $K$  (dark blue) or the saturation log threshold  $\epsilon$  (light blue). (b) Complementary cumulative distribution function (CCDF)  $G_X(x)$  of odor-OR affinities ( $x = \text{inverse EC}_{50}$ ). Empirical values from Si, *et al.*, 2019 [54], and best fit of the form  $G_X(x) = \tanh(1/bx^\alpha)$ . When sampling odors from  $G_X$ , we clip large affinities at the value indicated by the dotted vertical line. (c) Illustration of curved manifolds arising from OSN saturation, for a background manifold with  $N_B = 6$  odors in an olfactory space of  $N_S = 50$  OSN types, for moderate nonlinearity (saturation log threshold  $\epsilon = 6.0$ ). Points are colored according to which odors have a nonzero concentration; arrows indicate vectors of affinities  $\mathbf{K}_\gamma$ . (d) Example IBCM network simulation against the background in panel (c), showing that neurons still align with individual odor vectors  $\mathbf{K}_\gamma$  for moderately nonlinear manifolds. (e) Example BioPCA network simulation for the same background, with  $N_I = 12$  neurons. The network roughly picks out the principal values, although underestimating the largest one and compensating by having  $N_B + 1$  neurons covering  $N_B$  components. Small additional eigenvalues come from the manifold curvature. (f) New odor recognition performance after habituation as a function of the strength of nonlinearity controlled by  $\epsilon$ . The test procedure is similar to Fig. 2; the new odor is mixed in at concentration 0.5 (c). The “Optimal  $P$ ” model indicates the best possible performance for a linear manifold projection.

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 [Figs. 5(c) and 18], 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.

### F. Correlated and nonlinear background manifolds

For simplicity, we have introduced the concept of manifold learning using independent odor concentrations  $c_\gamma(t)$  and linear odor mixtures  $\mathbf{s}_b = \sum_\gamma c_\gamma(t) \hat{\mathbf{s}}_\gamma$ . However, real olfactory spaces have temporal correlations between sources; real OSNs have nonlinear responses to odor mixtures and fast adaptation dynamics. We now relax each of these assumptions to test biologically plausible manifold learning circuits in a wider range of biologically relevant conditions. First, we simulate IBCM and BioPCA habituation to turbulent backgrounds where a pair of odor concentrations have Pearson correlation  $\rho$  in Fig. 19 (Appendix B 1 a). The performance of both models for new odor recognition is not significantly impacted

by temporal correlations between odors [Fig. 19(b)]. For a complete correlation ( $\rho = 1$ ), the background is effectively  $N_B - 1$ -dimensional, and the models learn the corresponding decomposition.

Second, we extend our model to include a nonlinear saturating OSN response function [20,55] where the level of nonlinearity is controlled by a free-energy difference  $\epsilon$  [Fig. 6(a), Appendix I]. Odors are now defined by affinity vectors  $\mathbf{K}$  that we sample from our own fit of the empirical OR affinity distribution measured in Ref. [54] [Fig. 6(b)]. In Figs. 6(c)–6(e), we find that IBCM and BioPCA networks can approximate moderately nonlinear backgrounds. Strongly curved manifolds, however, do reduce their efficacy (Fig. 20). Our Discussion outlines possible improvements.

Third, we consider the impact of OSN adaptation on a fast timescale of  $\tau_a = 250$  ms, by including negative feedback from OSN activity onto the (now dynamical) free-energy differences  $\epsilon_i(t)$  in each receptor type  $i$  [Fig. 21(a)], as in Refs. [20,55]. We find that the combination of OSN adaptation and LN-mediated habituation improves odor recognition performance compared to adaptation alone [Fig. 21(c)]. Conversely, adaptation favors the convergence of IBCM neurons by maintaining OSN activity at a suitable amplitude [Fig. 21(e)]. In conclusion, manifold learning with BioPCA or IBCM interneurons emerges as a robust habituation mechanism in

naturalistic backgrounds displaying correlations, nonlinearities, and multiple timescales.

### III. 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 and correlations such as those encountered in turbulent environments, as well as moderate levels of nonlinear interactions. 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 (or weakly correlated, Fig. 19) varying components in the background mixture.

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 min of exposure to turbulent odor mixtures [Figs. 2(c) and 3(f)]. 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 [39]. 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. 2(c)], 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 LN activity should, conversely, closely track background stimuli in real time to keep inhibiting PN responses [Figs. 16(d) and 16(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 timescale of habituation as they learn the new background.

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  $s_{n,\perp}$  decreases; Figs. 8(c) and 8(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. Odor vectors  $\hat{s}$  and distances between them could be derived from measurements of glomerular responses to odors at standard concentrations, which provide an experimental proxy for OSN type activations (as in Ref. [15] or Ref. [39]), or from complete dose response curves (as in Ref. [54]).

Further theoretical work will also be necessary to refine our proposed implementation of manifold learning in olfactory circuits, and to couple this strategy with other odor recognition mechanisms. Schemes more sophisticated than a Hebbian rule for  $W$  would be necessary to reach the optimal performance promised by manifold learning (Fig. 5) or to fully exploit the biologically relevant projections learned by IBCM neurons. Improved schemes leveraging predictive filtering and temporal statistics could avoid subtracting the new odor component that lies in the background subspace, and thus bypass the limited efficacy of manifold learning for new odors very close to background ones [Fig. 8(c)]. Also, in our study, we mostly focused on linear background manifolds [Eq. (1)] that could be decomposed into linear projections ( $\mathbf{h} = \mathbf{L}\mathbf{M}\mathbf{s}$ ). Conceptually, manifold projection—the idea of removing input projections on the background manifold to isolate new odor components [Fig. 1(d)]—also applies to nonlinear backgrounds. While Fig. 6 shows that linear projection models work against moderate nonlinearities, strongly nonadditive mixtures will require olfactory circuit implementations of algorithms that can learn curved manifolds, such as manifold tiling [57].

The biological assumptions made in our model are supported by previous experimental studies, though further measurements are necessary to validate specific predictions. All connections in our network structure (OSN to PN, OSN to LN, and LN to PN) are abundant in the connectome [12]. History-dependent processing occurs at the level of PNs in flies [22,58] (in addition to OSNs) and the OB in mice [29]. Several studies found lateral inhibitory signals to be essential for 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 [59]. Of note, we neglected feedback of PNs on LNs [31] and instead considered a feedforward network for mathematical simplicity.

By choosing to focus on early olfactory processing, we neglected neuromodulatory inputs [60–62] and feedbacks from higher-level cognitive functions [63,64]. 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 [65] likely help mammals focus their attention on relevant cues rather than on

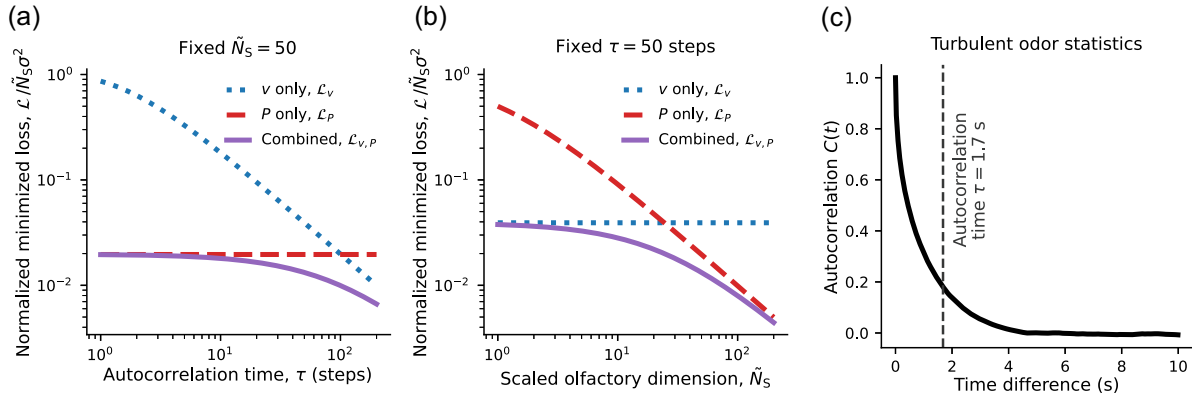


FIG. 7. Comparing manifold learning and predictive filtering. Supplement to Fig. 1(e). (a) Minimized loss function for new odor recognition in a simple background (Ornstein-Uhlenbeck), for the combined strategies (purple) and either single strategy (blue, red), as a function of the autocorrelation time  $\tau$ , 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,  $\tilde{N}_S = N_S\sigma^2/\sigma_{\text{new}}^2$ , 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 Figs. 1(b) and 1(c), showing an autocorrelation time of  $\sim 1.7$  s.

uninformative odors for, e.g., odor trail tracking [66,67]. As a counterpart to habituation, learning the manifold of target odors could be a strategy by which mice rapidly identify previously learned odors within novel backgrounds [68]. Future investigation on this aspect could draw upon recent advances on attention mechanisms in artificial learning models [69].

Beyond olfaction, the interplay between habituation and attention also arises in other biological systems performing chemodetection in fluctuating environments [70]. For instance, in T cell antigen recognition [71], both (immune) memory and (T cell receptor) signal processing networks play important roles for pathogen detection amid a sea of irrelevant (self) antigens. The concept of background manifold projection could also prove useful for algorithms performing figure-ground segregation in time-varying signals, such as in video object detection [72]. 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.

#### IV. MATERIALS AND METHODS

##### A. 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. 1(e), vectors for background ( $\hat{\mathbf{s}}_v$ ) and new ( $\hat{\mathbf{s}}_{\text{new}}$ ) 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  $c_{\text{new}}$ . Background concentrations  $c_v(t)$  follow a stochastic process, usually (Figs. 2 and 5) the turbulent process illustrated in Figs. 1(a) and 1(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  $c_0 = 0.6$ . To test the effect of temporal correlations between odors (Fig. 19), we mixed independent turbulent processes using the Cholesky decomposition of the desired covariance matrix. We also considered weakly non-Gaussian (Fig. 3) and log-normal (Fig. 12) background concentrations, by simulating a multivariate Ornstein-Uhlenbeck  $\{g_v(t)\}$ , then transforming these variables as  $c_v(t) = g_v(t) + v g_v(t)^2$  or  $c_v(t) = 10^{g_v(t)}$ , respectively. We used a short autocorrelation time  $\tau = 20$  ms, an average  $\langle g \rangle = 1/\sqrt{N_B}$ , and standard deviation  $\sigma_g = 0.3$ . For the non-Gaussian case, we set  $v = 0.2$ . Details of the stochastic simulation methods are provided in Appendix B.

##### B. 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,  $\mathbf{b}(T) + \mathbf{s}_n - \mathbf{u}(T)$ , and the target odor alone,  $\mathbf{s}_n$ . The corresponding loss is

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

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  $N_S \times N_S$  matrix  $P$ , as an upper bound on the mechanisms that real networks could learn during habituation. We minimized the loss by solving  $\frac{\partial \mathcal{L}}{\partial v_j} = \frac{\partial \mathcal{L}}{\partial P_{ij}} = 0$ . Appendix A 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.



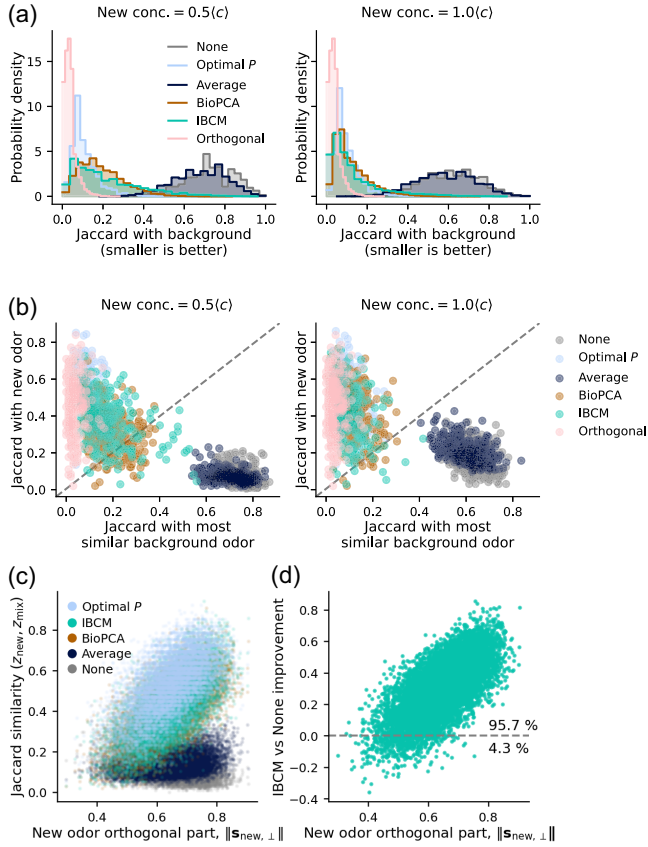


FIG. 8. New odor recognition compared to background recognition after habituation. Supplement to Fig. 2. (a) Histograms of the Jaccard similarity between the response to the mixture (background and new odor) and the most similar background odor, across all repeats (new odors, background samples, etc.) described in Fig. 2(a), 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  $z_{\text{mix}}$  and  $z_{\text{new}}$  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 ( $J_{\text{IBCM}} - J_{\text{None}}$ ). 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 panels (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).

For Fig. 1(e), 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  $\tau$ ,  $\langle c_\gamma(t)c_\rho(t+s) \rangle = \sigma^2 \delta_{\gamma\rho} e^{-|s|/\tau}$ . The minimized loss is then

$$\mathcal{L}_{v,P} = N_B \sigma^2 (1 - e^{-2/\tau}) / (1 + \tilde{N}_S (1 - e^{-2/\tau})), \quad (6)$$

where  $\tilde{N}_S = (\sigma^2/\sigma_n^2)N_S$  and  $\sigma_n^2$  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

$$\mathcal{L}_v = N_B \sigma^2 (1 - e^{-2/\tau}), \quad \mathcal{L}_P = N_B \sigma^2 / (1 + \tilde{N}_S). \quad (7)$$

For the “optimal” strategy in Figs. 2 and 5, we needed to derive the general solution (for nonorthogonal background and new odors drawn from  $\mathcal{P}_S$ ) for pure manifold learning ( $v = 0$ ) when the background has a nonzero average (as was the case in our simulations). In that case, the optimal projection matrix is

$$P = (\langle \mathbf{b}\mathbf{b}^\top \rangle + \langle \mathbf{b} \rangle \langle \mathbf{s}_{\text{new}} \rangle^\top) M^+, \quad (8)$$

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

$$M = \langle \mathbf{b}\mathbf{b}^\top \rangle + \langle \mathbf{s}_{\text{new}} \mathbf{s}_{\text{new}}^\top \rangle + \langle \mathbf{b} \rangle \langle \mathbf{s}_{\text{new}} \rangle^\top + \langle \mathbf{s}_{\text{new}} \rangle \langle \mathbf{b} \rangle^\top. \quad (9)$$

We evaluated numerically the moments  $\langle \mathbf{b}\mathbf{b}^\top \rangle$ ,  $\langle \mathbf{s}_{\text{new}} \mathbf{s}_{\text{new}}^\top \rangle$ , etc., by sampling unit vectors  $\mathbf{s}_\gamma$ ,  $\mathbf{s}_{\text{new}}$  from  $\mathcal{P}_S$  and background concentrations  $c_\gamma$  from the stationary distribution of the background process at hand—for Figs. 2 and 5, the turbulent statistics shown in Figs. 1(a) and 1(b).

### C. Mathematical model of the olfactory network

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

$$\bar{\mathbf{h}}(t) = \phi(L\mathbf{M}\mathbf{s}(t)) \quad (\text{interneurons}), \quad (10)$$

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

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

where  $\phi$  is an element-wise nonlinearity and  $R_\theta$  clips elements below threshold  $\theta$ . We used  $\phi(x) = A_{\text{sat}} \tanh(x/A_{\text{sat}})$  for IBCM neurons to saturate their activity at a large  $A_{\text{sat}} = 50$  for numerical stability; most of the time,  $x \ll A_{\text{sat}}$  is in the linear part of this function, so  $\mathbf{y} \approx \mathbf{s} - W\mathbf{L}\mathbf{M}\mathbf{s}$ . 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 Ref. [36]. First, PN activities are projected to the KC layer by the sparse  $N_K \times N_S$  binary matrix  $Q$ . Then,  $R_\theta$  clips Kenyon cells (KCs) with activity below threshold; we set  $\theta = \frac{1}{60} \times fN_S \times \langle s_i \rangle$ , 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 nonzero 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  $fN_S$  nonzero 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  $N_S/N_K = 50/2000$  found in *Drosophila*; hence, for mice with  $N_S = 1000$  OR types, we used  $N_K = 40\,000$  cortical cells (KC equivalent), and the  $Q$  matrix had  $fN_S = 120$  M/T cells (PN equivalent) projecting to each cortical cell. This ratio aligned with experimental estimates in mice giving  $\sim 200$  glomeruli connected to a cortical cell, or 10% sparsity in  $Q$  [73–75]. The other matrices ( $M$ ,  $W$ ,  $L$  in BioPCA) are slowly updated according to synaptic plasticity rules during a habituation run.

#### D. 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  $W_{ij}$ :

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

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  $\alpha$ ,  $\frac{dW_{ij}}{dt} = -\alpha \frac{\partial \mathcal{L}_W}{\partial W_{ij}}$ , yields the aforementioned Hebbian rule. The average  $\langle \rangle$  is replaced by time averaging over a time window  $1/\alpha$  using a slow rate  $\alpha$  to implement online averaging of fast background fluctuations.

#### E. Average subtraction model

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

$$\frac{d\mathbf{w}_{\text{avg}}}{dt} = \alpha(\mathbf{s}(t) - \mathbf{w}_{\text{avg}}) - \beta\mathbf{w}_{\text{avg}}, \quad (14)$$

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

#### F. 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 Ref. [37], 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  $\bar{\mathbf{h}} = \phi(L\mathbf{M}\mathbf{s})$ ) is

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

where we defined the inhibited weights  $\bar{\mathbf{m}}_i = \mathbf{m}_i - \sum_{j \neq i} \mathbf{m}_j$ , and where  $\phi$  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,

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

where  $\phi'$  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

$$\mu_{\bar{\Theta}_i} = \frac{\mu}{\bar{\Theta}_i + k_{\bar{\Theta}}} \quad (17)$$

This form is similar to the variant introduced in Ref. [46], but we added a constant  $k_{\bar{\Theta}}$  in the denominator to prevent blowups at  $t = 0$ , where we initialize  $\bar{\Theta} = 0$ . For simpler backgrounds (Figs. 3 and 12), we did not include this variant and simply used  $\mu_{\bar{\Theta}} = \mu$ . Moreover, the internal threshold of each neuron,  $\bar{\Theta}_i$ , evolves as

$$\frac{d\bar{\Theta}_i}{dt} = \frac{1}{\tau_{\bar{\Theta}}} ((\bar{h}_i)^2 - \bar{\Theta}_i), \quad (18)$$

such that it tracks the reduced neuron activity  $\bar{h}_i$  averaged on an intermediate timescale  $\tau_{\bar{\Theta}}$ .

#### G. Network of BioPCA neurons

We used the “inverse-free PSP” version of the BioPCA proposed in Ref. [38] 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 predefined diagonal matrix  $\underline{\Lambda}$  [[38], Lemma 3]. The model specifies dynamical update rules for  $M$  and  $L' = L^{-1}$ , the inverse of  $L$ , rather than  $L$  directly. To avoid nonbiological matrix inverse computations, the vector of interneuron activities  $\mathbf{h}$  is computed with a Taylor series for  $L$ ,

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

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

$$\frac{dM}{dt} = \mu_M (\bar{\mathbf{h}} \mathbf{s}^T - M), \quad (20)$$

$$\frac{dL'}{dt} = \mu_L (\bar{\mathbf{h}} \bar{\mathbf{h}}^T - \underline{\Lambda} L' \underline{\Lambda}). \quad (21)$$

In practice, we set  $\underline{\Lambda}_{kk} = \Lambda (1 - \frac{\lambda_r(k-1)}{(N_I-1)})$ , where  $\Lambda$  is the scaling factor for  $M$  weights described below to make IBCM and BioPCA perform similarly, and  $\lambda_r$  is the range of  $\Lambda$  values (between 0 and 1). We followed the original paper’s recommendation for the linear decrease of  $\Lambda_{kk}$  with  $k$  and for setting  $\mu_L = 2\mu_M$ . For further comparison with the original paper

TABLE I. 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           | $M$                      | $N_I \times N_S$ projection weights matrix, rows are $\mathbf{m}_j$                  |
|                           | $W$                      | $N_S \times N_I$ inhibitory weights matrix, columns are $\mathbf{w}_j$               |
|                           | $L$                      | $N_I \times N_I$ 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 $j$ th LN ( $j$ th column of $M$ )                   |
|                           | $h_j$                    | Uncoupled activity of the $j$ th LN neuron, given by $\mathbf{m}_j \cdot \mathbf{s}$ |
|                           | $\Theta_j$               | Uncoupled activity threshold of the $j$ th IBCM neuron                               |
|                           | $\bar{h}_j$              | Activity of the $j$ th LN neuron after feedforward coupling                          |
|                           | $\bar{\Theta}_j$         | Activity threshold of the $j$ th IBCM neuron after coupling                          |
|                           | $\mathbf{w}_j$           | Inhibitory weights out of the $j$ th LN neuron ( $j$ th row of $W$ )                 |
|                           | $z$                      | Neural tag of an odor projected on Kenyon cells                                      |
|                           | $\alpha$                 | Learning rate of inhibitory weights $W$                                              |
| W rates                   | $\beta$                  | Decay (regularization) rate of $W$                                                   |
|                           | $\mu_\Theta$             | Adaptive learning rate of projection weights $M$ (Law variant)                       |
| IBCM parameters           | $\eta$                   | Coupling strength of IBCM neurons (off-diagonal $L_{ij}$ )                           |
|                           | $\tau_\Theta$            | Averaging timescale of thresholds $\Theta_j$                                         |
|                           | $\Lambda_{\text{IBCM}}$  | Scale of $M$ weights                                                                 |
|                           | $A_{\text{sat}}$         | Saturating amplitude of IBCM neuron activity                                         |
|                           | $k_\Theta$               | Minimum denominator in the adaptive learning rate (Law variant)                      |
|                           | $\varepsilon$            | Decay rate of $M$ weights                                                            |
| BioPCA parameters         | $\mu$                    | Learning rate of projection weights $M$                                              |
|                           | $\mu_L$                  | Learning rate of coupling weights $L$                                                |
|                           | $\Lambda_{\text{PCA}}$   | Scale of weights $M$                                                                 |
|                           | $\lambda_{\text{range}}$ | Range over which $\Lambda$ 's diagonal decreases                                     |
|                           | $\mu_{\text{avg}}$       | Rate of averaging for the mean subtracted from BioPCA inputs                         |
| Background parameters     | $\langle c \rangle$      | Average odor concentration (from any distribution)                                   |
|                           | $\sigma^2$               | Variance of odor concentrations (from any distribution)                              |
|                           | $m_3$                    | Third moment of odor concentrations (from any distribution)                          |
|                           | $\tau_w$                 | Lower cutoff of whiff durations (turbulent)                                          |
|                           | $T_{\text{max},w}$       | Upper cutoff of whiff durations (turbulent)                                          |
|                           | $\tau_b$                 | Lower cutoff of blank durations (turbulent)                                          |
|                           | $T_{\text{max},b}$       | Upper cutoff of blank durations (turbulent)                                          |
|                           | $c_0$                    | Concentration scale in turbulent $p_c(c)$ (turbulent)                                |
|                           | $\alpha_c$               | Lower concentration plateau in $p_c(c)$ (turbulent)                                  |
|                           | $\langle g \rangle$      | Average of O-U process $g(t)$                                                        |
|                           | $\sigma_g^2$             | Variance of O-U process $g(t)$                                                       |
|                           | $\tau_b$                 | Autocorrelation time (O-U-based backgrounds)                                         |
|                           | $\nu$                    | Third moment factor in $c = g + \nu g^2$ (weakly non-Gaussian)                       |
|                           | $K_{i\gamma}$            | Affinity of OR type $i$ for odor $\gamma$                                            |
|                           | $\epsilon_i$             | Free-energy difference of OSN type $i$                                               |
|                           | $F_{\text{max}}$         | Maximum OSN activity                                                                 |
|                           | $\tau_a$                 | OSN adaptation time constant (250 ms)                                                |

TABLE II. Value of model and simulation parameters. Figure 8 uses the same parameters as 2. Figure 9 uses the same parameters as the simulations related to each panel ((a): Fig. 3, (b): Fig. 12, (c): Fig. 2). Figure 18 uses the same parameters as 5. Time steps in simulations correspond to 10 ms (such that 360 000 steps make 60 min).  $\Lambda^*$  is the  $\Lambda_{\text{PCA}}$  value predicted to make it equivalent to IBCM, given in Eq. (J4).

| Type            | Param.                                                  | Figure      |              |      |                    |             |          |              |             |      |              |       |             |             |             |      |      |
|-----------------|---------------------------------------------------------|-------------|--------------|------|--------------------|-------------|----------|--------------|-------------|------|--------------|-------|-------------|-------------|-------------|------|------|
|                 |                                                         | 2, 8        | 3            | 4    | 5, 18              | 6, 20       | 10       | 11           | 12          | 13   | 14           | 15    | 16          | 17          | 19          | 21   | 22   |
| Dims.           | $N_S$                                                   | 25          | 25           | 25   | 50–10 <sup>3</sup> | 50          | 25       | 25           | 25          | 25   | 25           | 25    | 100         | 25          | 50          | 50   | 25   |
|                 | $N_B$                                                   | 6           | 3            | 6    | 6                  | 6           | 2        | 3            | 6           | 2    | 3            | Vary  | 6           | 6           | 4           | 6    | 6    |
|                 | $N_K/1000$                                              | 1           | –            | –    | 1–40               | 2           | –        | –            | 1           | –    | –            | –     | 4           | 1           | 2           | 2    | 1    |
| Rates           | Duration (min)                                          | 60          | 53           | 53   | 60                 | 60          | 13       | 40           | 53          | 13   | 60           | 60    | 60          | 60          | 60          | 60   | 60   |
|                 | $\alpha/10^{-4}$                                        | 1           | 2.5          | 1    | 1                  | 0.5         | 2.5      | 2.5          | 1           | 2.5  | 1            | 0.5   | 1           | Vary        | 0.5         | 0.5  | 1    |
|                 | $\beta/10^{-4}$                                         | 0.2         | 0.5          | 0.2  | 0.2                | 0.1         | 0.5      | 0.5          | 0.2         | 0.5  | 0.2          | 0.1   | 0.2         | Vary        | 0.1         | 0.1  | 0.1  |
| IBCM            | $N_I$                                                   | 24          | 6            | 24   | 24                 | 24          | 2        | 32           | 24          | –    | 9            | 24    | 24          | 24          | 24          | 24   | 24   |
|                 | $\mu_{\text{IBCM}}/10^{-3}$                             | 1.25        | 1.5          | 1.25 | 0.75               | 0.75        | 2.5      | 2.5          | 0.75        | –    | 1.2          | Vary  | 0.75        | 1.25        | 1.0         | 1.0  | 1.25 |
|                 | $\tau_\Theta$ (steps)                                   | 1600        | 200          | 1600 | 2000               | 1600        | 300      | 150          | 200         | –    | 200, 150     | Vary  | 2000        | 1600        | 1200        | 1600 | 1600 |
|                 | $\eta/10^{-2}$                                          | 2.5         | 25/3         | 2.5  | 2.5                | 2.9         | 20       | 1.56         | 2.1         | –    | 0            | 2.5   | 2.5         | 2.5         | 2.5         | 2.9  | 2.5  |
|                 | $\Lambda_{\text{IBCM}}$                                 | 1           | 1            | 1    | 1                  | 1           | 1        | 1            | 1           | –    | 1            | 1     | 1           | 1           | 1           | 1    | Vary |
|                 | $A_{\text{sat}}$                                        | 50          | $\infty$     | 50   | 50                 | 50          | $\infty$ | $\infty$     | 50          | –    | inf          | 50    | 50          | 50          | 50          | 50   | 50   |
|                 | $k_\Theta$                                              | 0.1         | –            | 0.1  | 0.1                | 0.1         | –        | –            | –           | –    | –            | 0.1   | 0.1         | 0.1         | 0.1         | 0.1  | 0.1  |
| BioPCA          | $\varepsilon/10^{-2}$                                   | 0.5         | 0            | 0.5  | 0.5                | 0.5         | 0.5      | 0            | 0           | –    | 0.5          | 0     | 0.5         | 0.5         | 0.5         | 0.5  | 0.5  |
|                 | $N_I$                                                   | 6           | –            | –    | 6                  | 6, 7 (E)    | –        | –            | 6           | 1    | 3            | $N_B$ | 12          | 10          | 5           | 12   | 6    |
|                 | $\mu_{\text{PCA}}/10^{-4}$                              | 1           | –            | –    | 1                  | 1           | –        | –            | 5           | 5    | 0.2          | Vary  | 1           | 1           | 1           | 1    | 1    |
|                 | $\frac{\Lambda_{\text{PCA}}^2}{\mu_{\text{PCA}}} \mu_L$ | 2           | –            | –    | 2                  | 2           | –        | –            | 2           | 2    | 2            | 2     | 2           | 2           | 2           | 2    | 2    |
|                 | $\Lambda_{\text{PCA}}$                                  | $\Lambda^*$ | –            | –    | $\Lambda^*$        | $\Lambda^*$ | –        | –            | $\Lambda^*$ | 5    | 9            | 9     | $\Lambda^*$ | $\Lambda^*$ | $\Lambda^*$ | 9    | Vary |
|                 | $\lambda_{\text{range}}$                                | 0.5         | –            | –    | 0.5                | 0.5         | –        | –            | 0.8         | 0.5  | 0.3          | 0.3   | 0.5         | 0.5         | 0.5         | 0.5  | 0.5  |
|                 | $\mu_{\text{avg}}/10^{-4}$                              | 1           | –            | –    | 1                  | 1           | –        | –            | 5           | 5    | 1            | 1     | 1           | 1           | 1           | 1    | 1    |
| Turbulent back. | $\tau_w$ (step)                                         | 1           | –            | 1    | 1                  | 1           | –        | –            | –           | –    | –            | 1     | 1           | 1           | 1           | 1    | 1    |
|                 | $T_{\text{max},w}$ (step)                               | 500         | –            | 500  | 500                | 500         | –        | –            | –           | –    | –            | Vary  | 500         | 500         | 500         | 500  | 500  |
|                 | $\tau_b$ (step)                                         | 1           | –            | 1    | 1                  | 1           | –        | –            | –           | –    | –            | 1     | 1           | 1           | 1           | 1    | 1    |
|                 | $T_{\text{max},b}$ (step)                               | 800         | –            | 800  | 800                | 800         | –        | –            | –           | –    | –            | Vary  | 800         | 800         | 800         | 800  | 800  |
|                 | $c_0$                                                   | 0.6         | –            | 0.6  | 0.6                | 0.6         | –        | –            | –           | –    | –            | Vary  | 0.6         | 0.6         | 0.6         | 0.6  | 0.6  |
| O-U-based       | $\alpha_c$                                              | 0.5         | –            | 0.5  | 0.5                | 0.5         | –        | –            | –           | –    | –            | 0.5   | 0.5         | 0.5         | 0.5         | 0.5  | 0.5  |
|                 | $\langle g \rangle$                                     | –           | $1/\sqrt{3}$ | –    | –                  | –           | 0        | $1/\sqrt{3}$ | –0.5        | 0    | $1/\sqrt{3}$ | –     | –           | –           | –           | –    | –    |
|                 | $\sigma_g^2$                                            | –           | 0.09         | –    | –                  | –           | 0.09     | 0.09         | 0.09        | 0.09 | 0.16         | –     | –           | –           | –           | –    | –    |
|                 | $\tau_b$ (steps)                                        | –           | 2            | –    | –                  | –           | 2        | 2            | 2           | 2    | 4            | –     | –           | –           | –           | –    | –    |
|                 | $\nu$                                                   | –           | 0.2          | –    | –                  | –           | 0        | 0.02         | 0           | 0    | 0.2          | –     | –           | –           | –           | –    | –    |

[38], note the following equivalence between our notation  $\leftrightarrow$  theirs:  $M \leftrightarrow W$ ,  $L \leftrightarrow M^{-1}$ ,  $\mathbf{s} \leftrightarrow \mathbf{x}$ , and  $\mathbf{h} \leftrightarrow \mathbf{y}$ .

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  $(\mathbf{ss}^\top)$ , which still includes the average  $\langle \mathbf{s} \rangle$ . 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. 4(d), we computed, at each time point, the subspace alignment error proposed by Ref. [38],

$$E_{\text{Pro}}(M) = \min_{Q \in \mathbb{O}_{N_S}} \frac{\|FQ - U_{\text{PCA}}\|_{\text{Frob}}^2}{\|U_{\text{PCA}}\|_{\text{Frob}}^2}, \quad (22)$$

where columns of  $U_{\text{PCA}}$  contains the  $N_B$  PCA vectors with nonzero eigenvalues,  $F = (\underline{\Lambda}^{-1}LM)^\top$  contains the eigenvectors learned in the network's projection weights, and  $\|U\|_{\text{Frob}}^2 = \text{Tr}(U^\top U)$  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  $Q = U_F V^\top$  where  $U_F, V$  come from the SVD of  $U_{\text{PCA}} F^\top = U_F \Sigma V^\top$  [76].

## H. Nonlinear OSN response function

To examine nonlinear background manifolds in Figs. 6, 20, and 21, we describe an odor  $\gamma$  by its affinities  $K_{i\gamma}$  for each olfactory receptor (OR) type  $i$ . We consider a Michaelis-Menten saturation of OSN responses to a mixture of odors at concentrations  $c_\gamma(t)$ :

$$s_i(t) = F_{\text{max}} \frac{\sum_\gamma K_{i\gamma} c_\gamma(t)}{\exp(\epsilon_i) + \sum_\gamma K_{i\gamma} c_\gamma(t)}. \quad (23)$$

We sample affinities  $K_{i\gamma}$  independently from our fit of the empirical distribution measured by Si, *et al.* [54] [Fig. 20(b)]. In Figs. 6 and 20, we set  $\epsilon_i = \epsilon \forall i$  and control the level of nonlinearity with this parameter  $\epsilon$ .  $F_{\text{max}}$  is scaled with  $\epsilon$  to maintain the average OSN activity in response to a strong whiff of all odors at  $2.5 \langle c \rangle$ . For more details, see Appendix I.



### I. Scaling parameter $\Lambda$ for $M$ weights

In the BioPCA model, the scale parameter  $\Lambda$  in the  $\underline{\Delta}$  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 [ $\beta$  term in Eq. (13)] and thus further reduce PN activity. We set  $\Lambda$  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 posthabituation PN activity; see Appendix J for detailed expressions. Of note, we rescaled the  $\mu_L$  rate to  $\mu_L/\Lambda^2$  in the BioPCA model [Eq. (21)] to preserve exactly the same learning dynamics for any  $\Lambda$ , just with  $M$  weights scaled up or down.

For comparison, we introduced a similar scale parameter  $\Lambda_{\text{IBCM}}$  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. 22). Similar to BioPCA, scaling of the learning rate  $\mu$  was required for  $\Lambda_{\text{IBCM}} \neq 1$  (Appendix J for details).

### J. 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  $N_S = 25$  dimensions,  $N_K = 40N_S$  Kenyon cells,  $N_B = 6$  background odors,  $N_I = 24$  IBCM neurons or  $N_I = N_B$  BioPCA neurons. For  $W$  Hebbian learning, we used  $\alpha = 10^{-4}$  and  $\beta = 2 \times 10^{-5}$ . However, for Fig. 3 and Fig. 12, we used  $N_B = 3$ ,  $N_I = 6$  (IBCM),  $\alpha = 2.5 \times 10^{-4}$ , and  $\beta = 5 \times 10^{-5}$ .

For the IBCM weights, we used by default  $\mu = 1.25 \times 10^{-3}$ ,  $\tau_\Theta = 1600$ ,  $\eta = 0.6/N_I$ ,  $k_\Theta = 0.1$ ,  $\varepsilon = 0.005$ , and  $A = 50$  as the maximum amplitude of the nonlinearity  $\phi$ . For the simple background in Fig. 3, we used  $\mu = 1.5 \times 10^{-3}$ ,  $\tau_\Theta = 200$ ,  $\eta = 0.5/N_I$ , we did not apply the  $\phi$  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  $\underline{\Delta}_{kk}$  entries). For Fig. 4, we fixed  $\Lambda_{\text{PCA}} = 8$  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 I and II.

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  $\Lambda_{\text{PCA}}/\sqrt{N_S}$  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.

All simulation code supporting this study is publicly available on Github [77].

### ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation, through the Center for the Physics of Biological Function (Grant No. PHY-1734030). F.X.P.B. was supported in part by an award from the Fonds de recherche du Québec—Nature et technologies (Award No. 288527). 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. RFINS128865-01. G.R. acknowledges support from a joint research agreement between NTT Research Inc. and Princeton University.

All authors designed research, FXPB performed research, and F.X.P.B. and G.R. wrote the paper with input from P.F. and M.V.

The authors declare no competing interest.

### DATA AVAILABILITY

The data that support the findings of this article are openly available [78].

## APPENDIX A: OPTIMAL MODELS OF MANIFOLD LEARNING AND PREDICTIVE FILTERING

In this Appendix, we detail the optimization problem we solved to delineate regimes of predictive filtering and manifold learning, as shown in Figs. 1(e) and 7.

### 1. Definition of the loss function

We want to minimize the loss function

$$\mathcal{L}_{v,P} = \left\langle \left\| \mathbf{b}_t - \sum_{l=1}^{t-1} v_l \mathbf{b}_{t-l} - P(\mathbf{b}_t + \mathbf{x}) \right\|^2 \right\rangle_{\mathbf{b} \sim \mathcal{P}, \mathbf{x} \sim \mathcal{Q}} \quad (\text{A1})$$

as a function of the scalar coefficients  $v_l$  (predictive filtering) and of the matrix  $P$  (manifold learning). In this subsection, 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  $\mathbf{s}_b$  and  $\mathbf{s}_{\text{new}}$ . The background is a linear combination of predefined odor vectors,  $\hat{\mathbf{y}}_\rho$ , weighted by stochastic concentrations,  $\tilde{c}_{\rho,t'}$ , so  $\mathbf{b}_{t'} = \sum_{\rho=1}^{N_B} \tilde{c}_{\rho,t'} \hat{\mathbf{y}}_\rho$ . The background odor vectors are fixed because habituation is always relative to a given background; the loss is thus a function of the  $\hat{\mathbf{y}}_\rho$ . For simplicity in this conceptual calculation, 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  $\langle \tilde{c}_{\rho,t'} \tilde{c}_{\lambda,t'+s} \rangle = C(s) \delta_{\rho\lambda}$ , with  $C(0) = \sigma^2$ ; together, these concentrations statistics define the background vector distribution  $\mathcal{P}$ . The new odors  $\mathbf{x}$  come from a probability distribution  $\mathcal{Q}$  which we do not need to specify now beyond assuming it has a zero mean and finite covariance matrix  $\langle \mathbf{x} \mathbf{x}^T \rangle$ .

#### a. 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  $\underline{\cdot}$  (= $\underline{\cdot}$ ) for vectors (matrices) in time dimensions. We write out explicitly sums with Greek indices on background odor indices,  $\sum_{\rho=1}^{N_B}$ .

### b. 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

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

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

$$\begin{aligned}\mathcal{L} = & N_B \sigma^2 + N_B \sum_{l,m=1}^{t-1} v_l v_m C(l-m) + \sigma^2 \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho^\top P^\top P \hat{\mathbf{y}}_\rho \\ & + \langle \mathbf{x}^\top P^\top P \mathbf{x} \rangle - 2 N_B \sum_{l=1}^{t-1} v_l C(l) - 2 \sigma^2 \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho^\top P \hat{\mathbf{y}}_\rho \\ & + 2 \left( \sum_{l=1}^{t-1} v_l C(l) \right) \left( \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho^\top P \hat{\mathbf{y}}_\rho \right).\end{aligned}\quad (\text{A2})$$

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.

## 2. Solving for the optimal $P$ and $v$

### a. 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

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

To shorten the notation of terms involving the auto-correlation 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  $\underline{C}_{ij} = C(i-j)$ . We note that  $\underline{C}$  is a Toeplitz matrix (i.e.,  $C_{ij}$  depends on  $i-j$  only) and is symmetric since  $C(i-j) = C(j-i)$ . These properties help to express its inverse explicitly in some cases [79].

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:

$$0 = N_B \underline{C} \underline{v} - \left( N_B - \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho^\top P \hat{\mathbf{y}}_\rho \right) \underline{c}, \quad (\text{A3})$$

$$0 = P \left( \sigma^2 \sum_{\rho} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top + \langle \mathbf{x} \mathbf{x}^\top \rangle \right) - (\sigma^2 - \underline{v}^\top \underline{c}) \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top. \quad (\text{A4})$$

### b. Solving for $P$ in terms of $\underline{v}$

The best solution path is to first solve for  $P$  in terms of  $\underline{v}^\top \underline{c}$ , then solve for  $\underline{v}$ . We define the  $N_S \times N_S$  symmetric matrix

$$M = \sigma^2 \sum_{\rho} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top + \langle \mathbf{x} \mathbf{x}^\top \rangle, \quad (\text{A5})$$

which admits a spectral decomposition  $M = U \Sigma U^\top$  and a Moore-Penrose pseudoinverse  $M^+ = U \Sigma^+ U^\top$ , which is the actual inverse  $M^{-1}$  when  $M$  is invertible (i.e., no zero eigenvalue in  $\Sigma$ ). Equation (A4) thus takes the form  $PM = (\sigma^2 - \underline{v}^\top \underline{c}) \sum_{\rho} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top$ , which can be inverted for  $P_M = PU \Sigma \Sigma^+ U^\top$ , the component of  $P$  in the subspace spanned by  $M$ 's eigenvectors of nonzero 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,

$$P = (\sigma^2 - \underline{v}^\top \underline{c}) \sum_{\rho} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top M^+. \quad (\text{A6})$$

### c. Solving for $\underline{v}$

We can now insert the implicit solution for  $P$  in Eq. (A3) for  $\underline{v}$ . We first evaluate the term

$$\begin{aligned}\sum_{\rho} \hat{\mathbf{y}}_\rho^\top P \hat{\mathbf{y}}_\rho &= \sum_{\rho} \hat{\mathbf{y}}_\rho^\top \left[ (\sigma^2 - \underline{v}^\top \underline{c}) \sum_{\lambda} \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top M^+ \right] \hat{\mathbf{y}}_\rho \\ &= (\sigma^2 - \underline{v}^\top \underline{c}) N_B m_y,\end{aligned}$$

where we have defined

$$m_y = \frac{1}{N_B} \sum_{\rho, \lambda=1}^{N_B} (\hat{\mathbf{y}}_\rho^\top \hat{\mathbf{y}}_\lambda) \hat{\mathbf{y}}_\lambda^\top M^+ \hat{\mathbf{y}}_\rho. \quad (\text{A7})$$

We have not assumed that background odor directions  $\hat{\mathbf{y}}_\rho$  were orthogonal, but if that were the case,  $m_y$  would simplify to  $\frac{1}{N_B} \sum_{\rho} \hat{\mathbf{y}}_\rho^\top M^+ \hat{\mathbf{y}}_\rho$ . We insert this result in Eq. (A3) and divide by  $N_B$  to isolate  $\underline{v}$ . Here, we assume that the autocorrelation matrix  $\underline{C}$  is invertible, and defer the noninvertible case to Appendix A7, because it amounts to a deterministic

background. Here, using the inverse  $\underline{\underline{C}}^{-1}$ , we find

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

This is still an implicit expression because  $\underline{v}^\top \underline{c}$  appears on the right; taking the dot product of this expression with  $\underline{c}$ , we can isolate  $\underline{v}^\top \underline{c}$ , 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),

$$\underline{v} = \frac{1 - \sigma^2 m_y}{1 - \gamma m_y} \underline{\underline{C}}^{-1} \underline{c}, \quad (\text{A8})$$

where we have defined

$$\gamma = \underline{c}^\top \underline{\underline{C}}^{-1} \underline{c}. \quad (\text{A9})$$

#### d. Replacing $\underline{v}$ in the $P$ solution

Having solved for  $\underline{v}$ , we can put it back in Eq. (A6) to obtain an explicit solution for  $P$ ,

$$P = \frac{\sigma^2 - \gamma}{1 - \gamma m_y} \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top M^+. \quad (\text{A10})$$

### 3. Evaluating the loss function at the optimum

For simplicity, we first rewrite the loss function in Eq. (A2) using the underlined vector notation for  $\underline{c}$ ,  $\underline{v}$ ,  $\underline{\underline{C}}$ , giving

$$\mathcal{L}_{v,P} = N_B \sigma^2 + N_B \underline{v}^\top \underline{\underline{C}} \underline{v} + \sigma^2 \sum_{\rho} \hat{\mathbf{y}}_\rho^\top P^\top P \hat{\mathbf{y}}_\rho + \langle \mathbf{x}^\top P^\top P \mathbf{x} \rangle - 2 N_B \underline{v}^\top \underline{c} - 2 \sigma^2 \sum_{\rho} \hat{\mathbf{y}}_\rho^\top P \hat{\mathbf{y}}_\rho + 2 \underline{v}^\top \underline{c} \sum_{\rho} \hat{\mathbf{y}}_\rho^\top P \hat{\mathbf{y}}_\rho. \quad (\text{A11})$$

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. (A5) and (A10)], commuting scalars resulting from intermediate dot products, and renaming indices when appropriate. We find the following terms:

$$\begin{aligned} \underline{v}^\top \underline{\underline{C}} \underline{v} &= \left( \frac{1 - \sigma^2 m_y}{1 - \gamma m_y} \right)^2 \underline{c}^\top \underline{\underline{C}}^{-1} \underline{\underline{C}} \underline{\underline{C}}^{-1} \underline{c} = \left( \frac{1 - \sigma^2 m_y}{1 - \gamma m_y} \right)^2 \gamma, \\ \sigma^2 \sum_{\rho} \hat{\mathbf{y}}_\rho^\top P^\top P \hat{\mathbf{y}}_\rho &= \left( \frac{\sigma^2 - \gamma}{1 - \gamma m_y} \right)^2 \sum_{\lambda, \mu} \hat{\mathbf{y}}_\mu^\top M^+ \left\{ \sigma^2 \sum_{\rho} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top \right\} M^+ \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top \hat{\mathbf{y}}_\mu, \\ \langle \mathbf{x}^\top P^\top P \mathbf{x} \rangle &= \left( \frac{\sigma^2 - \gamma}{1 - \gamma m_y} \right)^2 \sum_{\lambda, \mu} \hat{\mathbf{y}}_\mu^\top M^+ \{ \langle \mathbf{x} \mathbf{x}^\top \rangle \} M^+ \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top \hat{\mathbf{y}}_\mu, \\ \sum_{\rho} \hat{\mathbf{y}}_\rho^\top P \hat{\mathbf{y}}_\rho &= \frac{\sigma^2 - \gamma}{1 - \gamma m_y} \sum_{\rho} \hat{\mathbf{y}}_\rho^\top \sum_{\lambda} \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top M^+ \hat{\mathbf{y}}_\rho = \frac{\sigma^2 - \gamma}{1 - \gamma m_y} N_B m_y. \end{aligned}$$

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

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

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,

$$\mathcal{L}_{v,P} = N_B \frac{(\sigma^2 - \gamma)(1 - \sigma^2 m_y)}{1 - \gamma m_y}. \quad (\text{A12})$$

We notice that the loss seems to scale proportionally with the background subspace dimensions,  $N_B$ . Terms  $\sigma^2$ ,  $\gamma$  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  $m_y = \frac{1}{N_B} \sum_{\rho, \lambda=1}^{N_B} (\hat{\mathbf{y}}_\rho^\top \hat{\mathbf{y}}_\lambda) \hat{\mathbf{y}}_\lambda^\top M^+ \hat{\mathbf{y}}_\rho$ , but we generally expect  $N_B m_y \sim N_B$ . This is especially clear if we assume orthogonality of the  $\hat{\mathbf{y}}$ , such that  $m_y = \frac{1}{N_B} \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho^\top M^+ \hat{\mathbf{y}}_\rho \sim \mathcal{O}(1)$ .

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  $\sim N_B$  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  $\gamma$ , and the dimensionality of the olfactory space, which enters  $m_y$  through the new odor statistics  $\langle \mathbf{x} \mathbf{x}^\top \rangle$  in  $M$ . We expect  $\gamma$  to increase with the autocorrelation timescale, and  $m_y$  to increase with the olfactory space dimension  $N_S$ . Hence, as the autocorrelation time increases,  $\gamma$  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 Appendix A6.

#### 4. Summary of the general optimal solution

We recapitulate the optimization results here. The optimal  $\underline{v}$  and  $P$  are

$$\underline{v} = \frac{1 - \sigma^2 m_y}{1 - \gamma m_y} \underline{C}^{-1} \underline{c}, \quad (\text{A8})$$

$$P = \frac{\sigma^2 - \gamma}{1 - \gamma m_y} \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top M^+, \quad (\text{A10})$$

and they give a minimum loss of

$$\mathcal{L}_{v,P} = N_B \frac{(\sigma^2 - \gamma)(1 - \sigma^2 m_y)}{1 - \gamma m_y}, \quad (\text{A12})$$

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

$$\gamma = \underline{c}^\top \underline{C}^{-1} \underline{c}, \quad (\text{A9})$$

$$m_y = \frac{1}{N_B} \sum_{\rho,\lambda=1}^{N_B} (\hat{\mathbf{y}}_\rho^\top \hat{\mathbf{y}}_\lambda) \hat{\mathbf{y}}_\lambda^\top M^+ \hat{\mathbf{y}}_\rho, \quad (\text{A7})$$

$$\text{in which } M = \sigma^2 \sum_{\rho} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top + \langle \mathbf{x} \mathbf{x}^\top \rangle, \quad (\text{A5})$$

$$\begin{aligned} \underline{c}_i &= C(i) = \langle \tilde{c}_{\rho,t'} \tilde{c}_{\rho,t'+i} \rangle \quad (i \in \{1, 2, \dots, t-1\}), \\ \underline{C}_{ij} &= C(i-j) = \langle \tilde{c}_{\rho,t'+i} \tilde{c}_{\rho,t'+j} \rangle. \end{aligned} \quad (\text{A13})$$

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

##### a. Predictive filtering only: $P = 0$

When  $P = 0$ , we can directly solve Eq. (A3) for  $\underline{v}$ , finding

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

which yields a loss of

$$\mathcal{L}_v = N_B (\sigma^2 - \gamma). \quad (\text{A14})$$

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

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

##### b. 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  $\sigma_{\text{new}}^2$  possibly different from the

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.

##### b. Manifold learning only: $\underline{v} = 0$

When  $\underline{v} = 0$ , we can directly solve Eq. (A4) in terms of  $M^+$ , resulting in

$$P_{v=0} = \sigma^2 \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top M^+,$$

which yields a loss of

$$\mathcal{L}_P = N_B \sigma^2 (1 - \sigma^2 m_y). \quad (\text{A15})$$

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.

#### 6. 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  $\gamma$ ,  $m_y$ , 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. 1(e) and 7.

##### a. Exponential autocorrelation kernel, to evaluate $\gamma$

We suppose that each odor concentration  $\tilde{c}_{\rho,t'}$  is independent of other odors and forms a Gaussian process with exponential autocorrelation kernel  $C(s) = \langle \tilde{c}_{\rho,t'} \tilde{c}_{\rho,t'+s} \rangle = \sigma^2 e^{-|s|/\tau}$  with autocorrelation time  $\tau$  (i.e., the Ornstein-Uhlenbeck process). A small  $\tau$  corresponds to fast fluctuations compared to the timescale of learning. In this case, the symmetric Toeplitz matrix  $\underline{C}_{ij} = \sigma^2 (e^{-1/\tau})^{|i-j|}$  is a Kac-Murdock-Szegő matrix (form  $A_{ij} = r^{|i-j|}$ ,  $r \neq 1$ ), which has an explicit inverse, provided in Ref. [[79], sec. 1.3]. This inverse is the tridiagonal matrix

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

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  $\langle \mathbf{x} \mathbf{x}^\top \rangle = \frac{\sigma_{\text{new}}^2}{N_S} \mathbb{I}$ , with  $\mathbb{I}$  the  $N_S \times N_S$  identity matrix. Additionally, we assume that background odors are orthogonal to each other:  $\hat{\mathbf{y}}_\rho^\top \hat{\mathbf{y}}_\lambda = \delta_{\rho\lambda}$ .

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

$$\tilde{N}_S = \frac{\sigma^2}{\sigma_{\text{new}}^2} N_S \quad (\text{A16})$$



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

$$M^+ = M^{-1} = \frac{\tilde{N}_S}{\sigma^2} \left( \mathbb{I} - \frac{\tilde{N}_S}{\tilde{N}_S + 1} \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top \right).$$

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

$$\begin{aligned} \sum_{\lambda} \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top M^+ &= \frac{\tilde{N}_S}{\sigma^2} \sum_{\lambda} \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top - \frac{\tilde{N}_S^2}{\sigma^2 (\tilde{N}_S + 1)} \sum_{\lambda} \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top \\ &= \frac{\tilde{N}_S}{\tilde{N}_S + 1} \frac{1}{\sigma^2} \sum_{\rho} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top, \end{aligned}$$

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

$$\begin{aligned} m_y &= \frac{1}{N_B} \sum_{\rho, \mu} \hat{\mathbf{y}}_\rho^\top \hat{\mathbf{y}}_\mu \hat{\mathbf{y}}_\mu^\top M^+ \hat{\mathbf{y}}_\rho \\ &= \frac{\tilde{N}_S}{\tilde{N}_S + 1} \frac{1}{\sigma^2} \frac{1}{N_B} \sum_{\rho, \lambda} \hat{\mathbf{y}}_\rho^\top \hat{\mathbf{y}}_\lambda \hat{\mathbf{y}}_\lambda^\top \hat{\mathbf{y}}_\rho = \frac{1}{\sigma^2} \frac{\tilde{N}_S}{\tilde{N}_S + 1}. \end{aligned}$$

### c. Optimal $\underline{v}$ , $P$ and loss function in this background choice

Inserting the above expressions for  $\gamma$ ,  $M^+$ ,  $m_y$ , etc., into the general optimal solution, we find

$$\begin{aligned} v_1 &= \frac{1}{1 + \tilde{N}_S(1 - e^{-2/\tau})} e^{-1/\tau}, \quad v_{j>1} = 0, \\ P &= \frac{\tilde{N}_S(1 - e^{-2/\tau})}{1 + \tilde{N}_S(1 - e^{-2/\tau})} \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top, \end{aligned}$$

and a minimum loss of

$$\mathcal{L}_{v,P} = N_B \sigma^2 \frac{1 - e^{-2/\tau}}{1 + \tilde{N}_S(1 - e^{-2/\tau})}. \quad (\text{A17})$$

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  $\tilde{N}_S$  decreases or  $\tau$  decreases. For small correlation times,  $1 - e^{-2/\tau} \approx 1$ , so the main reduction of the loss comes from the  $\frac{1}{\tilde{N}_S+1}$  factor (also the case if  $\tilde{N}_S$  is large). Meanwhile, for large correlation  $\tau$ , the reduction comes from the numerator  $1 - e^{-2/\tau} \approx 0$ , while the  $\tilde{N}_S$  term in the denominator is rendered ineffective—the same is true if  $\tilde{N}_S$  is small.

### d. 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

$$\begin{aligned} v_{1,P=0} &= e^{-1/\tau}, \quad v_{j \geq 2, P=0} = 0, \\ \mathcal{L}_v &= N_B \sigma^2 (1 - e^{-2/\tau}). \end{aligned} \quad (\text{A18})$$

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

$$P_{v=0} = \sigma^2 \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top M^+ = \frac{\tilde{N}_S}{\tilde{N}_S + 1} \sum_{\rho=1}^{N_B} \hat{\mathbf{y}}_\rho \hat{\mathbf{y}}_\rho^\top, \quad (\text{A19})$$

$$\mathcal{L}_P = \frac{N_B \sigma^2}{\tilde{N}_S + 1}. \quad (\text{A20})$$

### e. Plots of the loss function versus $\tilde{N}_S$ and $\tau$ for the different strategies

We notice that the loss is proportional to  $N_B \sigma^2$  in all strategies for our special background choice, hence we can illustrate the relative efficacy of predictive filtering and manifold learning by plotting  $\mathcal{L}/(N_B \sigma^2)$  as a function of  $\tilde{N}_S$  and  $\tau$ , the two background parameters for which there is actually a transition between the two strategies. Figures 7(a) and 7(b) show the single-strategy losses  $\mathcal{L}_v$  [Eq. (A18)] and  $\mathcal{L}_P$  [Eq. (A20)] compared to the loss for both strategies applied simultaneously,  $\mathcal{L}_{v,P}$  [Eq. (A17)].

We see that even for a small olfactory space dimension,  $\tilde{N}_S = 50$  as in the fruit fly, and moderately long correlation timescales, manifold learning performs much better than predictive filtering, 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  $\tilde{N}_S$ , small  $\tau$ ) or by predictive filtering (small  $\tilde{N}_S$ , large  $\tau$ ). The (smooth) transition occurs when  $\mathcal{L}_P = \mathcal{L}_v$ , which happens at  $1 - e^{-2/\tau} = \frac{1}{\tilde{N}_S+1}$ , which is at  $\tau \approx 2(\tilde{N}_S + 1)$  for large  $\tilde{N}_S$ . This is shown in Fig. 1(c). There is a large region (in red) where manifold learning is the most effective strategy.

## 7. Noninvertible autocorrelation matrix $\underline{C}$

In the above derivations, we assumed that the autocorrelation matrix  $\underline{C}$  was invertible. The case where  $\underline{C}$  has a zero determinant turns out to imply that the stationary process is not truly stochastic but rather a deterministic recurrence relation. To see this, recall that  $\underline{C}$  has elements  $\underline{C}_{ij} = C(i - j) = \langle c_{\tau-i} c_{\tau-j} \rangle$ , where  $c_{\tau-i}$  is the concentration of one of the (i.i.d.) odors, and  $\tau$  can be shifted to any time because the process is stationary. The matrix  $\underline{C}$  is symmetric, real, and positive semidefinite, so its eigenvectors form a complete orthonormal basis; if it is noninvertible, it has at least one zero eigenvalue with (nonzero) eigenvector  $\underline{u}$ . As a result,

$$\begin{aligned} 0 &= \underline{u}^\top \underline{C} \underline{u} = \sum_{l,m=1}^{N_t} u_l u_m C(l - m) \\ &= \sum_{l,m=1}^{N_t} u_l u_m \langle c_{l-l} c_{l-m} \rangle \\ &= \sum_{l,m=1}^{N_t} u_l u_m \langle c_{l+1-l} c_{l+1-m} \rangle \quad (\text{from stationarity}) \end{aligned}$$

$$\begin{aligned}
&= \left\langle \left( \sum_{l=1}^{N_t} u_l c_{t+1-l} \right)^2 \right\rangle \Rightarrow 0 = \sum_{l=1}^{t-1} u_l c_{t+1-l} \\
&\Rightarrow c_t = -\frac{1}{u_1} \sum_{l=2}^{N_t} u_l c_{t+1-l}, \quad (\text{A21})
\end{aligned}$$

showing that the odor concentration at time  $t$  can be exactly predicted from previous samples at times  $2, 3, \dots, t-1$ : this is a deterministic recurrence relation of order  $N_t - 1$ , from which a sequence of  $N_t - 1$  concentrations determines all future ones (due to stationarity). Hence, in this special case, predictive filtering will be able to exactly subtract the background at time step  $t$  based on the previous  $t-1$  samples. Note that if one finds  $u_{t-1} = 0$  above (so  $1/u_{t-1}$  isn't defined), and maybe  $u_{t-2}, u_{t-3}, \dots = 0$  as well, there must eventually be a  $u_{t-k} \neq 0$ ,  $k < t-1$ , since  $\underline{u}$  is not a zero vector. The derivation can then be repeated but shifting indices of  $c_l, c_m$  by  $+k$  instead of  $+1$ , to isolate  $c_t$  in terms of  $c_{k+1}, c_{k+2}, \dots, c_{t-1}$ ; the result is a deterministic recurrence relation of order  $N_t - k$ , based on which predictive filtering still works perfectly.

As a consequence, manifold learning is not required in the nonstochastic scenario where  $\underline{C}$  is not invertible, because predictive filtering can perfectly anticipate the background  $\mathbf{b}_t$ . However, this special case is not the general setting of fluctuating backgrounds we were interested in, and all background processes we considered in this study are truly stochastic, i.e., with an invertible  $\underline{C}$ .

#### a. Predictive filtering coefficients for noninvertible $\underline{C}$

For completeness, we derive the explicit expression of the coefficients  $v_l$  in terms of the null eigenvector elements  $u_l$ . If  $u_k$  is the first nonzero element  $u_l$ , then we have, for each i.i.d. background concentration,  $c_{\rho,t} = -\frac{1}{u_k} \sum_{l=k+1}^{t-1} u_l c_{\rho,t+k-l}$ . Defining the summation index  $m = l - k$ , multiplying by the odor vector  $\hat{y}_\rho$ , and summing over odors  $\rho$ , we recover full background vectors

$$\mathbf{b}_t = -\frac{1}{u_k} \sum_{\rho} \hat{y}_\rho \sum_{m=1}^{t-1-k} u_{m+k} c_{\rho,t-m} = \sum_{m=1}^{t-1-k} \frac{-u_{m+k}}{u_k} \mathbf{b}_{t-m}.$$

Comparing to the argument of the loss function,  $\mathbf{b}_t -$

$\sum_{l=1}^{t-1} v_l \mathbf{b}_{t-l}$ , we read off the coefficients  $v_l$  to make this argument zero and thus  $\mathcal{L} = 0$ :

$$v_l = \begin{cases} -\frac{u_{l+k}}{u_k} & \text{for } 1 \leq l \leq t-1-k, \\ 0 & \text{for } l > t-1-k. \end{cases} \quad (\text{A22})$$

Consequently,  $P = 0$  is the optimal manifold learning matrix in the absence of stochasticity. One can also show that this choice of  $\underline{v}$  is equal to  $\underline{C}^+ \underline{c}$ , where  $\underline{C}^+$  is the pseudoinverse of  $\underline{C}$  (inverse in the nonnull subspace), and  $\underline{c}$  is the vector of autocorrelation values  $C(1), C(2), \dots, C(t-1)$ . This results from setting  $\nabla_v \mathcal{L} = 0$  and using Eq. (A21) to establish that  $\underline{c}$  is orthogonal to zero-eigenvalue eigenvectors:  $\underline{c}^T \underline{u} = 0$ .

## APPENDIX B: SIMULATING BACKGROUND ODOR FLUCTUATIONS

### 1. Turbulent statistics

We use the statistics of whiff durations  $p_{t_w}(t_w)$ , blank durations  $p_{t_b}(t_b)$ , and whiff concentrations  $p_c(c)$  derived in Ref. [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 intrawhiff 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  $\tau_b, \tau_w$ . For numerical stability, we prevent abnormally large durations by imposing also an upper cutoff at  $T_{\max,b}, T_{\max,w}$ , and we use sharp cutoffs, corresponding to a probability density function

$$p_{t_x}(t_x) = \begin{cases} 0 & \text{if } t_x < \tau_x \text{ or } t_x > T_{\max,x} \\ \frac{1}{A_x} \left( \frac{t_x}{\tau_x} \right)^{-3/2} & \text{if } \tau_x \leq t_x \leq T_{\max,x} \end{cases} \quad (x = b \text{ or } w), \quad (\text{B1})$$

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

$$\langle t_x \rangle = \int_{\tau_x}^{T_{\max,x}} \left( \frac{t}{\tau_x} \right)^{-3/2} \frac{t}{A_x} dt = \sqrt{\tau_x T_{\max,x}} \quad (x = b \text{ or } w),$$

so the probability  $\chi$  to be in a whiff is

$$\chi = \frac{\langle t_w \rangle}{\langle t_w \rangle + \langle t_b \rangle} = \left( 1 + \sqrt{\frac{\tau_b T_{\max,b}}{\tau_w T_{\max,w}}} \right)^{-1}.$$

The probability distribution of whiff concentrations  $c$  is therefore 0 with probability  $1 - \chi$  [illustrated by the point at  $c = 0$  in Fig. 1(c), left] or, with probability  $\chi$ , the conditional distribution  $p_c$  given there is a whiff. Hence,  $p_c(c) = (1 - \chi)\delta(c) + \chi p_c(c|\text{whiff})$ . 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. 1(c), 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

$$p_c(c|\text{whiff}) = \begin{cases} \frac{1}{A_\alpha} \frac{e^{-c/c_0}}{c} & \text{if } c \geq \alpha_c c_0 \\ \frac{1}{A_\alpha} \frac{e^{-\alpha_c c}}{\alpha_c c_0} & \text{if } c < \alpha_c c_0 \end{cases}, \quad (\text{B2})$$

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

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

$$F_t(t_x) = \begin{cases} 0 & \text{if } t_x \leq \tau_x \\ \frac{(1-(t_x/\tau_x)^{-1/2})}{1-(T_{\max,x}/\tau_x)^{-1/2}} & \text{if } \tau_x < t_x < T_{\max,x} \\ 1 & \text{if } T_{\max,x} \leq t_x \end{cases} \quad (x = \text{b or w}).$$

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

$$t_x = \frac{\tau_x}{[1 - r(1 - (T_{\max,x}/\tau_x)^{-1/2})]^2} \quad (x = \text{b or w}). \quad (\text{B3})$$

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

The cdf for the conditional whiff concentrations is

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

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

$$c = F_c^{-1}(r) = \begin{cases} \alpha_c A_\alpha c_0 e^{\alpha_c} r & \text{if } r \leq F_c(\alpha_c c_0) = \frac{1}{e^{\alpha_c} A_\alpha} \\ c_0 E_1^{-1}(A_\alpha(1-r)) & \text{if } \frac{1}{e^{\alpha_c} A_\alpha} < r < 1, \end{cases} \quad (\text{B4})$$

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 [80] 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.,  $x = 2 \approx E_1(0.08237)$ ), so we solve for  $z = \log(y)$ . For  $x > 30$ ,  $y$  is small enough ( $y \sim 10^{-13}$ ) to use the approximation  $E_1(y) = -\gamma - \log(y) - \mathcal{O}(y)$  [[81], 6.6.1], where  $\gamma = 0.577 \dots$  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 uniform(0, 1) samples, one for  $t_w$  [using Eq. (B3)], one for  $c$  [using Eq. (B4)]; only one sample is needed when a blank starts, for  $t_w$  (B3). In our simulations, we use the following typical parameter values, the same for all background odors:  $\tau_b = \tau_w = 10 \text{ ms} = 1$  time step for the lower whiff or blank duration cutoff,  $T_{\max,w} = 5000 \text{ ms}$  and  $T_{\max,b} = 8000 \text{ ms}$  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 uniform(0, 1) sample  $r_1$  to

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

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

To sample from these distributions during a simulation, we use the inverse transform method: given a random 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

determine whether there is a whiff, with probability  $\chi$  (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. (B4).

#### a. Correlated turbulent odor sources

To correlate turbulent odor concentrations, we first simulate  $N_B$  independent turbulent concentration processes  $\mathbf{c}'$ . Second, we correlate these processes using the Cholesky decomposition  $R$  of the target covariance matrix, as

$$\mathbf{c} = R(\mathbf{c}' - \langle \mathbf{c}' \rangle) + \langle \mathbf{c}' \rangle \quad \text{where} \quad \sigma^2 R R^T = \text{target covariance}. \quad (\text{B5})$$

Specifically, in our simulations in Fig. 19, we correlated the last two odors, 2 and 3 by convention, while the first two (0 and 1) remain i.i.d. with variance  $\sigma^2$ . In this case, the Cholesky matrix has one  $2 \times 2$  diagonal block and one nondiagonal block equal to

$$R_{23} = \begin{pmatrix} 1 & 0 \\ \rho_{23} & \sqrt{1 - \rho_{23}^2} \end{pmatrix}, \quad (\text{B6})$$

where  $\rho_{23}$  is the Pearson correlation between these odors. Figure 19(a) shows a sample of this process when  $\rho_{23} = 0.7$ .

## 2. Univariate Ornstein-Uhlenbeck process

To simulate a univariate Ornstein-Uhlenbeck (O-U) process  $g(t)$  numerically, we use an exact update rule for finite time steps  $\Delta t$ , derived from the analytical solution of the O-U process. Taking the latest time step as a new deterministic initial condition Ref. [[82], Eq. (2.47)],

$$g(t + \Delta t) = g(t) e^{-\Delta t/\tau_b} + \sqrt{\sigma^2(1 - e^{-2\Delta t/\tau_b})} \xi(t), \quad (\text{B7})$$

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

### 3. Multivariate Ornstein-Uhlenbeck process

The multivariate Langevin equation for the Ornstein-Uhlenbeck process with zero stationary mean,  $\mathbf{g}(t)$ , is covered in Ref. [[83], Sec. 4.5.6], and can be simulated exactly using the same trick as in the univariate case, Eq. (B7). In practice, our simulations with O-U processes used independent and identically distributed background odors, so the update matrices were diagonal and the general simulation method reduced to simulating  $N_B$  zero-mean univariate processes in parallel [using Eq. (B7)], then adding the desired mean vector  $\mathbf{g}_0$  to it.

### 4. Weakly non-Gaussian fluctuating background

We simulate a multivariate Ornstein-Uhlenbeck process,  $\mathbf{g}$  as in Appendix B 3, with zero mean and identically distributed variables with stationary variance  $\sigma_g^2$ . Then, we take  $c_\gamma = g_0 + g_\gamma + \nu g_\gamma^2$ , where  $\nu$  should be chosen small and  $g_0$  is the desired zeroth-order mean concentration. Then, the concentrations have the following moments:

$$\begin{aligned}\langle c \rangle &= g_0 + \nu \sigma_g^2, \\ \text{Var}[c] &= \sigma_g^2 + 2\nu^2 \sigma_g^4, \\ \langle (c - \langle c \rangle)^3 \rangle &= 6\nu \sigma_g^4 + 8\nu^3 \sigma_g^6.\end{aligned}$$

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  $\nu$ .

### 5. Log-normal background fluctuations

As above, we simulate identically distributed O-U variables with variance  $\sigma_g^2$ , 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 [84], the concentrations themselves have a log-normal distribution with moments

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

We test habituation to this background with IBCM and BioPCA networks in Fig. 12.

### 6. Numerical stability

Numerical integration of the  $W$  equations displays instabilities when  $\bar{\mathbf{h}}$  reaches large magnitudes, e.g., when increasing the scale parameter  $\Lambda$  (Fig. 22 and Appendix J). 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,

$$\begin{aligned}W_{t+1} &= W_t + \Delta t(\alpha \mathbf{y} \bar{\mathbf{h}}^T - \beta W_t) \\ &= W_t + \Delta t(\alpha \bar{\mathbf{h}} \bar{\mathbf{h}}^T - \alpha W_t \bar{\mathbf{h}} \bar{\mathbf{h}}^T - \beta W_t)\end{aligned}$$

$$= B \Delta t + W_t(\mathbb{I} + \Delta t J) \quad \text{where } B = \alpha \bar{\mathbf{h}} \bar{\mathbf{h}}^T,$$

$$J = -\alpha \bar{\mathbf{h}} \bar{\mathbf{h}}^T - \beta \mathbb{I}.$$

We consider the worst-case scenario, when  $\bar{\mathbf{h}}$  reaches its maximal magnitude encountered in a simulation,  $\bar{\mathbf{h}} = \mathbf{h}_{\max}$  and  $J_{\max} = -\alpha \mathbf{h}_{\max} \mathbf{h}_{\max}^T - \beta \mathbb{I}$ , 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  $A = U D U^\dagger$  with  $D = \text{diag}(\lambda_i)$ . Then, if  $A$  has at least one eigenvalue with magnitude  $|\lambda_j| > 1$ , then  $A^N = U \text{diag}(\lambda_i^N) U^\dagger$  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  $A = (1 - \beta \Delta t) \mathbb{I} - \alpha \Delta t \|\mathbf{h}_{\max}\|^2 \hat{\mathbf{h}}_{\max} \hat{\mathbf{h}}_{\max}^T$ ,

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

To see this, consider the rotation matrix  $R$  that aligns  $\mathbf{h}_{\max}$  with one of the unit vectors, e.g.,  $(1, 0, \dots, 0)$ :  $R A R^T$  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,  $\lambda_1$  is first to reach a magnitude  $> 1$ , by becoming negative. Hence, we can predict that numerical instabilities arise in the Euler integrator when

$$\lambda_1 < -1 \Rightarrow \|\mathbf{h}_{\max}\|^2 > \frac{2 - \beta \Delta t}{\alpha \Delta t} \quad \text{or} \quad \Delta t > \frac{2}{\beta + \alpha \|\mathbf{h}_{\max}\|^2}. \quad (\text{B8})$$

Given a simulation of the  $M$  weights and  $\bar{\mathbf{h}}$  (which are not destabilized), we can thus anticipate whether the  $W$  integration should be unstable by extracting  $\|\mathbf{h}_{\max}\|$  from the simulation. The vertical lines in Fig. 22 indicate the smallest  $\Lambda$  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 solution is linearly stable for any  $\Lambda$ . Hence, the observed divergences are numerical limitations rather than true model performance drops and would be remedied by decreasing the time step,  $\Delta t$ .

### APPENDIX C: AVERAGE SUBTRACTION MODEL

In this Appendix, we examine the average subtraction model from Ref. [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

$$\frac{d\mathbf{w}}{dt} = \alpha \mathbf{s}(t) - (\alpha + \beta) \mathbf{w}(t) \quad (\text{C1})$$

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  $\mathbf{s}_{b,0}$ , then the inhibition vector  $\mathbf{w}$  converges to  $\mathbf{w}_{ss} = \frac{\alpha}{\alpha + \beta} \mathbf{s}_{b,0}$ ,



so the background is then perfectly subtracted. However, this strategy fails if the background vector  $\mathbf{s}_b(t)$  fluctuates randomly over time. In this case, Eq. (C1) 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,

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

assuming  $\mathbf{s}_b(t)$  started at a given initial value  $\mathbf{s}_{b,0}$ . 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,  $\langle \mathbf{w} \rangle = \frac{\alpha}{\alpha+\beta} \langle \mathbf{s}_b \rangle$ . If the background process has an autocorrelation timescale much faster than the learning rate,  $\tau_b \ll \frac{1}{\alpha}$ , with its elements approximately obeying  $\langle \Delta s_i(t_1) \Delta s_i(t_2) \rangle = \sigma_{s_i}^2 e^{-|t_2-t_1|/\tau_b}$ , then to leading order, the inhibitory weights have a small variance

$$\text{Var}[w_i] = \alpha \tau_b \frac{\alpha}{\alpha + \beta} \sigma_{s_i}^2.$$

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

$$\begin{aligned} \langle \mathbf{y} \rangle &= \langle \mathbf{s}_b \rangle - \langle \mathbf{w} \rangle = \frac{\beta}{\alpha + \beta} \langle \mathbf{s}_b \rangle \quad (\text{reduction of the average}), \\ \text{but } \text{Var}[y_i] &\sim \sigma_{s_i}^2 \quad \forall i \quad (\text{no reduction of the fluctuations}). \end{aligned} \quad (\text{C2})$$

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 Figs. 2, 5, 8, and others.

#### APPENDIX D: 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  $\mathbf{s}_b(t)$ , assuming perfect separation of timescales between  $\mathbf{s}$ ,  $\Theta$ , and  $\mathbf{m}$ . Then, we obtain exact solutions to these approximate equations.

##### 1. Establishing the IBCM fixed-point equations

Let us 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  $\phi$  is the identity function (instead of a nonlinearity like  $\tanh$ ), such that

$\bar{\mathbf{h}} = L\mathbf{m}\mathbf{s}$ . We also set the decay term  $-\varepsilon\mu\mathbf{m}_i$  to zero. Then, for each neuron  $i$ ,

$$\frac{d\mathbf{m}_i}{dt} = \mu_{\bar{\Theta}_i} \bar{h}_i (\bar{h}_i - \bar{\Theta}_i) \mathbf{s}(t) - \eta \sum_{j \neq i} \mu_{\bar{\Theta}_j} \bar{h}_j (\bar{h}_j - \bar{\Theta}_j) \mathbf{s}(t), \quad (\text{D1})$$

$$\frac{d\bar{\Theta}_i}{dt} = \frac{1}{\tau_{\Theta}} ((\bar{h}_i)^2 - \bar{\Theta}_i). \quad (\text{D2})$$

The reduced activity of interneuron  $i$  is  $\bar{h}_i = \bar{\mathbf{m}}_i \cdot \mathbf{s}(t)$ , where the reduced weights  $\bar{\mathbf{m}}_i = \mathbf{m}_i - \eta \sum_{j \neq i} \mathbf{m}_j$ . Hence,  $\bar{h}_i$  varies in time on two separate scales: rapidly with sensory inputs  $\mathbf{s}(t)$  fluctuating on timescale  $\tau_b$ , and gradually as the synaptic weights  $\bar{\mathbf{m}}_i$  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 ( $\mathbf{M}$ ,  $\bar{\Theta}$ ) remain unchanged. Moreover, we make a quasistatic approximation on the thresholds: we assume  $\tau_{\Theta}$  is slow enough to average over background fluctuations, yet fast enough that  $\bar{m}_i$  remains unchanged over the averaging time window (i.e., we neglect correlations between  $\bar{m}$  and  $\bar{\Theta}$ ). Hence, as in the original IBCM models [45], we assume a perfect separation between the background, threshold, and synaptic weight timescales:  $\tau_b \ll \tau_{\Theta} \ll \frac{1}{\mu}$ .

Thus, averaging Eq. (D2), and setting  $\frac{d\langle \bar{\Theta}_i \rangle}{dt} = 0$ , we find

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

and we replace  $\bar{\Theta}_i$  in the  $\mathbf{m}$  Eq. (D1) with this average. Averaging that equation as well, and setting  $\frac{d\langle \bar{\mathbf{m}}_i \rangle}{dt} = 0$ , we find the IBCM fixed-point equations:

$$\begin{aligned} 0 &= \mu_{\langle \bar{\Theta}_i \rangle} \langle \bar{h}_i (\bar{h}_i - \langle \bar{\Theta}_i \rangle) \mathbf{s}(t) \rangle \\ &\quad - \eta \sum_{j \neq i} \mu_{\langle \bar{\Theta}_j \rangle} \langle \bar{h}_j (\bar{h}_j - \langle \bar{\Theta}_j \rangle) \mathbf{s}(t) \rangle \quad i \in \{1, 2, \dots, N_I\}. \end{aligned} \quad (\text{D3})$$

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

$$0 = L\varphi,$$

where  $L$  is the  $N_I \times N_I$  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  $\eta = \frac{1}{N_I-1}$  (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  $\varphi_i$  term to zero individually:

$$0 = \mu_{\langle \bar{\Theta}_i \rangle} \langle \bar{h}_i (\bar{h}_i - \langle \bar{\Theta}_i \rangle) \mathbf{s}(t) \rangle \quad \forall i \in \{1, 2, \dots, N_I\}. \quad (\text{D4})$$

Hence, in terms of the reduced synaptic weights  $\bar{\mathbf{m}}_i$  and activities  $\bar{h}_i$ , 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  $\bar{\mathbf{m}}_i$

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

$$L_{ij}^{-1} = \begin{cases} \frac{(N_I-2)\eta-1}{(N_I-1)\eta^2+(N_I-2)\eta-1} & \text{on the diagonal,} \\ \frac{-\eta}{(N_I-1)\eta^2+(N_I-2)\eta-1} & \text{off-diagonal,} \end{cases}$$

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

$$\mathbf{m}_i = \frac{[(N_I-2)\eta-1]\overline{\mathbf{m}}_i - \eta \sum_{k \neq i} \overline{\mathbf{m}}_k}{(N_I-1)\eta^2 + (N_I-2)\eta - 1} \quad \text{if } \eta \neq \frac{1}{N_I-1} \text{ or } -1. \quad (\text{D5})$$

## 2. 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,

$$\overline{h}_{i\gamma} = \overline{\mathbf{m}}_i \cdot \hat{\mathbf{s}}_\gamma, \quad (\text{D6})$$

$$\langle h(t)^2 \rangle = \langle (\mathbf{m} \cdot \mathbf{s})^2 \rangle = \sum_{\gamma, \rho} \langle c_\gamma(t) c_\rho(t) \rangle h_\gamma h_\rho = h_d^2 + \sigma^2 u^2,$$

$$\langle h(t) \mathbf{s}(t) \rangle = h_d \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma + \sum_{\gamma, \rho} \langle \tilde{c}_\gamma \tilde{c}_\rho \rangle h_\rho \hat{\mathbf{s}}_\gamma = \sum_\gamma (\langle c \rangle h_d + \sigma^2 h_\gamma) \hat{\mathbf{s}}_\gamma,$$

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

where we have defined

$$h_d = \mathbf{m} \cdot \mathbf{s}_d = \langle c \rangle \sum_\gamma h_\gamma \quad \text{and} \quad u^2 = \sum_{\gamma=1}^{N_B} h_\gamma^2. \quad (\text{D7})$$

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

$$\begin{aligned} 0 &= \sum_{\gamma=1}^{N_B} \left[ \langle c \rangle (h_d^2 + \sigma^2 u^2) (1 - h_d) \right. \\ &\quad \left. - \sigma^2 h_\gamma (h_d^2 + \sigma^2 u^2 - 2h_d) + m_3 h_\gamma^2 \right] \hat{\mathbf{s}}_\gamma. \end{aligned}$$

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,  $h_\gamma$ . These are the fixed-point equations to solve for the  $h_\gamma$ s:

$$\begin{aligned} 0 &= \langle c \rangle (h_d^2 + \sigma^2 u^2) (1 - h_d) - \sigma^2 h_\gamma (h_d^2 + \sigma^2 u^2 - 2h_d) \\ &\quad + m_3 h_\gamma^2 \quad \forall \gamma \in \{1, 2, \dots, N_B\}, \quad \forall \text{ neuron.} \end{aligned} \quad (\text{D8})$$

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. (D1)]. 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. (D4) are identical and solved independently for each neuron.

Since we have decoupled the  $\Theta$  and  $\mathbf{m}$  fluctuations in Eq. (D3), 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 quasistatic approximation. We now assume that the odor concentrations  $c_\gamma(t)$  are independent, identically distributed stationary processes. We write  $c_\gamma(t) = \langle c \rangle + \tilde{c}_\gamma(t)$  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  $\langle \tilde{c}_\gamma^3 \rangle = m_3$ . We also let  $\mathbf{s}(t) = \mathbf{s}_d + \sum_{\gamma=1}^{N_B} c_\gamma(t) \hat{\mathbf{s}}_\gamma$  and  $h(t) = h_d + \sum_\gamma h_\gamma \tilde{c}_\gamma(t)$ , where we have defined  $\mathbf{s}_d = \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$  and  $h_d = \mathbf{m} \cdot \mathbf{s}_d$ . We proceed to compute the averages appearing in the fixed-point equation:

These equations are cubic polynomials in the  $h_\gamma$ s, since  $h_d = \langle c \rangle \sum_\gamma h_\gamma$  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.

## 3. 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,

$$0 = \langle c \rangle (h_d^2 + \sigma^2 u^2) (1 - h_d) - \sigma^2 h_\gamma (h_d^2 + \sigma^2 u^2 - 2h_d),$$

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

$$\sum_\gamma \overline{h}_\gamma = \frac{1}{\langle c \rangle}, \quad (\text{D9})$$

$$\sum_\gamma \overline{h}_\gamma^2 = \frac{1}{\sigma^2}, \quad (\text{D10})$$

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  $(N_B - 2)$ -dimensional surface in the background subspace are nonisolated fixed points when  $N_B > 2$ ; Fig. 11(a) 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  $h_\gamma$ , we find it must have the same value for every odor,  $h_\gamma = h_0$ . Then,  $h_d = N_B h_0$ ,  $u^2 = N_B h_0^2$ , and we can solve to find

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

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 Appendix D4a).

#### 4. Solution for a general background with nonzero third moment

We return to solving the full fixed-point equations, Eq. (D8), when  $m_3 \neq 0$ . Now, because of the  $m_3 h_\gamma$  term, setting  $h_d = 1$  and  $u^2 = 1/\sigma^2$  does not satisfy the equations, so we must solve for individual  $h_\gamma$ s. This leads to a finite number of isolated fixed points: as shown in Fig. 11(b), the third moment of the background breaks the degeneracy seen in the Gaussian case.

First, we notice that the dot products  $\bar{h}_\gamma$  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  $h_\gamma$  and some other  $h_\alpha$ , which gives

$$0 = m_3(h_\gamma + h_\alpha)(h_\gamma - h_\alpha) - (h_d^2 \sigma^2 + \sigma^4 u^2 - 2h_d \sigma^2)(h_\gamma - h_\alpha).$$

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

$$h_\gamma + h_\alpha = \frac{h_d^2 \sigma^2 + \sigma^4 u^2 - 2h_d \sigma^2}{m_3}. \quad (\text{D11})$$

Note that constraint (D11) is removed for Gaussian backgrounds  $m_3 = 0$ , explaining the nonisolated fixed points in that special case. For any pair  $\gamma, \alpha$ , the r.h.s. is the same; if we imagine a third dot product  $h_\beta$ , 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  $h_\gamma$ s are equal, or they each take one of two possible values.

##### a. All $h_\gamma$ s are equal

First, consider the case where all  $h_\gamma = y$ , a unique dot product value. Then  $u^2 = N_B y^2$  and  $h_d = N_B \langle c \rangle y$ . Inserting in Eq. (D8), we can factor out  $y^2$ , giving

$$0 = y^2 [N_B^2 \langle c \rangle^3 + 3\sigma^2 N_B \langle c \rangle + m_3 - y(N_B^3 \langle c \rangle^4 + 2\sigma^2 N_B^2 \langle c \rangle^2 + \sigma^4 N_B)].$$

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

$$y = \frac{N_B^2 \langle c \rangle^3 + 3\sigma^2 N_B \langle c \rangle + m_3}{N_B^3 \langle c \rangle^4 + 2\sigma^2 N_B^2 \langle c \rangle^2 + \sigma^4 N_B}. \quad (\text{D12})$$

We conjecture that this fixed point is always unstable, based on the linear stability analysis of Appendix D6 below. Figure 9 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.

##### b. Two different $h_\gamma$ 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  $k_2 = N_B - k_1$  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,

$$h_d = \langle c \rangle (k_1 y_1 + k_2 y_2) \quad (\text{D13})$$

$$u^2 = k_1 y_1^2 + k_2 y_2^2. \quad (\text{D14})$$

Then, the set of  $N_B$  equations (D8) 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. (D11)—the difference between the two values of  $h_\gamma$ . We rewrite the equation in terms of  $y_1$  and  $y_2$  as

$$0 = h_d^2 - 2h_d + \sigma^2 u^2 - \frac{m_3}{\sigma^2} (y_1 + y_2) \quad (\text{D15})$$

by letting one of the  $h_\gamma$ s 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 (D8),

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

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

$$0 = \sigma^2 u^2 - h_d^2 - \frac{m_3}{\sigma^2} h_d (y_1 + y_2) - \frac{m_3}{\langle c \rangle} y_1 y_2. \quad (\text{D16})$$

Together, Eqs. (D15) and (D16) 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  $y_i$  (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). This allows us to easily isolate  $y_2$  in terms of  $y_1$  (or viceversa). Indeed, writing  $h_d$  and  $u^2$  in terms of the  $y_i$  [Eqs. (D13) and (D14)], we find it takes the simple, symmetric form

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

where

$$a_i = \sigma^2 k_i - \langle c \rangle^2 k_i^2 - \frac{m_3 \langle c \rangle}{\sigma^2} k_i \quad (i \in \{1, 2\}),$$

$$b = 2 \langle c \rangle^2 k_1 k_2 + \frac{m_3 \langle c \rangle}{\sigma^2} (k_1 + k_2) + \frac{m_3}{\langle c \rangle}. \quad (\text{D17})$$

This equation makes clear the symmetry of solutions under exchange of labels  $1 \leftrightarrow 2$ , confirming that we can keep only the roots where  $y_1 > y_2$ , knowing that other roots would be

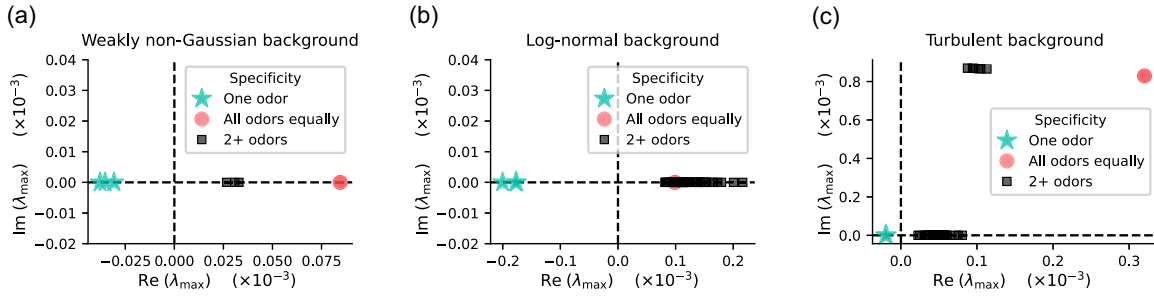


FIG. 9. Selective states are the only stable fixed points in the IBCM model. Eigenvalue with the largest negative real part in the Jacobian matrix obtained in the linear stability analysis of the IBCM model (Appendix D6), evaluated for each possible fixed point in (a) the weakly non-Gaussian background of Fig. 3, (b) the log-normal background of Fig. 12, and (c) the turbulent background of Figs. 1(b), 1(c), 2, and 5. The Jacobian matrix derived in Eq. (D21) is diagonalized numerically for each possible choice of specific and nonspecific 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.

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  $y_1$ ,

$$y_2 = \left( \frac{b \pm \sqrt{b^2 - 4a_1a_2}}{2a_2} \right) y_1 = \alpha_{\pm} y_1. \quad (\text{D18})$$

Formally, the numerator should be  $by_1 \pm \sqrt{b^2 - 4a_1a_2}|y_1|$ , but we can absorb the absolute value into  $\pm$ , compute the solution for both  $\alpha$  values, and keep the one with  $y_1 > y_2$  at the end. We now insert  $y_2 = \alpha y_1$  into Eq. (D15); another root  $y_1 = y_2 = 0$  can be factored out, and we find the nontrivial solution

$$y_1 = \frac{2 \langle c \rangle (k_1 + \alpha k_2) + \frac{m_3}{\sigma^2} (1 + \alpha)}{\langle c \rangle^2 (k_1 + \alpha k_2)^2 + \sigma^2 (k_1 + \alpha^2 k_2)}. \quad (\text{D19})$$

Equations (D17), (D18), and (D19) form our analytical solution for the (approximate) fixed points of IBCM neurons in terms of the dot products  $h_\gamma$  taking values  $y_1$  and  $y_2$ .

### c. Condition of existence of nontrivial fixed points

The fixed points with  $y_1$  and  $y_2$  given by Eqs. (D18) and (D19) will only exist in  $\mathbb{R}$  if the discriminant  $b^2 - 4a_1a_2$  in  $\alpha$  is nonnegative. Writing this discriminant explicitly,

$$\begin{aligned} b^2 - 4a_1a_2 &= 4\sigma^2 \langle c \rangle^2 k_1 k_2 \left( N_B - \frac{\sigma^2}{\langle c \rangle^2} \right) \\ &\quad + 12m_3 \langle c \rangle k_1 k_2 + m_3^2 \left( \frac{\langle c \rangle^2}{\sigma^4} (k_1 - k_2)^2 + \frac{1}{\langle c \rangle^2} + \frac{2N_B}{\sigma^2} \right). \end{aligned} \quad (\text{D20})$$

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

## 5. Summary of the general fixed-point solutions

The fixed-point solution for i.i.d. odor concentrations with mean  $\langle c \rangle$ , variance  $\sigma^2$ , third moment  $m_3$ , is summarized here. The fixed points of  $\bar{\mathbf{m}}$  are characterized by their dot products with the  $N_B$  background odors,  $\bar{h}_\gamma = \bar{\mathbf{m}} \cdot \hat{\mathbf{s}}_\gamma$ . There are  $2^{N_B}$  fixed points in total: one with all  $h_\gamma = 0$ , one with all  $h_\gamma$ s equal to (Appendix D4a)

$$\bar{h}_\gamma = \frac{N_B^2 \langle c \rangle^3 + 3\sigma^2 N_B \langle c \rangle + m_3}{N_B^3 \langle c \rangle^4 + 2\sigma^2 N_B^2 \langle c \rangle^2 + \sigma^4 N_B}, \quad (\text{D12})$$

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

$$\begin{aligned} a_i &= \sigma^2 k_i - \langle c \rangle^2 k_i^2 - \frac{m_3 \langle c \rangle}{\sigma^2} k_i \quad (i \in \{1, 2\}), \\ b &= 2 \langle c \rangle^2 k_1 k_2 + \frac{m_3 \langle c \rangle}{\sigma^2} (k_1 + k_2) + \frac{m_3}{\langle c \rangle}, \end{aligned} \quad (\text{D17})$$

$$\alpha = \frac{b \pm \sqrt{b^2 - 4a_1a_2}}{2a_2}. \quad (\text{D18})$$

Compute solutions for each sign in  $\alpha$

$$\begin{aligned} y_1 &= \frac{2 \langle c \rangle (k_1 + \alpha k_2) + \frac{m_3}{\sigma^2} (1 + \alpha)}{\langle c \rangle^2 (k_1 + \alpha k_2)^2 + \sigma^2 (k_1 + \alpha^2 k_2)}, \\ y_2 &= \alpha y_1, \quad \text{keep the pair where } y_1 > y_2. \end{aligned} \quad (\text{D19})$$

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  $\mathbf{m} \cdot \hat{\mathbf{s}}_\gamma = y_1 \equiv h_{\text{sp}}$  (specific) for that odor, and  $y_2 \equiv h_{\text{ns}}$  (nonspecific) for all other odors. We observe close quantitative agreement between these fixed point predictions and numerical simulations in the weakly non-Gaussian background, Fig. 3(c). For log-normal [Figs. 12(d) and 12(e)] and turbulent backgrounds [Figs. 4(a) and 4(b)], we also observe that IBCM neurons



align with individual background odors, but due to stronger correlations between  $\mathbf{m}$  and  $\Theta$ , the dot product values do not exactly match Eqs. (D17)–(D19).

### 6. Linear stability analysis of IBCM fixed points

To support our empirical results on IBCM neuron selectivity, we linearize the dynamical equations (D1) and (D2) 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

$$\begin{aligned} \left\langle \frac{d\mathbf{m}}{dt} \right\rangle &= \mu \langle h^2 \mathbf{s}_b \rangle - \mu \langle \Theta \rangle \langle h \mathbf{s}_b \rangle, \\ \left\langle \frac{d\Theta}{dt} \right\rangle &= \frac{1}{\tau_\Theta} (\langle h^2 \rangle - \langle \Theta \rangle). \end{aligned}$$

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

$$\begin{aligned} \frac{\partial}{\partial m_i} \left\langle \frac{dm_j}{dt} \right\rangle &= 2\mu \langle h s_i s_j \rangle - \mu \langle \Theta \rangle \langle s_i s_j \rangle, \\ \frac{\partial}{\partial \Theta} \left\langle \frac{dm_j}{dt} \right\rangle &= -\mu \langle h s_j \rangle, \\ \frac{\partial}{\partial m_i} \left\langle \frac{d\Theta}{dt} \right\rangle &= \frac{2}{\tau_\Theta} \langle h s_i \rangle, \\ \frac{\partial}{\partial \Theta} \left\langle \frac{d\Theta}{dt} \right\rangle &= -\frac{1}{\tau_\Theta}. \end{aligned}$$

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

$$D\mathbf{f}(\mathbf{m}, \Theta) = \left( \begin{array}{c|c} 2\mu \langle h \mathbf{s}_b \mathbf{s}_b^T \rangle & -\mu \langle h \mathbf{s}_b \rangle \\ \hline -\mu \langle \Theta \rangle \langle \mathbf{s}_b \mathbf{s}_b^T \rangle & -\frac{1}{\tau_\Theta} \end{array} \right). \quad (\text{D21})$$

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

$$\begin{aligned} \langle h \mathbf{s}_b \rangle &= h_d \mathbf{s}_d + \sigma^2 \sum_\gamma h_\gamma \hat{\mathbf{s}}_\gamma, \\ \langle \mathbf{s}_b \mathbf{s}_b^T \rangle &= \mathbf{s}_d \mathbf{s}_d^T + \sigma^2 \sum_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^T, \\ \langle h \mathbf{s}_b \mathbf{s}_b^T \rangle &= h_d \langle \mathbf{s}_b \mathbf{s}_b^T \rangle + \sigma^2 \sum_\gamma h_\gamma (\mathbf{s}_d \hat{\mathbf{s}}_\gamma^T + \hat{\mathbf{s}}_\gamma \mathbf{s}_d^T) \\ &\quad + 2m_3 \sum_\gamma h_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^T. \end{aligned}$$

The last two lines, along with  $\Theta = h_d^2 + \sigma^2 u^2$ , mean that the main block of the matrix is

$$2\mu \langle h \mathbf{s}_b \mathbf{s}_b^T \rangle - \mu \langle \Theta \rangle \langle \mathbf{s}_b \mathbf{s}_b^T \rangle$$

$$\begin{aligned} &= \mu \left[ (2h_d - h_d^2 - \sigma^2 u^2) (\mathbf{s}_d \mathbf{s}_d^T + \sigma^2 \sum_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^T) \right. \\ &\quad \left. + 2m_3 \sum_\gamma h_\gamma \hat{\mathbf{s}}_\gamma \hat{\mathbf{s}}_\gamma^T + 2\sigma^2 \sum_\gamma h_\gamma (\mathbf{s}_d \hat{\mathbf{s}}_\gamma^T + \hat{\mathbf{s}}_\gamma \mathbf{s}_d^T) \right]. \end{aligned}$$

We have defined here  $\mathbf{s}_d = \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$ . 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. 9. In all these examples, we find that the only stable fixed points are those where the neuron has one dot product  $h_\gamma = h_{sp}$  (specific) and the  $N_B - 1$  others are equal to  $h_{ns}$  (nonspecific). This property is robust against OSN noise, as shown in Fig. 16. 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. 16(h)].

### 7. 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,  $N_I \geq N_B$ . We call  $\gamma_j$  the background odor to which IBCM neuron  $j$  is specific; thus,  $\mathbf{m}_j \cdot \hat{\mathbf{s}}_\gamma = h_{sp}$  if  $\gamma = \gamma_j$ ,  $h_{ns}$  otherwise. There will be in general some number  $n_\gamma$  of neurons specific to odor  $\gamma$ .

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

$$\frac{dW}{dt} = \alpha \mathbf{y}(t) \bar{\mathbf{h}}(t)^T - \beta W. \quad (\text{D22})$$

We define the (constant) matrix  $\Gamma$ , whose columns are the  $\hat{\mathbf{s}}_\gamma$ , and  $\mathbf{c}$  the vector of odor concentrations. Then,  $\mathbf{s}_b(t) = \Gamma \mathbf{c}$ . We also define the matrix  $H = L M \Gamma$ , in which row  $j$  gives the alignment of IBCM neuron  $j$  with each odor,  $h_{j\gamma}$ . Since each neuron is selective for one odor, each row contains  $h_{sp}$  once and  $h_{ns}$   $N_B - 1$  times. Averaging the  $W$  equation over fast  $c$  fluctuations, and neglecting correlations between  $c$ ,  $W$ , and  $H$ , we have

$$\left\langle \frac{dW}{dt} \right\rangle = \alpha (\Gamma - \langle W \rangle \langle H \rangle) \langle \mathbf{c} \mathbf{c}^T \rangle \langle H \rangle^T - \beta \langle W \rangle.$$

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

$$N = \langle \mathbf{c} \mathbf{c}^T \rangle = \langle c \rangle^2 O_{N_B} + \sigma^2 \mathbb{I}_{N_B}, \quad (\text{D23})$$

where  $\mathbb{I}$  is the  $N_B \times N_B$  identity matrix and  $O_{N_B}$  is a  $N_B \times N_B$  matrix filled with ones. We now set  $dW/dt = 0$  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  $H$  ( $\mathbf{h}_j^T$  is the row  $j$  of

H). The set of equations to solve for the  $\mathbf{w}_j$  is then

$$\frac{\beta}{\alpha} \mathbf{w}_j + WHN\mathbf{h}_j = \Gamma N\mathbf{h}_j. \quad (\text{D24})$$

We notice here that all IBCM neurons with the same specificity  $\gamma_j$  will have the same  $\mathbf{h}_j$ , thus all columns  $\mathbf{w}_j$  with the same  $\gamma_j$  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

$$\mathbf{h}_k^T \mathbf{h}_j = \begin{cases} h_{\text{sp}}^2 + (N_B - 1)h_{\text{ns}}^2 = u^2 & \text{if } \gamma_j = \gamma_k, \\ 2h_{\text{sp}}h_{\text{ns}} + (N_B - 2)h_{\text{ns}}^2 = h_{\text{ns}}(\bar{h}_d / \langle c \rangle + h_{\text{sp}} - h_{\text{ns}}) & \text{else,} \end{cases} \quad (\text{D25})$$

and that

$$\begin{aligned} \mathbf{h}_k^T O_{N_B} \mathbf{h}_j &= \mathbf{h}_k^T \mathbf{v} (h_{\text{sp}} + (N_B - 1)h_{\text{ns}}) \\ &= (h_{\text{sp}} + (N_B - 1)h_{\text{ns}})^2 \\ &= \frac{\bar{h}_d^2}{\langle c \rangle^2}, \end{aligned} \quad (\text{D26})$$

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

$$N\mathbf{h}_j = \sigma^2 \mathbf{h}_j + \langle c \rangle^2 (h_{\text{sp}} + (N_B - 1)h_{\text{ns}}) \mathbf{v} = \sigma^2 \mathbf{h}_j + \langle c \rangle \bar{h}_d \mathbf{v},$$

where we used Eq. (D23). We use the above results to evaluate the two terms involving  $N$  in Eq. (D24). With some algebra to combine coefficients efficiently, we find

$$WHN\mathbf{h}_j = \sigma^2 (h_{\text{sp}} - h_{\text{ns}})^2 n_{\gamma_j} \mathbf{w}_{\gamma_j} + A \sum_{\gamma} n_{\gamma} \mathbf{w}_{\gamma}, \quad (\text{D27})$$

$$\Gamma N\mathbf{h}_j = \sigma^2 (h_{\text{sp}} - h_{\text{ns}}) \hat{\mathbf{s}}_{\gamma_j} + \left( \frac{\sigma^2}{\langle c \rangle} h_{\text{ns}} + \bar{h}_d \right) \mathbf{s}_d, \quad (\text{D28})$$

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

$$A = \bar{h}_d^2 + \frac{\sigma^2}{\langle c \rangle} \bar{h}_d h_{\text{ns}} + \sigma^2 h_{\text{ns}} (h_{\text{sp}} - h_{\text{ns}}). \quad (\text{D29})$$

We now insert Eqs. (D27) and (D28) into Eq. (D24) for  $\mathbf{w}_j$  (or equivalently,  $\mathbf{w}_{\gamma_j}$ ), to find

$$B_{\gamma_j} \mathbf{w}_j + A \sum_{\gamma} n_{\gamma} \mathbf{w}_{\gamma} = \sigma^2 (h_{\text{sp}} - h_{\text{ns}}) \hat{\mathbf{s}}_{\gamma_j} + \left( \frac{\sigma^2}{\langle c \rangle} h_{\text{ns}} + \bar{h}_d \right) \mathbf{s}_d, \quad (\text{D30})$$

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

$$B_{\gamma_j} = \frac{\beta}{\alpha} + n_{\gamma_j} \sigma^2 (h_{\text{sp}} - h_{\text{ns}})^2. \quad (\text{D31})$$

Now, to solve Eq. (D30) 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:

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

Doing this for each  $\gamma_k \neq \gamma_j$ , and noticing the expression reduces to  $\mathbf{w}_{\gamma_j}$  for  $\gamma_k = \gamma_j$ , we express  $\sum_{\gamma} n_{\gamma} \mathbf{w}_{\gamma}$  in terms of

$\mathbf{w}_{\gamma_j}$  only, and insert into Eq. (D30) to isolate  $\mathbf{w}_{\gamma_j}$ . Dividing the result by  $B_{\gamma_j} (1 + A \sum_{\gamma} \frac{n_{\gamma}}{B_{\gamma}})$ , and rearranging with further algebra gives our final expression for columns of the matrix  $W$ ,

$$\begin{aligned} \mathbf{w}_{\gamma_j} &= \frac{\sigma^2 (h_{\text{sp}} - h_{\text{ns}})}{B_{\gamma_j}} \hat{\mathbf{s}}_{\gamma_j} + \frac{\bar{h}_d + \frac{\sigma^2}{\langle c \rangle} h_{\text{ns}}}{B_{\gamma_j} (1 + AK)} \mathbf{s}_d \\ &\quad - \frac{A \sigma^2 (h_{\text{sp}} - h_{\text{ns}})}{B_{\gamma_j} (1 + AK)} \sum_{\gamma} \frac{n_{\gamma}}{B_{\gamma}} \hat{\mathbf{s}}_{\gamma}, \end{aligned} \quad (\text{D32})$$

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 Fig. 3(d) (weakly non-Gaussian) and Fig. 12(f) (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,  $n_{\gamma_j}$ , are inferred from the  $M$  weights numerical results, then used to evaluate Eq. (D32).

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  $B_{\gamma}$ . However, we find a more compact expression for the average PN response: after simplifying some terms,

$$\begin{aligned} \langle \mathbf{y} \rangle &= \langle c \rangle \sum_{\gamma} \frac{\hat{\mathbf{s}}_{\gamma}}{B_{\gamma}} \frac{1}{1 + AK} \\ &\quad \times \left( \frac{\beta}{\alpha} - \sigma^2 h_{\text{ns}} (h_{\text{sp}} - h_{\text{ns}}) (n_{\gamma} N_B - K B_{\gamma}) \right). \end{aligned} \quad (\text{D33})$$

The factor  $n_{\gamma} N_B - K B_{\gamma}$  is zero when all  $n_{\gamma}$ s 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  $n_{\gamma} N_B - K B_{\gamma} = 0$  for some  $n_{\gamma}$  is  $n^* = \frac{\beta / \alpha K}{N_B - K \sigma^2 (h_{\text{sp}} - h_{\text{ns}})^2}$ ; 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  $n_{\gamma} = n_I / N_B$ , the threshold  $n^* < n_I / N_B$ ; hence, all components which have  $n_{\gamma} > n_I / N_B$  surely have a positive factor  $(n_{\gamma} N_B - K B_{\gamma})$ , and since  $h_{\text{ns}} < 0$  in general, they have a negative bias in Eq. (D33); 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 nonspecific response of other neurons,  $h_{\text{ns}} < 0$ , but at the cost of less

efficient inhibition of all other odors, and without overall reduction of fluctuations since the factor  $\frac{\beta/\alpha}{B} = 1$  if  $n_\gamma = 0$ .

When all  $n_\gamma$ s are equal, the mean PN response to the background, Eq. (D33), is minimized for the IBCM network, and it reduces to a simple expression,

$$\langle \mathbf{y} \rangle = \frac{\beta/\alpha}{B + N_B A} \mathbf{s}_d, \quad (\text{D34})$$

where, recall,  $A$  is given by Eq. (D29),  $B$  by Eq. (D31) with all  $n_\gamma$  equal, and  $\mathbf{s}_d = \langle c \rangle \sum_\gamma \hat{\mathbf{s}}_\gamma$  is the average background. We use this simplified expression of the maximal habituation by IBCM neurons to determine the  $\Lambda$  factor needed to match the BioPCA and IBCM performances (Appendix J 2).

#### APPENDIX E: ANALYTICAL FIXED-POINT SOLUTIONS OF THE BIOPCA MODEL

We assume that the average background  $\langle \mathbf{s} \rangle = \mathbf{s}_d = \langle c \rangle \sum_{\gamma=1}^{N_B} \hat{\mathbf{s}}_\gamma$  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  $\tilde{\mathbf{s}} = \sum_{\gamma=1}^{N_B} \tilde{c}_\gamma \hat{\mathbf{s}}_\gamma$ , with  $\langle \tilde{c}_\gamma \rangle = 0$  and  $\langle \tilde{c}_\gamma^2 \rangle = \sigma^2$  for all components  $\gamma$ . The LN activity is  $\bar{\mathbf{h}}(t) = LM\tilde{\mathbf{s}}(t)$  and the PN response is  $\mathbf{y}(t) = \tilde{\mathbf{s}}(t) - W\bar{\mathbf{h}}(t)$ . The covariance matrix, from which a PCA with  $N_B$  components can be computed, is

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

where  $U$  is  $N_s \times N_B$ , with its columns containing the  $N_B$  principal component vectors with nonzero eigenvalue, and  $D$  is  $N_B \times N_B$  and diagonal with the principal values in it,  $\sigma_i^2$ ,  $i \in \{1, 2, \dots, N_B\}$ . Since the  $\hat{\mathbf{s}}_\gamma$  are not orthogonal in general, these eigenvalues are not all equal to the variance  $\sigma^2$ , 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.* [38], we expect the BioPCA model with  $N_I = N_B$  neurons (one per background component) to have the following stationary solution:

$$L' = D \quad (\text{principal values}), \quad (\text{E1})$$

$$LM = \underline{\Lambda} U^\top \quad (\text{projection on principal components}), \quad (\text{E2})$$

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

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

We insert the BioPCA stationary solution into the  $W$  equation (D22), where we average over fast  $\tilde{\mathbf{s}}$  fluctuations, assuming a separation of timescales between these fluctuations and slow  $M, L, W$  learning, as in the IBCM case. Writing the  $W$  equation in matrix form, this leads to

$$\begin{aligned} \frac{dW}{dt} &= \alpha \langle \mathbf{y}(t) \bar{\mathbf{h}}^\top(t) \rangle - \beta W \\ &= \alpha (\mathbb{I} - W \underline{\Lambda} U^\top) \langle \tilde{\mathbf{s}} \tilde{\mathbf{s}}^\top \rangle U \underline{\Lambda}^\top - \beta W. \end{aligned}$$

Setting the  $W$  derivative to zero, replacing  $\langle \tilde{\mathbf{s}} \tilde{\mathbf{s}}^\top \rangle = UDU^\top$ , we can rearrange to isolate  $W$ ,

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

Since  $D$  and  $\underline{\Lambda}$  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  $N_I \leq N_B$ , so we can invert the equation explicitly. Since  $D \underline{\Lambda}^\top$  is also diagonal, we find

$$W = U \text{diag} \left( \frac{\Lambda_{ii} \sigma_i^2}{\beta/\alpha + \Lambda_{ii}^2 \sigma_i^2} \right). \quad (\text{E3})$$

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

$$\mathbf{y} = \tilde{\mathbf{s}} - W \underline{\Lambda} U^\top \tilde{\mathbf{s}} = U \text{diag} \left( \frac{\beta/\alpha}{\beta/\alpha + \Lambda_{ii}^2 \sigma_i^2} \right) U^\top \tilde{\mathbf{s}}, \quad (\text{E4})$$

where we have used the fact that  $UU^\top$  is a projector on the background subspace to write  $\tilde{\mathbf{s}} = UU^\top \tilde{\mathbf{s}}$ , and where  $\Lambda_{ii} = \Lambda_{\text{PCA}}(1 - \lambda_r(i+1)/N_B)$  for  $i = 0, 1, \dots, N_I$ . Hence, we see that the BioPCA network projects the inputs on the principal directions ( $U^\top \tilde{\mathbf{s}}$ ), 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 Eq. (D34) for the IBCM model, the latter has a better reduction by a factor of approximately  $(h_{\text{sp}} - h_{\text{ns}})^2$  in the denominator, hence we need to increase the  $\Lambda$  scale in the BioPCA network to match the performance of these models, as explained in Appendix J 2. We check some of these predictions against numerical simulations in a background with log-normal [Figs. 12(b) and 12(c)] and turbulent [Figs. 4(c) and 4(d)] concentration statistics. We also check that these properties are relatively robust against OSN noise (Fig. 16); 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).

#### APPENDIX F: 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 nontrivial background, illustrated in Figs. 10(a) and 10(b). It consists of two odors ( $\mathbf{s}_a, \mathbf{s}_b$ ) with fluctuating proportion  $\bar{v}(t)$  following a Ornstein-Uhlenbeck process (Appendix B 2),

$$\mathbf{s}(t) = \left( \frac{1}{2} + \bar{v}(t) \right) \mathbf{s}_a + \left( \frac{1}{2} - \bar{v}(t) \right) \mathbf{s}_b. \quad (\text{F1})$$

We start by calculating the average fixed points of the IBCM neurons synaptic weights,  $\bar{\mathbf{m}}_i$ . The fixed-point equations (D4) are identical for all neurons, so we focus on one neuron and omit index  $i$ . As in Appendix D, we assume that timescales are well separated, replace  $\Theta = \langle h^2 \rangle$ , and neglect correlations between  $\bar{v}$ ,  $\Theta$ , and  $\mathbf{m}$ . We work with reduced weights and activities  $\bar{h}_i$ , 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 Eq. (D5). Moreover, for this simple background, the learning rate can be chosen constant,  $\mu_{\langle \bar{\Theta} \rangle} = \mu$ . Hence, the fixed-point equation to solve here is

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

Now, we rewrite

$$\mathbf{s}(t) = \mathbf{s}_d + \bar{v}(t) \mathbf{s}_s,$$

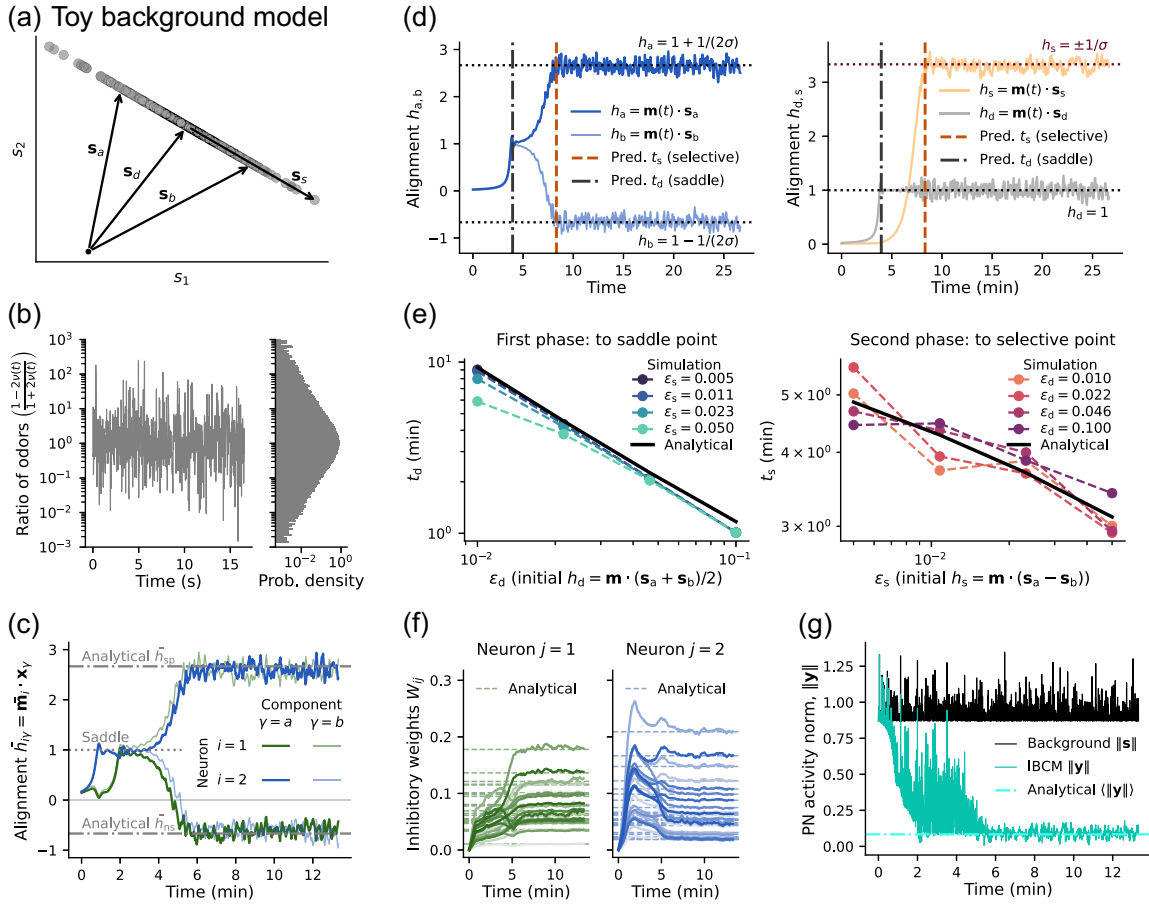


FIG. 10. IBCM learning dynamics on a simplified background model. (a) Schematic of the simplified background model described in Appendix F: a 1D manifold generated by two odors  $\mathbf{s}_a, \mathbf{s}_b$  with fluctuating proportion  $\bar{v}(t)$  following an Ornstein-Uhlenbeck process with  $\langle \bar{v} \rangle = 0$ ,  $\sigma^2 = 0.09$ , autocorrelation time  $\tau_b = 20$  ms. The deterministic part is  $\mathbf{s}_d = \frac{1}{2}(\mathbf{s}_a + \mathbf{s}_b)$  and the stochastic part,  $\mathbf{s}_s = \mathbf{s}_a - \mathbf{s}_b$ . 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  $\mathbf{s}_a$ , the other, to  $\mathbf{s}_b$ . The average dot products at steady-state match the analytical fixed points calculated in Eq. (F2),  $\bar{h}_{\pm} = 1 \pm 1/(2\sigma)$ . (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  $t_d$  and  $t_s$  closely matching their analytical values [Eqs. (F8) and (F9)]. Right: visualized in terms of the alignments  $h_d$  and  $h_s$  with the deterministic and stochastic components,  $\mathbf{s}_d$  and  $\mathbf{s}_s$ , 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 reaches 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,  $\epsilon_d$  and  $\epsilon_s$ . The numerical simulations match the analytical expressions, valid for small  $\epsilon_s$ . (f) Time series of the inhibitory weights (elements of  $W$ ), following the Hebbian dynamics of Eq. (F3), for the two neurons in the network (weights of LN  $j$  are in column  $j$  of  $W$ ). The steady-state values match closely the Hebbian dynamics derived in Eq. (F4). (g) Reduction in PN activity after habituation, compared to the background norm. The steady-state reduction matches the analytical prediction [Eq. (F5)] and is reached once the two neurons select one odor each.

where we have defined  $\mathbf{s}_d = \frac{1}{2}(\mathbf{s}_a + \mathbf{s}_b)$  (deterministic part) and  $\mathbf{s}_s = \mathbf{s}_a - \mathbf{s}_b$  (stochastic part). We examine the dot products of synaptic weights with these components,  $h_d = \mathbf{m} \cdot \mathbf{s}_d$  and  $h_s = \mathbf{m} \cdot \mathbf{s}_s$ , such that  $h(t) = h_d + \bar{v}(t)h_s$ . We can solve for the two dot products  $h_d$  and  $h_s$  because they specify the fixed points completely for  $N_B = 2$  background components. In term of these quantities, the fixed-point equation becomes

$$\begin{aligned}
 0 &= \langle (h_d + \bar{v}h_s)^2 (\mathbf{s}_d + \bar{v}\mathbf{s}_s) \rangle \\
 &\quad - \langle (h_d + \bar{v}h_s)^2 \rangle \langle (h_d + \bar{v}h_s)(\mathbf{s}_d + \bar{v}\mathbf{s}_s) \rangle, \\
 0 &= (h_d^2 + \sigma^2 h_s^2 - h_d^3 - h_d h_s^2 \sigma^2) \mathbf{s}_d \\
 &\quad + (2h_s h_d - h_d^2 h_s - \sigma^2 h_s^3) \sigma^2 \mathbf{s}_s.
 \end{aligned}$$

Since  $\mathbf{s}_d$  and  $\mathbf{s}_s$  are linearly independent, both coefficients must be zero, leading to a system of two equations for  $h_d$  and  $h_s$ ,

$$\begin{aligned}
 0 &= h_d^3 - h_d^2 - \sigma^2 h_s^2 + h_d h_s^2 \sigma^2, \\
 0 &= \sigma^2 h_s (h_d^2 + \sigma^2 h_s^2 - 2h_d).
 \end{aligned}$$

There is a trivial solution  $h_s = h_d = 0$ , which is unstable. The other solutions are, by inspection,

$$h_d = \mathbf{m} \cdot \mathbf{s}_d = 1 \text{ and } h_s = \mathbf{m} \cdot \mathbf{s}_s = \pm \frac{1}{\sigma},$$



## (a) Gaussian background

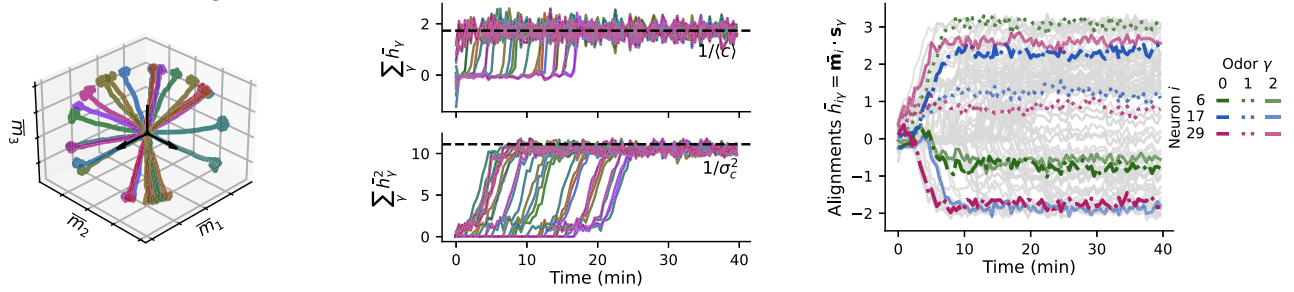
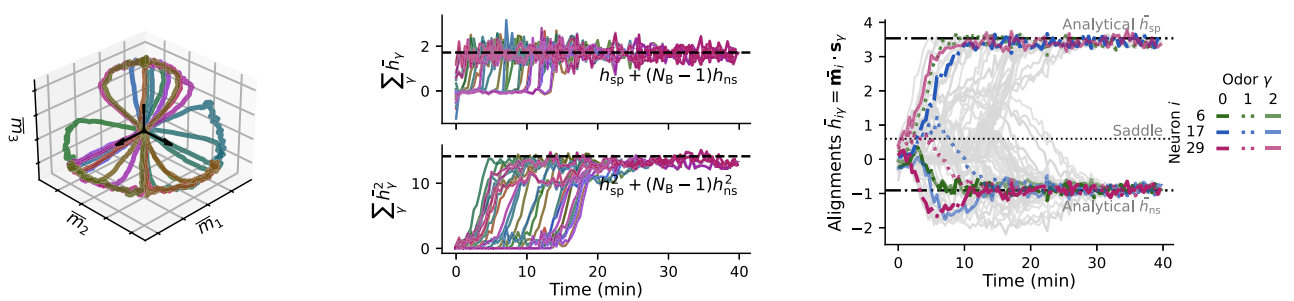
(b) Weakly non-Gaussian background ( $\nu = 0.2$ )

FIG. 11. IBCM learning depends on higher-order moments of the background distribution. (a) Three-odor background with Gaussian concentration fluctuations: each odor has  $c_\gamma = g_\gamma$  where  $g_\gamma$  follows a Ornstein-Uhlenbeck process (with  $\langle g \rangle = 1/\sqrt{N_B}$ ,  $\sigma_c^2 = 0.09$ , correlation time  $\tau = 20$  ms). Left: when the third moment is zero, synaptic weights  $\bar{\mathbf{m}}_i = \mathbf{m}_i - \eta \sum_{j \neq i} \mathbf{m}_j$  have nonisolated fixed points on the codimension-two ring defined by constraints (D9) (hyperplane) and (D10) (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 nonspecific odors for each neuron. (b) Same as panel (a), but with small nonzero third moment added to the background concentrations statistics, by taking  $c = g + \nu g^2$  for  $\nu = 0.2$ , 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  $\bar{h}_y$ s and  $\bar{h}_y^2$ 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  $N_B - 1$  odors); three neurons highlighted. Odor vectors were chosen to be symmetric around the origin in the first three dimensions, to clarify the geometric picture.

or, in terms of the dot products with  $\mathbf{s}_a$  and  $\mathbf{s}_b$ ,

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

Hence, we have two different stable fixed points, which we call  $\mathbf{m}_+$  and  $\mathbf{m}_-$  to indicate which sign the dot product with  $\mathbf{s}_s$  takes. Figure 10(c) 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  $\mathbf{s}(t)$ :

$$h_\pm(t) = \mathbf{m}_\pm \cdot (\mathbf{s}_s + \bar{v}(t)\mathbf{s}_d) = 1 \pm \frac{\bar{v}(t)}{\sigma}. \quad (\text{F2})$$

We notice that  $h_\pm = 0$  when  $\bar{v} = \mp\sigma$ , that is, the IBCM neuron is nonresponsive 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.

## 1. 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  $N_I = 2$  neurons, one at each fixed point  $\pm$ .

Averaging the  $W$  equation (D22) over background fluctuations, writing out  $\mathbf{y} = \mathbf{s} - WLM\mathbf{s}$ , and focusing on the column for one neuron  $j$ , we have

$$\frac{d\langle \mathbf{w}_j \rangle}{dt} = \alpha \langle \bar{h}_j(\mathbf{s} - WLM\mathbf{s}) \rangle - \beta \langle \mathbf{w}_j \rangle \quad \forall j. \quad (\text{F3})$$

To solve for  $\mathbf{w}_j$ , we set the derivative to zero, and we assume there are  $N_I = 2$  IBCM neurons, one at each fixed point  $\pm$ . We still assume a separation of timescales, and assume the  $\bar{\mathbf{m}}$  are equal to their average fixed-point values, so at any time  $t$ , the IBCM neuron activity  $\bar{h}(t)$  is given by (F2). 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

$$\begin{aligned} \frac{d\mathbf{w}_\pm}{dt} = 0 = \alpha \left\langle \left( 1 \pm \frac{\bar{v}(t)}{\sigma} \right) \left[ \mathbf{s}_d + \bar{v}(t)\mathbf{s}_s - \mathbf{w}_+ \left( 1 + \frac{\bar{v}(t)}{\sigma} \right) \right. \right. \\ \left. \left. - \mathbf{w}_- \left( 1 - \frac{\bar{v}(t)}{\sigma} \right) \right] \right\rangle - \beta \mathbf{w}_\pm. \end{aligned}$$

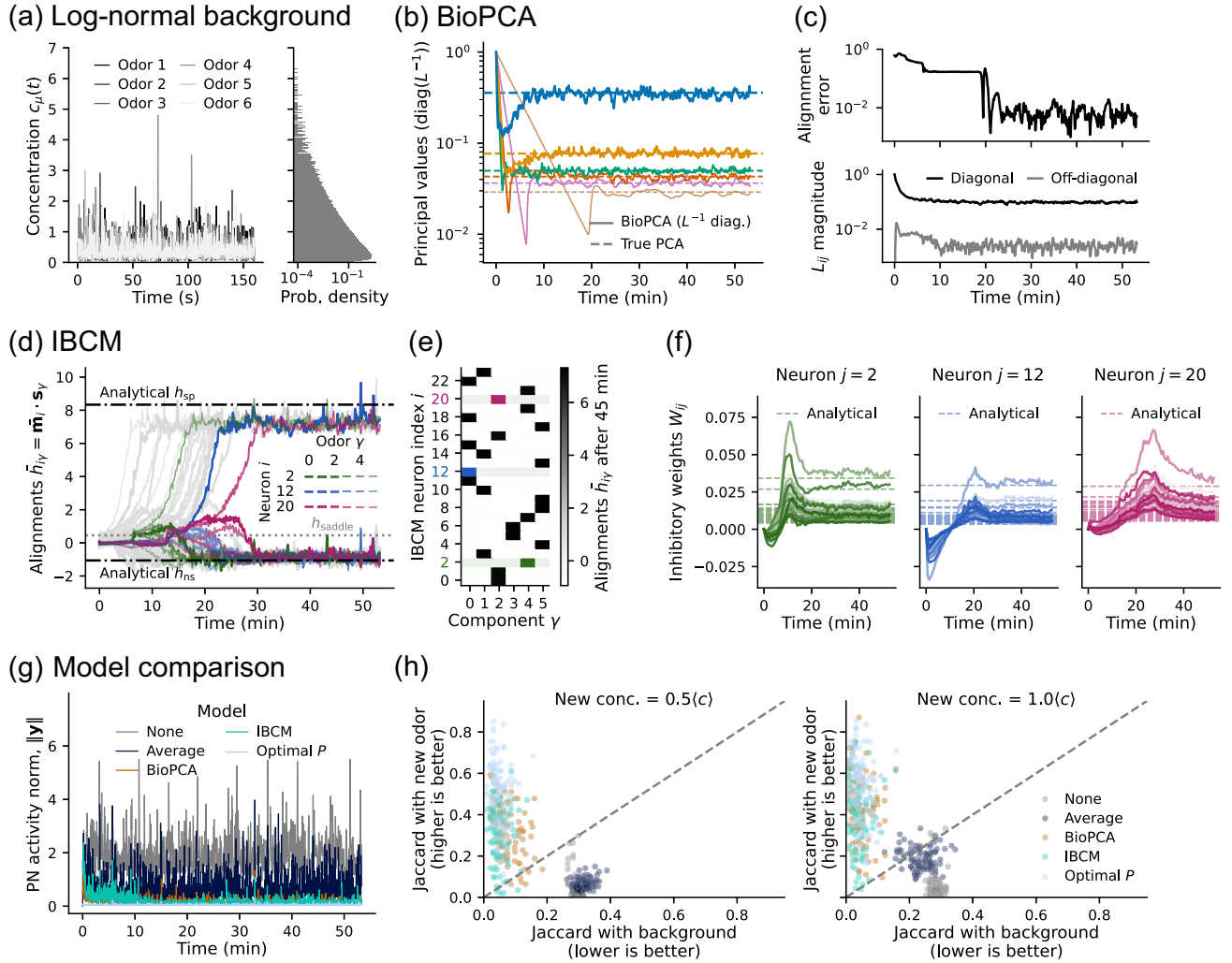


FIG. 12. Habituation to log-normal background statistics. (a) (Left) Excerpt from the concentration time series of  $N_B = 6$  odors, and (right) histogram of the concentrations showing they follow a log-normal distribution. (b) In a six-neuron BioPCA network, learning dynamics of the  $L$  matrix diagonal entries, which converge to the background's principal values, as predicted in Appendix E (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  $N_I = 24$  IBCM neurons' alignment with the background odors  $\hat{\mathbf{s}}_y$ , with three neurons highlighted ( $N_B = 6$  dot products each). Each neuron aligns with one odor and reaches dot product magnitudes close to the analytical expressions for  $h_{sp}$ ,  $h_{ns}$  (Appendix D 5). Different neurons select different odors. (e) Table summarizing the alignment of each IBCM neuron (with three highlighted). Plotted values are the dot products  $\tilde{h}_{iy}$  averaged over the last 15 min of the simulation example. (f) Time series of the  $W$  weights in the IBCM network, showing their convergence to the analytical predictions [Eq. (D32)]. (g) Example PN time series during habituation by IBCM, BioPCA, and optimal manifold learning networks, compared to the absence of habituation (OSN input) and average subtraction. (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 or with average subtraction for most samples (below diagonal).

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

$$\mathbf{w}_{\pm} = \frac{\alpha}{2\alpha + \beta} (\mathbf{s}_d \pm \sigma \mathbf{s}_s). \quad (\text{F4})$$

Hence, each IBCM neuron inhibits the off-average component for which it is selective,  $\mathbf{s}(\bar{v} = \pm\sigma)$ . Combining the two

IBCM neurons, the instantaneous PN activity is reduced to

$$\begin{aligned} \mathbf{s}(t) &= \mathbf{s}_d + \bar{v}(t) \mathbf{s}_s - \bar{h}_+ \mathbf{w}_+ - \bar{h}_- \mathbf{w}_- \\ &= \mathbf{s}_d - \frac{\alpha}{2\alpha + \beta} \left( 1 + \frac{\bar{v}}{\sigma} \right) (\mathbf{s}_d + \sigma \mathbf{s}_s) \\ &\quad - \frac{\alpha}{2\alpha + \beta} \left( 1 - \frac{\bar{v}}{\sigma} \right) (\mathbf{s}_d - \sigma \mathbf{s}_s) = \frac{\beta/\alpha}{2 + \beta/\alpha} \mathbf{s}(t). \quad (\text{F5}) \end{aligned}$$

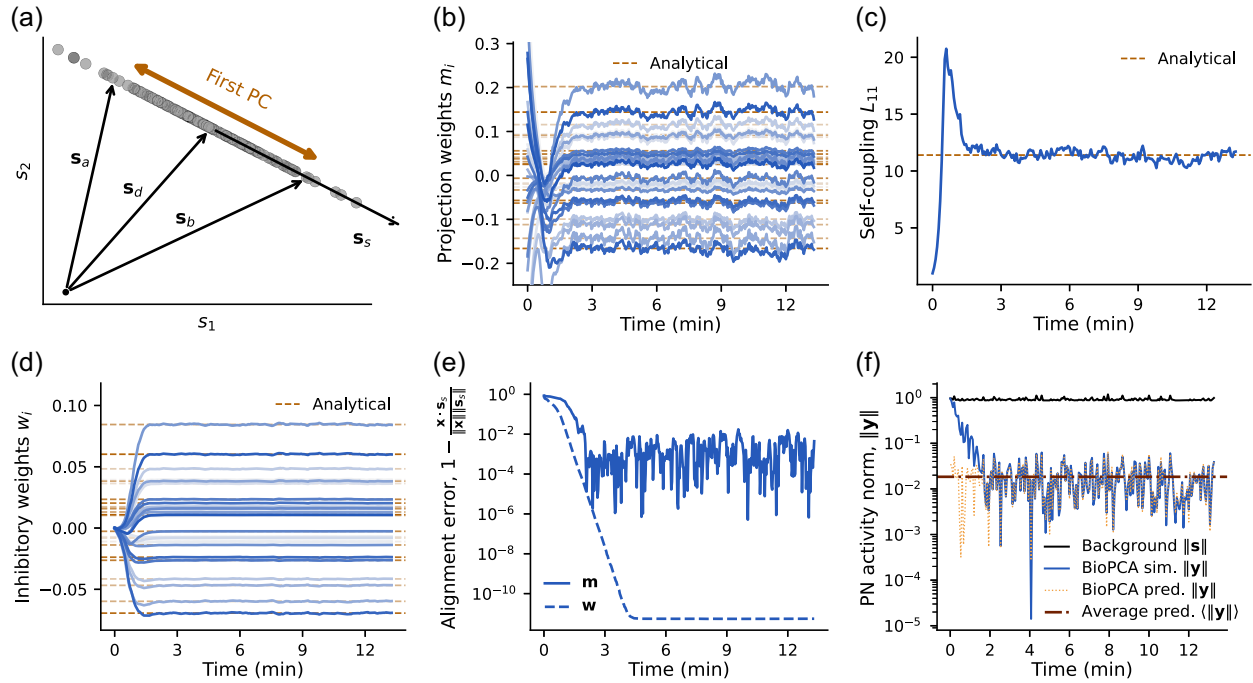


FIG. 13. BioPCA model learning dynamics on a simplified background model. We simulate a network with one BioPCA interneuron habituating to the same two-odor toy model used in Fig. 10. (a) The first principal component (PC) corresponds to the stochastic component  $\mathbf{s}_s$  along which the background 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  $\sigma^2 \|\mathbf{s}_s\|$  (dashed lines), as predicted in Eq. (F11). (c) Time series of the self-coupling  $L_{11}$  of this neuron, which converges to the inverse of the eigenvalue (dashed), Eq. (F10). (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. (F12). (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. (F13). The response to the background is reduced to  $\sim 2\%$  of its original amplitude.

Figures 10(f) and 10(g) show close agreement at steady-state between numerical simulations and Eqs. (F4) and (F5). Hence, by learning  $N_B = 2$  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 the learned background.

## 2. 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  $\alpha, \beta$  rates chosen,  $W$  weights converge at a similar pace.

We again make a quasistatic approximation on the threshold  $\Theta$ , assuming it averages over the fast background fluctuations but also converges fast enough to track the slow

variations of  $\mathbf{m}$ ,

$$\Theta = \langle (\mathbf{m} \cdot \mathbf{s}_d + \bar{\mathbf{m}} \cdot \mathbf{s}_s)^2 \rangle = h_d^2 + \sigma^2 h_s^2,$$

where we made use of  $\langle \bar{\mathbf{v}} \rangle = 0$  and  $h(t) = h_d + \bar{\mathbf{v}}(t)h_s$ . Then, we derive dynamical equations for the slow variables  $h_d$  and  $h_s$ , by taking the dot product of  $\frac{d\mathbf{m}}{dt}$  with  $\mathbf{s}_s$  and  $\mathbf{s}_d$ , averaging over fast timescales of  $\bar{\mathbf{v}}(t)$ , and using the quasistatic  $\Theta$  above. To simplify calculations, we assume that the two odor vectors,  $\mathbf{s}_a$  and  $\mathbf{s}_b$ , have the same norm (like the  $\hat{\mathbf{s}}_\gamma$  in the general case are unit normed); in this case, the vectors  $\mathbf{s}_d$  and  $\mathbf{s}_s$  are orthogonal. Making use of these properties, we calculate, for instance, for  $h_d$ ,

$$\begin{aligned} \frac{dh_d}{dt} &= \left\langle \frac{d\mathbf{m}}{dt} \cdot \mathbf{s}_d \right\rangle = \mu \langle (h_d + \bar{\mathbf{v}}h_s)(h_d + \bar{\mathbf{v}}h_s - \Theta) \rangle \mathbf{s}_d^2 \\ &= \mu (h_d^2 + \sigma^2 h_s^2 - h_d(h_d^2 + \sigma^2 h_s^2)) \mathbf{s}_d^2 \\ &= \mu \mathbf{s}_d^2 (1 - h_d)(h_d^2 + \sigma^2 h_s^2). \end{aligned} \quad (\text{F6})$$

By a similar calculation, we find for  $h_s$

$$\frac{dh_s}{dt} = \mu \sigma^2 \mathbf{s}_s^2 h_s (2h_d - (h_d^2 + \sigma^2 h_s^2)). \quad (\text{F7})$$

From Eqs. (F6) and (F7), we can conclude there will be two phases to the dynamics if the initial values of  $h_s(0) = \epsilon_s$  and  $h_d(0) = \epsilon_d$  are small, and  $\sigma^2$  is small also. The only positive

term in  $\frac{dh_s}{dt}$  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,  $\epsilon_s$ . 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  $t_d + t_s$ .

We compute  $t_d$  (first phase duration) by integrating Eq. (F6) from 0 to some fraction  $\xi$  (close to unity; we use  $\xi = 0.9$  in practice) of the steady-state  $h_d = 1$ , with the assumption that  $h_s^2$  is approximately constant and subdominant in that phase; i.e.,  $h_s^2 \approx \epsilon_s^2 \approx 0$ . We find

$$\int_{\epsilon_d}^{\xi} \frac{dh_d}{h_d^2(1-h_d)} = \int_0^{t_d} \mu s_d^2 dt \Rightarrow t_d = \frac{1}{\mu s_d^2} \left[ \frac{1}{\epsilon_d} - \frac{1}{\xi} + \ln \left( \frac{\xi(1-\epsilon_d)}{\epsilon_d(1-\xi)} \right) \right], \quad (\text{F8})$$

where  $s_d^2 = \|\mathbf{s}_d\|^2$ .

Then, once  $h_d \approx 1$ , the second phase starts. We neglect subdominant terms and we integrate from  $t_d$  (time at which  $h_d \approx 1$  but  $h_s \approx \epsilon_s$  still) to  $t_d + t_s$ . We integrate  $h_s$  from  $\epsilon_s$  to  $\pm \xi/\sigma$ : depending on the sign of the initial value  $\epsilon_s$ , the system goes to either fixed point  $\pm 1/\sigma$  (same sign as the initial value). Hence,

$$\begin{aligned} \frac{dh_s}{dt} &\approx \mu s_s^2 \sigma^2 h_s \\ \Rightarrow \int_{\epsilon_s}^{\pm \xi/\sigma} \frac{dh_s}{h_s} &= \mu s_s^2 \sigma^2 h_s \int_{t_d}^{t_d+t_s} dt \\ \Rightarrow t_s &= \frac{1}{\mu s_s^2 \sigma^2} \ln \left| \frac{\xi}{\sigma \epsilon_s} \right|. \end{aligned} \quad (\text{F9})$$

Figures 10(d) and 10(e) show that the approximations (F8) and (F9) hold well in a range of initial values  $\epsilon_s$  between 0.005 and 0.05, and  $\epsilon_d$  between 0.01 and 0.1. Importantly, these analytical expressions show that convergence time is faster if initial conditions are larger (note inverse  $\epsilon$  terms) and the noise  $\sigma^2$  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.

### 3. 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,  $\mathbf{s}_s$  [Fig. 13(a)], a single BioPCA neuron is needed to capture it. Then, the matrix  $M$  is only a row vector  $\mathbf{m}^\top$  and the matrix  $L$  is only a scalar  $\ell$ , so  $L' = \ell' = 1/\ell$ . Likewise,  $W$  is a column vector  $\mathbf{w}$  and the matrix  $\underline{A}$  is just the scalar parameter  $\Lambda$ . Since we assume BioPCA receives the input with the average subtracted, the background we consider is  $\tilde{\mathbf{s}} = \bar{\mathbf{v}}(t)\mathbf{s}_s$ . We illustrate how to solve the BioPCA steady-state equations in this simple case. With  $\bar{h} = \bar{\mathbf{v}}(t)\mathbf{m}^\top\mathbf{s}_s/\ell'$ , the

dynamical equations simplify to

$$\begin{aligned} \frac{1}{\mu_M} \frac{d\mathbf{m}^\top}{dt} &= (\bar{\mathbf{v}}(t)\mathbf{m}^\top\mathbf{s}_s/\ell')\bar{\mathbf{v}}(t)\mathbf{s}_s^\top - \mathbf{m}^\top, \\ \frac{1}{\mu_L} \frac{d\ell'}{dt} &= (\bar{\mathbf{v}}(t)\mathbf{m}^\top\mathbf{s}_s/\ell')^2 - \Lambda^2\ell'. \end{aligned}$$

Averaging over  $\bar{\mathbf{v}}(t)$  and setting the derivatives equal to zero, we have two fixed-point equations,

$$\begin{aligned} 0 &= \sigma^2 h_s / \ell' \mathbf{s}_s - \mathbf{m}, \\ 0 &= \sigma^2 h_s^2 / \ell'^2 - \Lambda^2 \ell', \end{aligned}$$

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

$$L_{11} = 1/L'_{11} = \frac{1}{\sigma^2\|\mathbf{s}_s\|^2}. \quad (\text{F10})$$

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

$$\mathbf{m} = \Lambda\sigma^2\|\mathbf{s}_s\|\mathbf{s}_s. \quad (\text{F11})$$

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

$$\frac{d\mathbf{w}}{dt} = \alpha(\Lambda\|\mathbf{s}_s\|\bar{\mathbf{v}}(t))(\bar{\mathbf{v}}(t)\mathbf{s}_s - \mathbf{w}(\Lambda\|\mathbf{s}_s\|\bar{\mathbf{v}}(t))) - \beta\mathbf{w}.$$

Averaging over  $\bar{\mathbf{v}}(t)$  and solving for  $\mathbf{w}$ , we find that it is also a vector parallel to  $\mathbf{s}_s$ ,

$$\mathbf{w} = \frac{\sigma^2\Lambda\|\mathbf{s}_s\|}{\sigma^2\|\mathbf{s}_s\|^2\Lambda^2 + \beta/\alpha} \mathbf{s}_s. \quad (\text{F12})$$

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

$$\mathbf{y}(t) = \tilde{\mathbf{s}}(t) - WLM\tilde{\mathbf{s}}(t) = \frac{\beta/\alpha}{\beta/\alpha + \sigma^2\Lambda^2\|\mathbf{s}_s\|^2} \bar{\mathbf{v}}(t)\mathbf{s}_s. \quad (\text{F13})$$

Figure 13 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  $\frac{\beta/\alpha}{\beta/\alpha + \sigma^2\Lambda^2\|\mathbf{s}_s\|^2}$ , to be compared with the reduction achieved by the IBCM network, in Eq. (F5).

## APPENDIX G: ANALYSIS OF IBCM AND BIOPCA CONVERGENCE DYNAMICS

The goal of this Appendix is to derive analytical approximations of the full IBCM and BioPCA convergence dynamics to habituation. We check these predictions quantitatively in a simplified (weakly non-Gaussian) background in Fig. 14 and qualitatively in the full turbulent background simulations of Fig. 15. Figure 14 uses the default IBCM model, while Fig. 15 uses the Law and Cooper variant where the learning



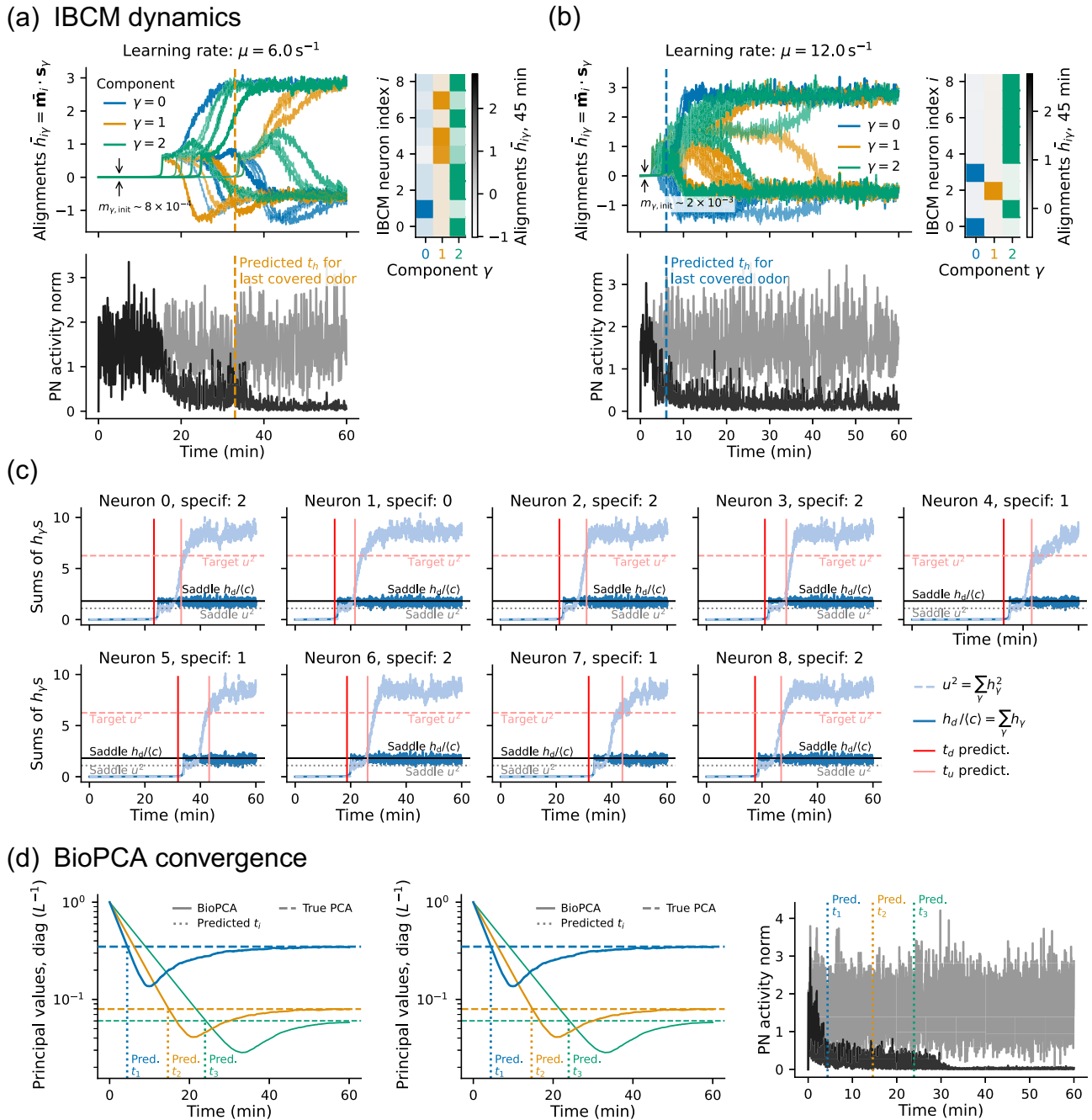


FIG. 14. Convergence time analysis of IBCM and BioPCA neurons in the weakly non-Gaussian background of Fig. 3. (a) Top: Alignments  $h_y = \mathbf{m}^T \mathbf{s}_y$  as a function of time, in a network of 9 uncoupled IBCM neurons, when the learning rate and the initial  $\mathbf{m}$  weights are small. Right: table summarizing the specificity reached by each neuron. Bottom: PN response over time, showing that the network completes habituation once each background odor is covered. (b) Same as panel (a), but for a larger learning rate and initial  $\mathbf{m}$  synaptic weights, resulting in faster convergence. (c) Trajectories of individual neurons from the simulation in panel (a), plotted in terms of the mean alignment  $h_d = \langle c \rangle \sum_y h_y$  and the sum of squared alignments,  $u^2 = \sum_y h_y^2$ . Red vertical lines mark the analytical prediction for the duration of the first phase of convergence, during which neurons reach the saddle point (expression in Appendix D). The dynamics take the form  $dh_d/dt \sim h_d^2$ , leading to a convergence time  $t_h$  derived in Appendix G 1. Pink vertical lines mark the analytical prediction for the duration  $t_u$  of the second phase of convergence, during which neurons diverge from the saddle point exponentially with a rate given by the largest eigenvalue of the Jacobian matrix at the saddle point (Appendixes D and G 1). The phase II duration prediction  $t_u$  is made after the actual point at which the saddle is reached. (d) Convergence of  $N_1 = N_B = 3$  BioPCA neurons in the same non-Gaussian background. The convergence time is primarily set by the exponential decay of each diagonal entry of the coupling matrix  $L^{-1}$  from their initial value of 1. Left:  $L$  diagonal entries vs true principal values (horizontal lines); the vertical lines are the analytical predictions of the times  $t_i$  at which  $L_i^{-1}$  converge to the true principal values  $\sigma_i^2$  with rates  $\underline{\Lambda}_{ii}$ . Middle: soon after  $t_i$ , the synaptic weights  $\mathbf{m}_i$  for neuron  $i$  increase in norm, aligning with the principal vector  $i$ . Right: habituation occurs soon after the  $N_B$ th neuron has converged to its principal value.

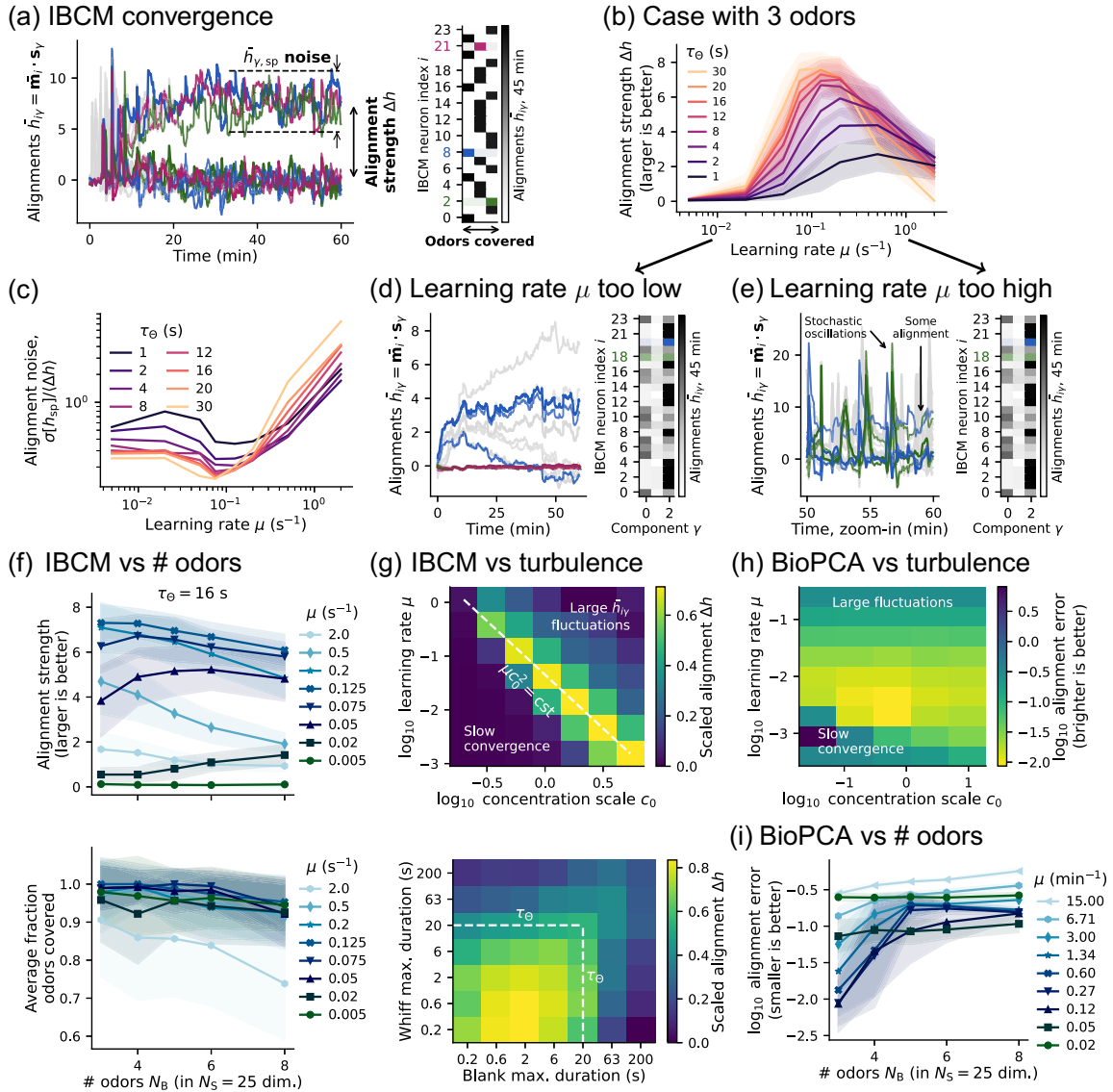


FIG. 15. Convergence of biologically plausible manifold learning networks as a function of various model and turbulence parameters. (a) Definition of IBCM convergence metrics: the alignment strength  $\Delta h$  is the difference between the largest and second largest  $h_y$  of a neuron. We also consider the variance over time of the  $h_y$ s and the fraction of odors covered by (i.e., aligning with) at least one neuron (right). We perform simulations in  $N_S = 25$  OR type dimensions. (b) Alignment strength as a function of the learning rate  $\mu$  and threshold averaging time  $\tau_\Theta$ , in a 3-odor turbulent background. For each choice of parameters, the line and shaded area are the mean and standard deviation across 32 simulations (different backgrounds) of the alignment strength averaged across 24 neurons in each simulation. (c) Alignment noise normalized by the alignment strength, averaged across 32 simulations and all neurons. (d) Example IBCM dynamics when  $\mu$  is too low: most neurons do not reach the saddle point to converge to specific alignment. (e) Example IBCM dynamics when  $\mu$  is too high: neurons reach a nonzero average alignment but with large stochastic oscillations, similar to oscillatory fixed points reported by Ref. [53] when  $\tau_\Theta$  is not  $\ll 1/\mu_\Theta$ . (f) Top: alignment strength as a function of the number of background odors and the learning rate. Bottom: average fraction of odors covered. Statistics across 32 simulations as in panels (b), (c). (g) Top: average alignment strength as a function of  $\mu$  and the absolute concentration scale  $c_0$ , which sets the OSN inputs magnitude.  $\Delta h$  is scaled by the theoretical alignment strength ( $h_{sp} - h_{ns}$ ) to keep the convergence metric comparable across conditions. IBCM convergence depends on this magnitude, such that compensating  $c_0$  with  $\mu \sim 1/c_0^2$  maintains convergence; see Fig. 14 and Appendix G. Bottom: as a function of whiff and blank maximum durations  $T_{\max,w}$ ,  $T_{\max,b}$ , when  $\tau_\Theta = 20$  s. In the simulations at each grid point of this panel, the learning rate  $\mu$  is adjusted (default:  $\mu = 0.125$  s<sup>-1</sup> at  $T_{\max,w} = 5$  s,  $T_{\max,b} = 8$  s) to keep  $\mu \langle c \rangle^2$  constant as  $\langle c \rangle$  varies with the choice of whiff and blank durations. (h) Convergence of a BioPCA network with  $N_I = N_B = 3$  neurons, measured by the subspace alignment error (lower is better) defined in the Methods, as a function of the concentration scale  $c_0$  and the learning rate  $\mu$ . BioPCA is less sensitive than IBCM to the background magnitude, depending only logarithmically on it (see Appendix G), but slow convergence and large fluctuations occur in both models outside of an optimal range of  $\mu$ . (i) BioPCA convergence as a function of the number of background odors; BioPCA appears less robust to manifold dimensionality than IBCM. Its performance as duration of whiffs and blanks is varied (not shown) behaves similarly to that of the IBCM model, with good convergence for  $t_w$  and  $t_b$  up to 20 s. Values reported in panels (g)–(i) are averages across 32 simulations performed for each set of parameters.

rate is divided by  $k_\Theta + \Theta$  [Eq. (J3)]. This variant does not qualitatively change the behavior as a function of  $\mu$ .

### 1. IBCM convergence occurs in two phases

To make progress, we analyze an individual IBCM neuron without coupling to other LNs ( $\eta = 0$ ); the scaling of the derived expressions generalize to complete networks in turbulent backgrounds (Fig. 15), clarifying how background and model parameters influence convergence of IBCM neurons to selective fixed points.

We start from the dynamical equation for the alignment of the synaptic weights  $\mathbf{m}$  with (unit-normed) odor  $\hat{\mathbf{s}}_\gamma$ ,  $h_\gamma = \mathbf{m}^\top \hat{\mathbf{s}}_\gamma$ . We average over fast background fluctuations and neglect temporal correlations between  $\Theta$  and  $\mathbf{m}$  as in Appendix D. The resulting equation reads

$$\frac{1}{\mu} \frac{dh_\gamma}{dt} = \sum_v [\langle c \rangle (h_d^2 + \sigma^2 u^2) (1 - h_d) - \sigma^2 h_v (h_d^2 + \sigma^2 u^2 - 2h_d) + m_3 h_v^2] \hat{\mathbf{s}}_\gamma^\top \hat{\mathbf{s}}_v, \quad (\text{G1})$$

where  $h_d = \langle c \rangle \sum_{\gamma=1}^{N_B} h_\gamma$ ,  $u^2 = \sum_\gamma h_\gamma^2$ , and  $\langle c \rangle$ ,  $\sigma^2$ , and  $m_3$  are the mean, variance, and central third moment of the i.i.d. background concentrations  $c_\gamma$ .

Numerical simulations show that these dynamics follow two phases. First, the neuron approaches the saddle point (Appendix D4a) where all alignments are equal. Second, the neuron diverges exponentially from the saddle point and toward one of the selective fixed points, at a rate given by the largest eigenvalue of the Jacobian matrix at the saddle point. Habituation (reduction in PN activity  $\mathbf{y}$ ) is finalized once at least one IBCM neuron has aligned with each background odor. In other words, background subtraction is maximal once the  $M$  weights provide a complete basis for the background subspace, such that the  $W$  matrix can subtract the background component by combining the complete set of projections stored in LN activities  $\mathbf{h}$ .

#### a. Phase I

For phase I, the convergence to the saddle point can be analyzed using the equation for  $h_d$ . We obtain this equation by summing Eq. (G1) over components  $\gamma$  and multiplying by  $\langle c \rangle$ . Sums such as  $\sum_{\gamma,v} h_v d_{\gamma v}$  arise, involving dot products between odors,  $d_{\gamma v} = \hat{\mathbf{s}}_\gamma^\top \hat{\mathbf{s}}_v$ . These terms depend on the specific background odor vectors selected; to make progress, we make the mean-field approximation that  $f_v = \sum_\gamma d_{\gamma v} \approx f_0$ , equal to  $1 + (N_B - 1)d_{\cos}$ , where  $d_{\cos}$  is the average cosine distance between odors from the background ensemble. For unit-normed odors with exponentially distributed elements,  $d_{\cos} = 0.516 \pm 0.002$  for  $N_S = 50$  dimensions,  $d_{\cos} = 0.531 \pm 0.002$  for  $N_S = 25$  dimensions. Then,  $\sum_v f_v \approx N_B f_0$ ,  $\sum_v h_v f_v \approx f_0 h_d / \langle c \rangle$ , and  $\sum_v h_v^2 f_v \approx f_0 u^2$ , so the  $h_d$  dynamics obey approximately

$$\frac{1}{\mu} \frac{dh_d}{dt} \approx \langle c \rangle^2 (h_d^2 + \sigma^2 u^2) (1 - h_d) N_B f_0 - \sigma^2 h_d (h_d^2 + \sigma^2 u^2 - 2h_d) f_0 + m_3 u^2 f_0.$$

Initially, the alignments are small,  $h_\gamma \ll 1$ , so  $h_d \gg u^2$ ,  $h_d^2$ . Keeping only the terms of lowest order in  $h_\gamma$ , we find

$$\frac{dh_d}{dt} = \mu f_0 (N_B \langle c \rangle^2 + 2\sigma^2) h_d^2 + \mathcal{O}(h_\gamma^3) \approx A h_d^2. \quad (\text{G2})$$

This separable, nonlinear ODE can be integrated exactly, giving  $1/h_d(t) = 1/h_{d,0} - At$ . We can thus invert for the convergence time to the saddle point,  $h_{d,\text{saddle}}$ :

$$t_h = \frac{1}{A} \left( \frac{1}{h_{d,0}} - \frac{1}{h_{d,\text{saddle}}} \right), \quad (\text{G3})$$

where  $A = \mu f_0 (N_B \langle c \rangle^2 + 2\sigma^2)$  and  $f_0 = 1 + (N_B - 1)d_{\cos}$  was defined above. This expression closely matches the actual convergence time in simulations with uncoupled neurons and small initial weights  $\mathbf{m}$  in Fig. 14. In these simulations, each neuron is initialized with random weights  $m_j = \frac{1}{2} m_0 (1 + u_j)$ , where  $u_j$  are uniform (0, 1) samples and  $m_0$  is a small amplitude parameter.

From the expression for  $t_h \sim 1/A$ , we understand why the IBCM model is sensitive to the absolute magnitude of the background concentration in Fig. 15(g), top: in A, both  $\langle c \rangle^2$  and  $\sigma^2$  are proportional to the concentration scale  $c_0$ , so  $t_h \sim 1/(\mu c_0^2)$ . Thus, keeping  $\mu c_0^2$  constant leaves the convergence time unchanged. Note that if the background odor vectors are not unit-normed (e.g., for the affinity vectors used for nonlinear OSN responses in Appendix I), their magnitude also influences convergence and can be absorbed into  $c_0$  to set the compensated learning rate.

#### b. Phase II

Once the neuron has approached the saddle point, a second phase in the dynamics starts: the trajectory starts to move away from the saddle point along one of its unstable directions, toward one of the selective fixed points. The linearized dynamics near the saddle predict an exponential escape with a rate set by the largest eigenvalue of the Jacobian matrix computed at the saddle point. To evaluate this rate, we use our analytical expressions for the IBCM fixed points [Eqs. (D17), (D18), and (D19)] and the Jacobian matrix [Eq. (D21)] derived in Appendix D. We diagonalize the matrix numerically to extract the exponential escape rate.

For this phase, we study the  $u^2 = \sum_\gamma h_\gamma^2$  coordinate, since it encompasses the specific  $h_\gamma$  converging to  $h_{\text{sp}}$  and the other alignments  $h'_\gamma$  approaching the nonspecific value  $h_{\text{ns}}$ . If  $u^2$  has an initial value  $u_{\text{saddle}}^2 + \Delta u_s^2$  at the end of phase I (i.e., at  $t_h$ ), then  $u^2$  increases as

$$u^2(t) = u_{\text{saddle}}^2 + \Delta u_s^2 e^{(t-t_h)/\tau_{\text{jac}}},$$

where  $\tau_{\text{jac}}$  is the inverse of the largest eigenvalue real part of the Jacobian matrix. To obtain a duration for phase II, we solve for the time  $t_u$  (measured from  $t = 0$ ) at which  $u^2$  reaches some target value. In practice, we take  $u_{\text{target}}^2 = 1/\sigma^2$ , which is below the fixed point for any background  $m_3 > 0$ . This gives a total convergence time (phase I + phase II) of

$$t_u = t_h + \tau_{\text{jac}} \log \left( \frac{u_{\text{target}}^2 - u_{\text{saddle}}^2}{\Delta u^2} \right). \quad (\text{G4})$$

In Fig. 14, this prediction is made for each neuron by using its actual phase I time  $t_h$  and its initial  $\Delta u^2$  at that time. The predictions are still accurate, but less than for phase I, since the exponential dynamics are only valid near the saddle point.

## 2. BioPCA dynamics

The main determinant of BioPCA convergence is the exponential decay time of the inverse coupling matrix  $L'$ 's diagonal elements to principal values of the background distribution. We initialize these elements  $L'_{ii} = 1$ , so they must decay to the principal values  $\sigma_i^2 < 1$  for the network to reach a fixed point. Initially, when  $L' = L^{-1}$  is too large, the  $L$  matrix is too small, so the LN activities  $\mathbf{h} = L\mathbf{M}\mathbf{s}$  are small too; the BioPCA dynamics for  $L'$  and  $M$  [Methods, Eqs. (20) and (21)] are thus dominated by decay terms  $-\mu M$  and  $-\mu_L \underline{\Delta} L' \underline{\Delta}$ . Once an entry  $L'_{ii}$  reaches  $\sigma_i$ , the corresponding neuron's synaptic weights  $\mathbf{m}_i$  grow to align with the principal component  $i$ ; see the simulations in Fig. 14(d). Habituation is complete once the  $N_B$ th neuron has converged to the corresponding principal component; in other words, once  $M$  and  $L$  provide a complete set of projections, as in the IBCM network.

Based on the  $L'$  dynamical equation, the decay rate of  $L'_{ii}$  is  $\mu_L \underline{\Delta}_{ii}^2$ . In simulations, we set  $\mu_L$  to some multiple  $1/\tau_L > 1$  of  $\mu$ , we set the  $\underline{\Delta}_{ii}$  matrix entries to  $\Lambda(1 - \lambda_r \frac{i-1}{N_B-1})$ , and we divide the chosen learning rate  $\mu$  by the scale factor  $\Lambda^2$ . The range of entries is  $\lambda_r = 0.5$  by default (Methods). Hence, the actual exponential decay rate is  $\lambda_i = \mu/\tau_L(1 - \lambda_r \frac{i-1}{N_B-1})$ —independent of the scale parameter  $\Lambda$ . The time  $t_i$  for the  $i$ th entry to decay exponentially from 1 to  $\sigma_i^2$  is thus

$$\sigma_i^2 = L'_{ii}(t_i) \approx 1 \times e^{-\lambda_i t_i} \Rightarrow t_i = -\frac{1}{\lambda_i} \log \sigma_i^2. \quad (\text{G5})$$

The exponential decay approximation of the initial BioPCA dynamics is excellent and predicts nearly perfectly the time at which  $L'_{ii}$  crosses the principal value  $\sigma_i^2$ ; see Fig. 14(d).

When  $L'_{ii}$  first reaches  $\sigma_i^2$ , it keeps decaying for some time while other terms in the BioPCA equation become significant and the  $\mathbf{m}$  vectors align with principal components. We do not have an analytical expression for the duration of that phase, but we get a lower bound on the total time to complete habituation from the exponential decay time of the  $i = N_B$ th neuron, assuming there is one BioPCA neuron per background odor,  $N_I = N_B$ :

$$t_{\text{BioPCA}} \geq t_{N_B} = -\frac{\tau_L}{\mu(1 - \lambda_r)} \log(\sigma_{N_B}^2). \quad (\text{G6})$$

We observe, from Eq. (G5), that the BioPCA convergence time depends on the background concentration scale only logarithmically as  $\log(\sigma_{N_B}^2) \sim 2 \log(c_0)$ . This property makes it robust against changes in  $c_0$  without needing to compensate by a change in the learning rate  $\mu$ , as opposed to the IBCM model which has  $t_d \sim c_0^2$ .

## APPENDIX H: TESTING DIFFERENT $L$ -norms FOR THE 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 suboptimal performance, limited to the similarity between the new odor and its orthogonal component (Figs. 5 and 18). 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

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

$$\|W\|_q = \left( \sum_{i=1}^{N_S} \sum_{j=1}^{N_I} \|W_{ij}\|^q \right)^{1/q}.$$

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,

$$\frac{dW_{ij}}{dt} = \alpha \|\mathbf{y}\|_p^{2-p} |y_i|^{p-1} \text{sgn}(y_i) \bar{h}_j - \beta \|W\|_q^{2-q} |W_{ij}|^{q-1} \text{sgn}(W_{ij}). \quad (\text{H1})$$

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  $\alpha$  and  $\beta$  learning rate values, centered on relevant windows (e.g., a smaller  $p$  requires a smaller  $\alpha$  to prevent numerical instabilities). Unfortunately, different  $p, q$  choices did not fundamentally alter the model performance for habituation [Figs. 17(a)–17(c)] or new odor recognition [Figs. 17(d)–17(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 Appendix A, would be needed to further improve new odor recognition performance.

## APPENDIX I: NONLINEAR OSN RESPONSES AND OSN ADAPTATION

In this Appendix, we describe the saturating OSN response model and numerical simulations used to examine habituation to nonlinear background manifolds in Figs. 6 and 20 and the effects of OSN adaptation in Fig. 21.

### 1. OSN response model

The saturating function introduced in the Methods,

$$s_i(t) = F_{\max} \frac{\sum_{\gamma} K_{i\gamma} c_{\gamma}(t)}{\exp(\epsilon_i) + \sum_{\gamma} K_{i\gamma} c_{\gamma}(t)}, \quad (\text{23})$$



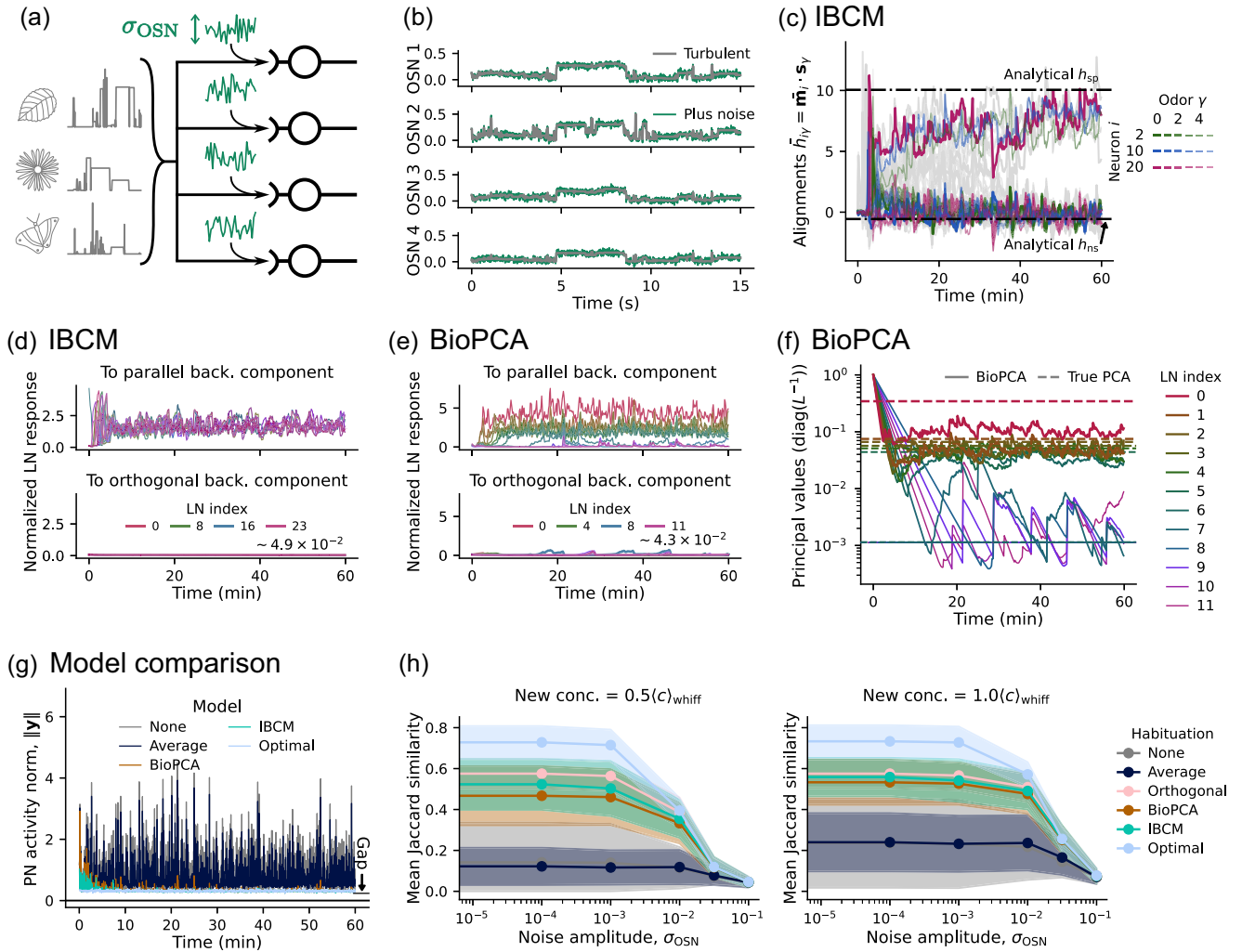


FIG. 16. Robustness against OSN noise in the IBCM and BioPCA manifold learning models. (a) Illustration of the OSN noise considered. Turbulent backgrounds with six odors are simulated as usual, then independent Gaussian white noise is added to each OSN input at each time point. We consider simulations in  $N_S = 100$  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  $\sigma_{OSN} = 0.01$ ; 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,  $\bar{h}_j(t)$ , to the parallel (top) and orthogonal (bottom) components of the background, normalized by the average norm of these components,  $\|\bar{s}_j\|$  or  $\|\bar{s}_\perp\|$ . 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 panel (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  $N_B = 7$  neurons roughly capturing the background subspace, while remaining neurons capture the small PCs along the pure Gaussian noise directions (eigenvalues equal to  $\sigma_{OSN}$ ). (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  $\sigma_{OSN}$ . Numerical experiments similar to those of Fig. 2, repeated for each  $\sigma_{OSN}$  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  $N_I = 24$  IBCM neurons, and  $N_I = 12$  BioPCA neurons.

constitutes a simplified version of the OSN response in the adaptation model from Gorur-Shandilya, *et al.* [55] and Kadakia and Emonet [20]. In that model, olfactory receptors (OR) form complexes with Orco co-receptors. OR-Orco

complexes have an active and an inactive state, with affinities  $K_{i\gamma}$  and  $K_{i\gamma}^*$ , respectively (we swap the  $*$  notation from Kadakia and Emonet), between odor  $\gamma$  and receptor type  $i$ . The quasistatic OSN firing rate in response to a mixture is, in

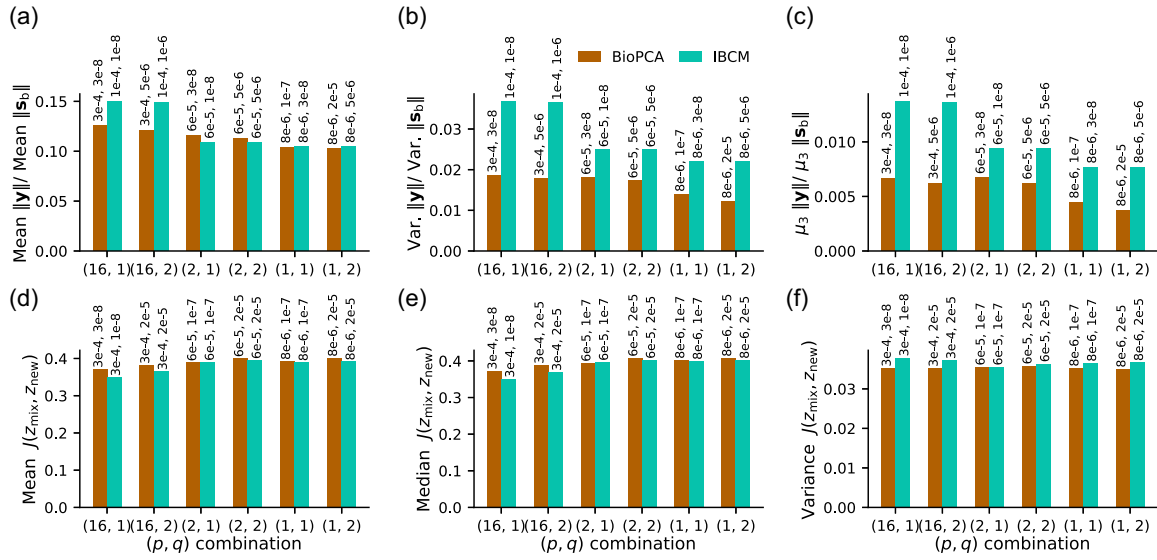
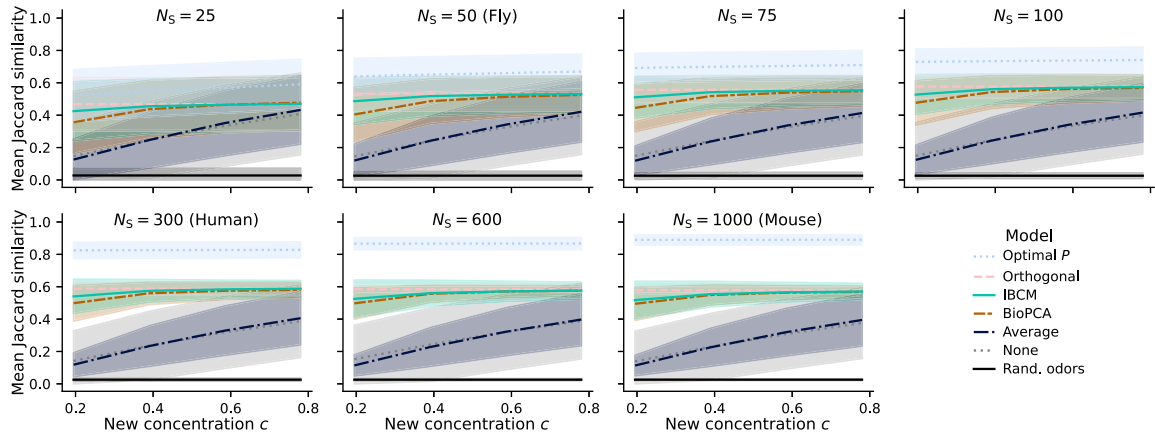


FIG. 17. Performance for various  $L^p$  norms in  $W$ 's cost function learning rule. (a) Reduction of the mean, (b) variance, and (c) third central moment of the norm of the PN response after habituation by IBCM of BioPCA networks. We consider various  $p, q$  norm choices in the generalized Hebbian learning rules derived in Appendix H. For each  $p, q$  combination, we performed a grid search over learning rates  $\alpha$  and  $\beta$ ; the bar graph reports the best performance for each  $p, q$ ; the  $\alpha, \beta$  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 panels (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.

(a) Each dimension separately



(b) Each concentration separately

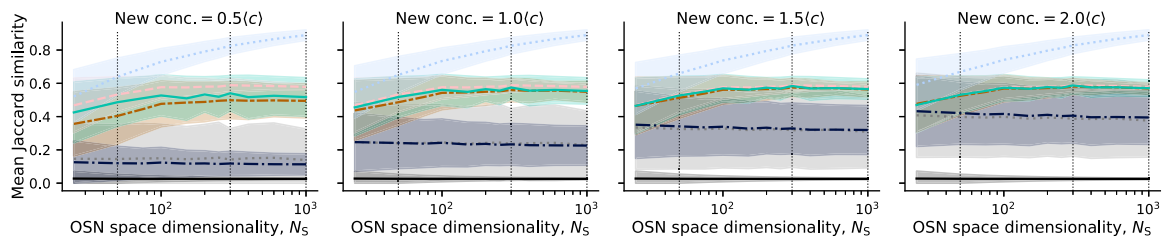


FIG. 18. Odor recognition performance as a function of dimensionality and new odor concentration. Related to Fig. 5. (a) Jaccard similarity as a function of new odor concentration, for each tested dimensionality separately. (b) Jaccard similarity as a function of dimensionality, for each tested new odor 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.

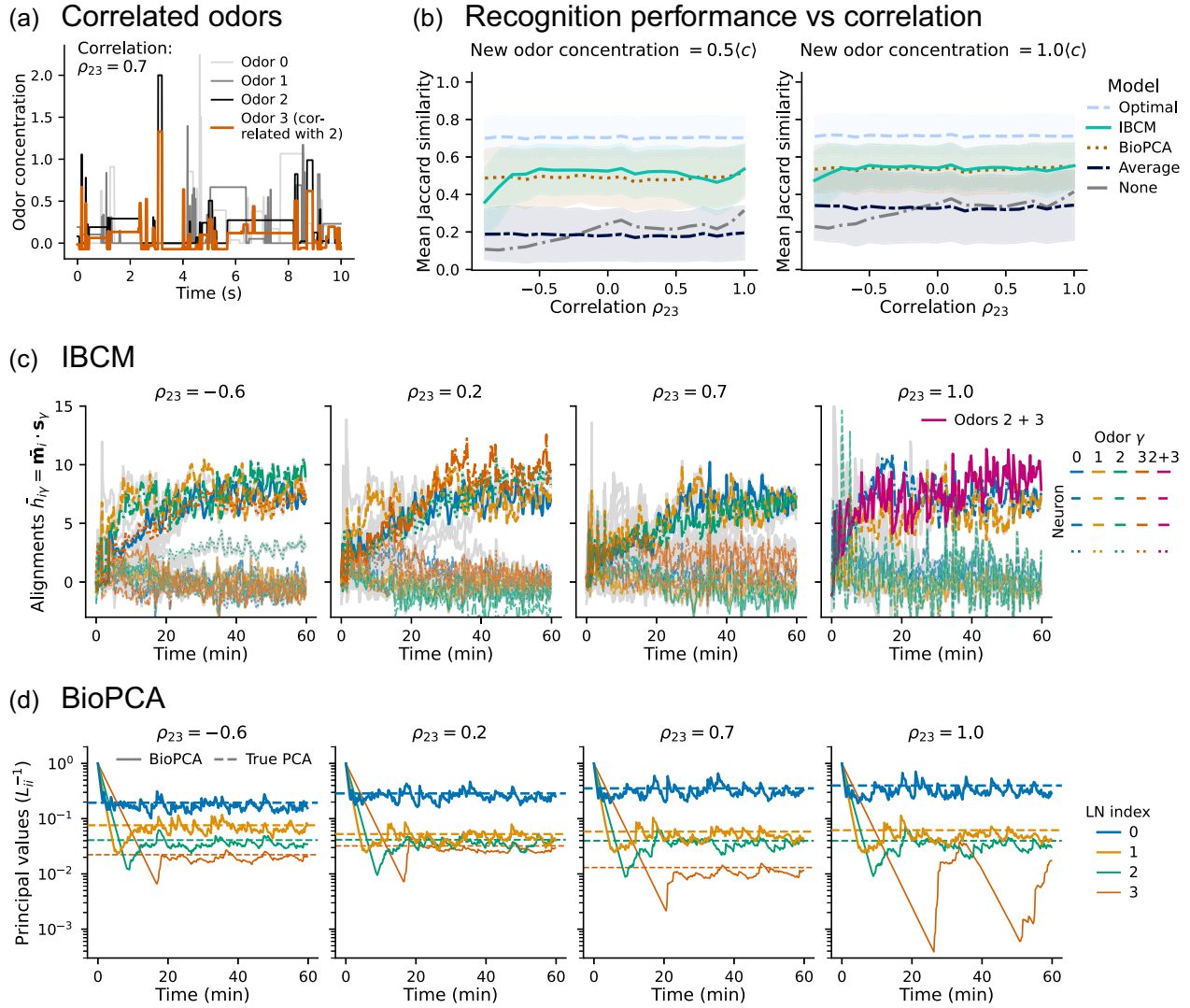


FIG. 19. Manifold learning in the presence of correlated odor concentrations. (a) Excerpt from the concentration time series of a four-odor background with turbulent statistics and nonzero Pearson correlation  $\rho_{23}$  (here,  $\rho_{23} = 0.7$ ) between the last two odors. The correlation is introduced with the Cholesky decomposition, as in Eq. (B5). (b) Performance of different models for new odor recognition, following a protocol similar to Fig. 2: statistics across 32 background seeds, testing 100 new odors at 10 test times, against 10 background samples at each time; new odor introduced at a concentration of either 0.5 ( $\langle c \rangle$ ) or 1.0 ( $\langle c \rangle$ ) after 60 min of habituation. Dimension  $N_S = 50$  OSNs. (c) Alignment of IBCM neurons with background odors for a range of  $\rho_{23}$  values. Trajectories are highlighted for three neurons, with colors indicating the different background odors  $\gamma$ . For  $\rho_{23} = 0.7$ , no neuron is aligned with odor 3, and for  $\rho_{23} = 1$ , neurons align with the combined odors  $s_2 + s_3$  rather than with either of the individually. (d) BioPCA network learning of the background's principal values for a few different  $\rho_{23}$ . At  $\rho_{23} = 1$ , the background is  $N_B - 1$ -dimensional, so one LN (neuron 3) does not have any eigenvalue to learn and thus fluctuates near small, nonzero values.

that model,

$$s_i = F_{\max} \left[ 1 + \exp(\epsilon_i) \left( \frac{1 + \sum_{\gamma} K_{i\gamma}^* c_{\gamma}}{1 + \sum_{\gamma} K_{i\gamma} c_{\gamma}} \right) \right]^{-1},$$

where  $\epsilon_i(t)$  is the free-energy difference between the inactive and active conformations in the unbound state of OR type  $i$ , in units of  $k_B T$ .

In this paper, we first fix this free-energy difference  $\epsilon_i = \epsilon$  for all OSN types to control the strength of nonlinearity in Figs. 6 and 20 (Appendix 12). Then, we explore the effects of  $\epsilon_i(t)$  adaptation on short timescales  $\tau \sim 250$  ms, as in Kadakia and Emonet's model, in Fig. 21 (Appendix 14).

The full response function simplifies to Eq. (23) if we assume a regime where affinities  $K_{i\gamma}^*$  are very small while  $K_{i\gamma}$  affinities are very large. In that case,  $1 + \sum_{\gamma} K_{i\gamma}^* c_{\gamma} \approx 1$  and  $1 + \sum_{\gamma} K_{i\gamma} c_{\gamma} \approx \sum_{\gamma} K_{i\gamma} c_{\gamma}$ , and we recover Eq. (23). Then, odors are defined only by their active affinities  $K$ , while the inactive affinities do not matter. Kadakia and Emonet use parameters in this regime in their full model, setting  $K^* = 0.01$  (in dilution units) for all odorants and  $K$  in the range  $[10^3, 10^4]$ .

However, note that the approximation is not precise for very small  $c_{\mu}$ : the full function had a nonzero baseline  $1/(1 + e^{\epsilon})$ , whereas our version introduces the simplification of a zero

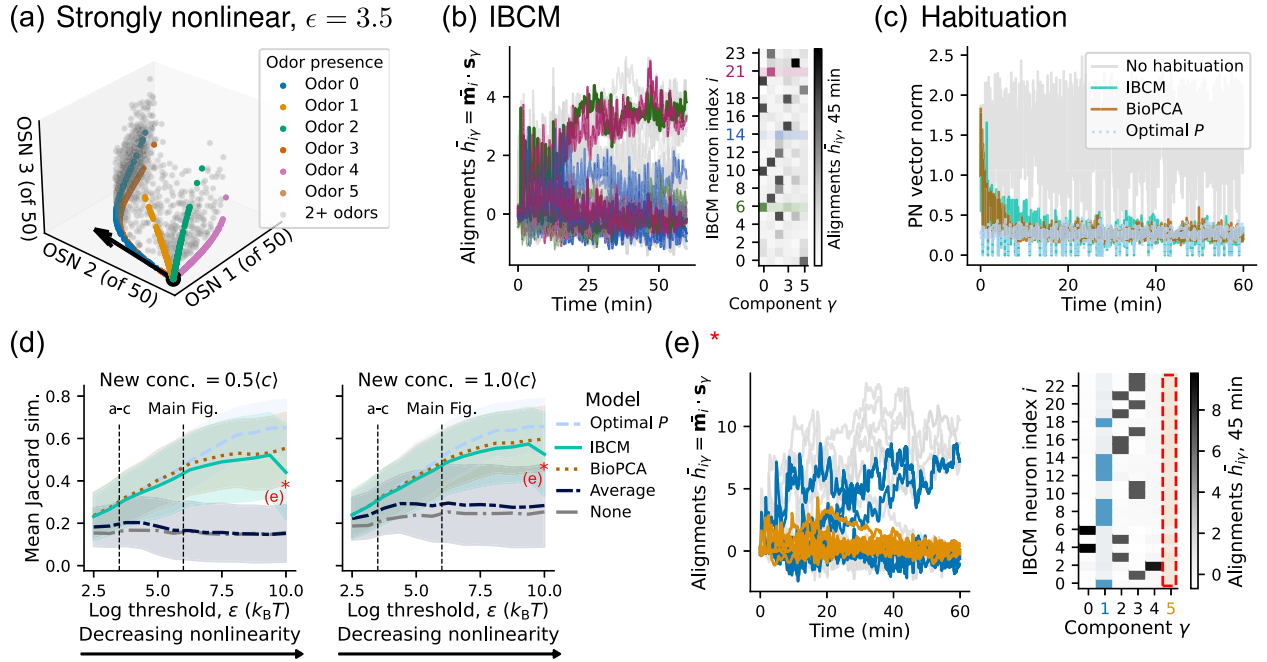


FIG. 20. Supplementary results with nonlinear OSN responses. Related to Fig. 6. (a) 3D slice of a background manifold with strong nonlinearity ( $\epsilon = 3.5$ ). (b) IBCM network habituation to the strongly nonlinear manifold in panel (a). Neurons partially align with the linear odor basis; in reality, they likely align with a distorted version of that basis due to the nonlinearities. (c) Habituation when OSN saturation is strong, as in panel (a), leaves a nonnegligible response to the background, because the linear models cannot perfectly subtract the curved manifold. (d) New odor recognition performance after habituation as a function of  $\epsilon$ , for two new odor concentrations (left panel is also in the main figure). The test procedure is similar to Fig. 2.  $F_{\max}$  is scaled with  $\epsilon$  to maintain the OSN activity magnitude in response to a strong whiff, as detailed in Appendix 12. (e) The IBCM network's dip in performance at  $\epsilon = 10$  (nearly linear manifold) occurs because neurons converge too fast and sometimes miss a background odor. This happens because (1) the  $K$  distribution in Fig. 6(b) creates large asymmetries between background odors in the linear regime especially, and (2) scaling  $F_{\max}$  does not perfectly compensate the impact of  $\epsilon$  on the input magnitude, to which IBCM convergence is sensitive (Appendix G).

baseline OR firing when  $c_i = 0$ . We make this simplification because a nonzero baseline would be removed by a simple average filter in the circuit prior to manifold learning—as we have in fact already included in the BioPCA network.

## 2. Linear regime and controlling nonlinearity strength with $\epsilon$

When  $\exp(\epsilon) \gg \sum_{\gamma} K_{iy} c_{\gamma}$ , the simplified OSN saturation function (23) reduces to the linear case we considered in the rest of our study,

$$s_i(t) \approx F_{\max} / \exp(\epsilon) \sum_{\gamma} K_{iy} c_{\gamma}(t).$$

Scaled affinity vectors  $F_{\max} / \exp(\epsilon_i) \mathbf{K}_{\gamma}$  then play the role of the constant odor vectors  $\hat{\mathbf{s}}_{\gamma}$  in  $N_S$ -dimensional OSN type space. This regime occurs when concentrations  $c$  are small or, equivalently, when the free-energy difference  $\epsilon_i$  is large.

Thus, we can use  $\epsilon$  to control the strength of OSN nonlinearity. In Figs. 6 and 20, we perform simulations at various  $\epsilon$  levels, interpolating between linear (large  $\epsilon$ ) and strongly nonlinear (small  $\epsilon$ ) background manifolds. Importantly, as we vary  $\epsilon$ , we scale the OSN response amplitude  $F_{\max}$  inversely to preserve the overall magnitude of OSN activity transmitted to LNs and PNs. Specifically, for each simulation, we set  $F_{\max}$  such that the average OSN activity,  $\frac{1}{N_S} \sum_{i=1}^{N_S} s_i(t)$  is equal to  $2.5/\sqrt{N_S}$  in response to a strong whiff of all odors at 2.5 times the average concentration. Since the IBCM convergence

time depends on the input magnitude, scaling  $F_{\max}$  ensures convergence of IBCM neurons [imperfectly; see Fig. 20(e)] to selective fixed points without having to adjust the learning rate  $\mu = 7.5 \times 10^{-4}$  for each simulation. In practice, this scale would be adjusted by adaptation [20] and normalization mechanisms [17,19] in the olfactory circuit to match the biologically determined value of  $\mu$ .

## 3. Odor affinities distribution

Following Kadakia and Emonet [20], we generate background odor affinities based on the empirical distribution of OSN sensitivities (inverse  $EC_{50}$ ) measured in *Drosophila* larvae by Si, *et al.* [54]. The affinity  $K$  is the ratio of binding and unbinding rates with the OR,  $k_{\text{on}}/k_{\text{off}}$ , and has inverse concentration units; therefore, it corresponds to the sensitivities  $x = 1/EC_{50}$  reported by Si, *et al.* We fit the empirical distribution of  $x$  with a complementary cumulative distribution function of the form

$$G_X(x) = \mathbb{P}[X > x] = \tanh\left(\frac{1}{bx^{\alpha}}\right).$$

This simple form captures not only the power-law tail  $G_X \sim x^{-\alpha}$  fitted by Si, *et al.* on the largest  $x$  samples (they find an exponent  $\alpha = 0.42$ ), but also the bend (cutoff) and plateau of the empirical CCDF at low  $x$ ; see Fig. 6(b). We estimated parameters  $\alpha$  and  $\log b$  using nonlinear least-squares curve



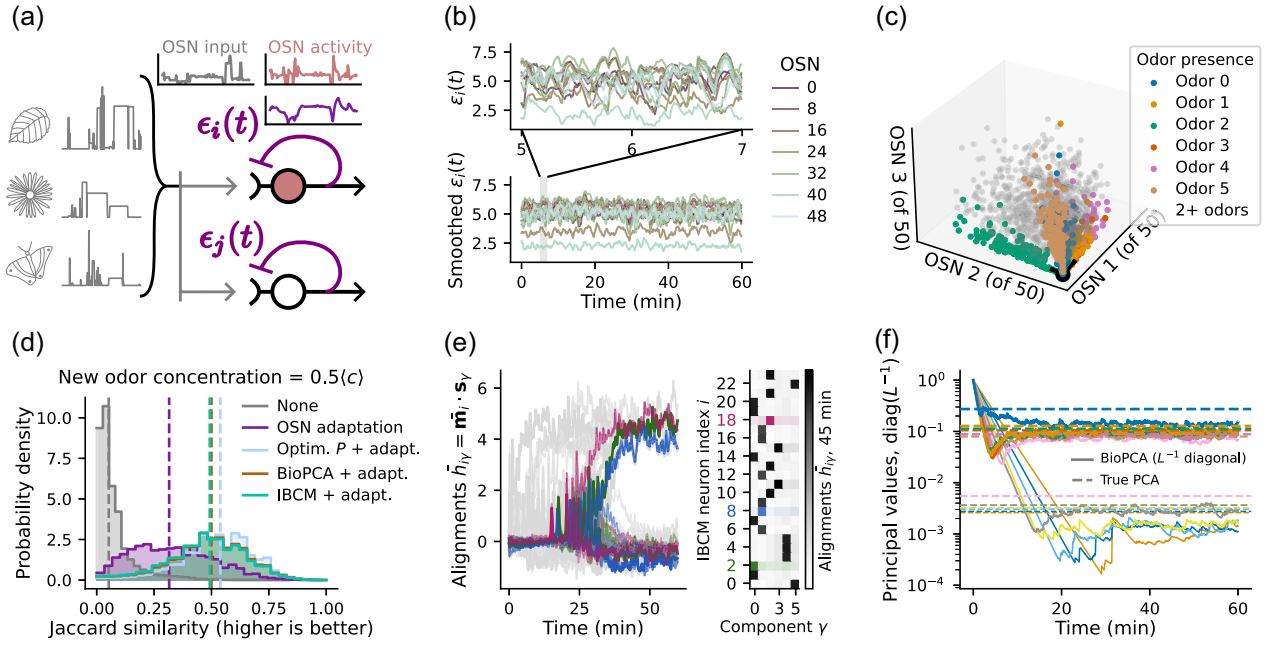


FIG. 21. Impact of fast, nonlinear OSN adaptation on habituation by manifold learning. (a) Illustration of OSN adaptation by a negative feedback of OSN activity on the OR-Orco complexes' free-energy difference, as explained in Appendix I 4. (b) Dynamics of  $\epsilon_i(t)$  for  $N_S = 50$  OSN types (24 shown) in response to a turbulent background with  $N_B = 6$  odors. Bottom: Time series smoothed with a moving average (50 s window) to show that different OSNs'  $\epsilon_i(t)$  fluctuate around different average values. Top: zoom-in (no smoothing) to show the fast adaptation dynamics within a 2-min period. (c) 3D slice of the background manifold with OSN adaptation ( $N_B = 6$ ,  $N_S = 50$ ). Adaptation maintains the background in a moderately curved regime, similar to the case  $\epsilon = 6$  in Fig. 6. (d) New odor recognition performance after habituation with OSN adaptation and manifold learning combined, compared to OSN adaptation only and no adaptation. Statistics come from 96 random backgrounds, 100 new odors tested in each background, each independently at 100 time points in the last 5 min of a 1-h habituation period. New odors are presented at  $0.5 \langle c \rangle$ . Jaccard similarity is computed with the neural tag in response to the new odor alone when the  $\epsilon_i(t)$  are set to their average values for each OSN. (e) Example IBCM habituation simulation showing that neurons still align with individual odor vectors  $\mathbf{K}_\gamma$  when there is OSN adaptation. (f) Example BioPCA network simulation for the same background, showing that the network still picks out approximately the principal components of the background distribution.

fitting of  $\log G_X$ , instead of  $G_X$  directly, to better capture the tail. We found best parameter fits of  $\log b = -1.29 \pm 0.02$  and  $\alpha = 0.388 \pm 0.003$  (fit with *SciPy's* *curve\_fit* function [80], uncertainties based on the Fisher information matrix).

For our simulations, we then generate odor vectors  $\mathbf{K}_\gamma$  by sampling i.i.d. elements from the fitted affinity CCDF, using the inverse transform method. We scale the (relative) inverse dilution units of the empirical data by  $K_{\text{scale}} = 5 \times 10^{-4}$  to match the scale of (arbitrary) concentration units in our simulations. We also clip large  $K$  to  $500x_{\text{knee}}$ , where the bend (point where the plateau is reached) in the CCDF is at  $x_{\text{knee}} = b^{-1/\alpha}$  [dotted vertical line in Fig. 6(b)]. This prevents excessively strong OR-odor affinities from dominating all activation.

#### 4. OSN adaptation

To include OSN adaptation in our network for Fig. 21, we followed the adaptation model of Gorur-Shandilya, *et al.* [55], 2017; Kadakia and Emonet, 2019 [20]. In the saturating OSN activation function, we promote the free-energy differences  $\epsilon_i(t)$  to dynamical variables receiving a negative feedback from the current activity  $s_i(t)$  of that OSN type [Fig. 21(a)], with dynamics of the form

$$\tau_a \frac{d\epsilon_i(t)}{dt} = s_{i,0} - s_i(t). \quad (11)$$

As a result,  $\epsilon_i(t)$  adapts to maintain a target OSN activity  $s_{i,0}$  on a timescale  $\tau_a = 250$  ms [Fig. 21(b)]. Moreover, as in Kadakia and Emonet, 2019,  $\epsilon_i$  is clipped between limits  $[\epsilon_L, \epsilon_H]$ , such that adaptation is only possible on a finite range.

For our simulations, we set the target activity  $s_{i,0} = s_0 = 1/\sqrt{N_S}$  (same for all OSNs) and use bounds  $\epsilon_L = 1$ ,  $\epsilon_H = 10$  (units of  $k_B T$ ). Since different OSNs have different magnitudes of affinities for the different odors  $K_{i\gamma}$ , they will also converge to different average free-energy differences, as shown in Fig. 21(b) bottom, to maintain activity near the same target  $s_0$ . Since the target amplitude depends on the maximum OSN amplitude chosen, we set  $F_{\text{max}}$  as previously for the nonlinear OSN simulations, that is, such that the average OSN response to a strong whiff at concentration  $2.5 \langle c \rangle$  is equal to  $1/\sqrt{N_S}$  when  $\epsilon$  is set to the midpoint of the allowed range,  $\frac{1}{2}(\epsilon_L + \epsilon_H)$ . This ensures that the average  $\epsilon_i$  needed for each OSN type to reach the target  $s_0$  is generally within the  $\epsilon$  bounds.

#### APPENDIX J: PROJECTION WEIGHTS SCALE FACTOR, $\Lambda$

As explained in Methods, we defined a parameter  $\Lambda$  controlling the scale of  $M$  weights and compensating for the regularization on  $W$ . Here, we explain how to introduce  $\Lambda$

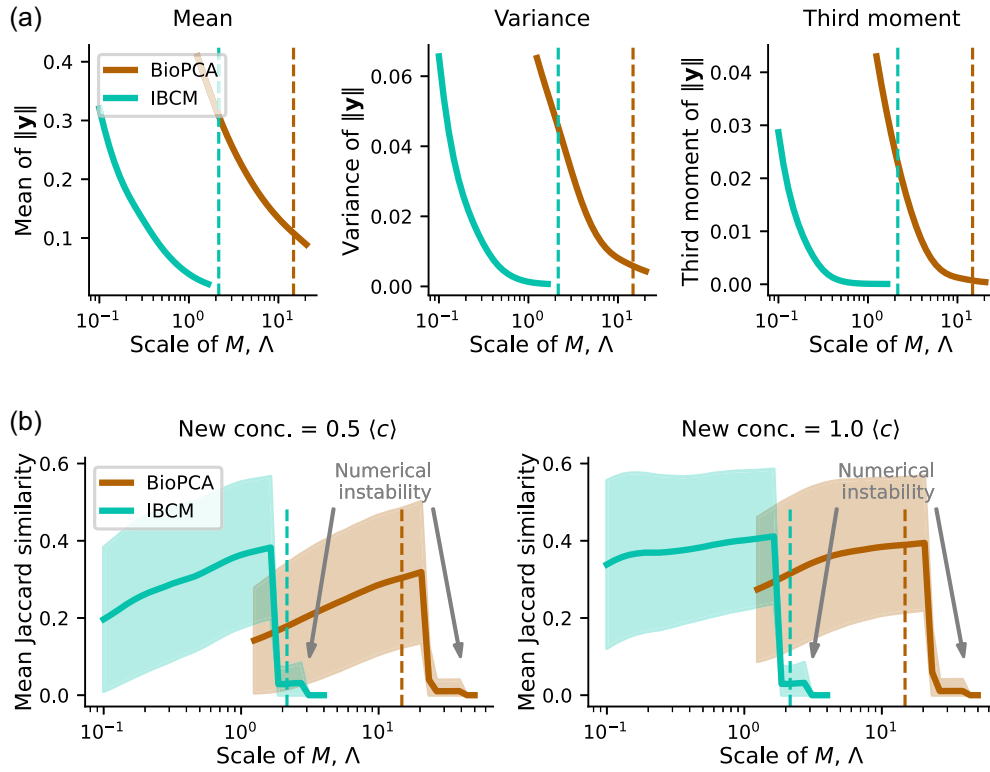


FIG. 22. Effect of the  $M$  weights scaling parameter  $\Lambda$ , to match the IBCM and BioPCA models. (a) Habituation performance, quantified by reduction in the mean, variance, and third moment of the PN response to the background,  $y$ . Habituation runs are performed on the same background time series, with different  $\Lambda$  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 panel (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  $\Lambda$  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 (Appendix B 6). Simulations were run in  $N_S = 25$  dimensions and with default parameter values.

in the IBCM model, and how to set this parameter to make BioPCA and IBCM perform equivalently.

### 1. Introducing a scale factor $\Lambda$ in the IBCM model

We start with the equations for a single neuron. We seek to introduce  $\Lambda$  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  $\Lambda$ . 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  $\Theta \rightarrow h^2/\Lambda$ . Hence, we start by modifying the threshold equation to

$$\frac{d\Theta}{dt} = \frac{1}{\tau_\Theta} (h^2/\Lambda - \Theta). \quad (J1)$$

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  $\frac{d\mathbf{m}}{dt} \sim \Lambda$ , we modify the  $\mathbf{m}$  equation to

$$\frac{d\mathbf{m}}{dt} = \frac{\mu_\Theta}{\Lambda} h(h - \Theta)\mathbf{s}(t) - \varepsilon \frac{\mu_\Theta}{\Lambda} \mathbf{m}. \quad (J2)$$

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

$$\mu_\Theta = \frac{\mu}{\Theta/\Lambda + k_\Theta}. \quad (J3)$$

The generalization to a network of IBCM neurons is straightforward, since  $\Lambda$  is a unique scale parameter for all neurons. Each  $\bar{\Theta}_i$  equation has the rescaled term  $\bar{h}_i^2/\Lambda$  as in Eq. (J1). All terms in neuron  $i$ 's  $\mathbf{m}_i$  equation, including those from coupling with other neurons  $j$ , have their learning rates  $\mu_{\bar{\Theta}_j}$  divided by  $\Lambda$ , as in Eq. (J2). Also,  $\bar{\Theta}_i$  is divided by  $\Lambda$  in the denominator of each learning rate  $\mu_{\bar{\Theta}_j}$ , as in Eq. (J3).

These are the IBCM equations we use for general  $\Lambda$  values. In Fig. 22, we characterize the performance of the network for habituation and new odor recognition as a function of the scale  $\Lambda$ , and we observe that  $\Lambda_{\text{IBCM}} = 1$  is large enough to maximize the performance.

### 2. Scaling the BioPCA model for performance equivalent to IBCM

As explained in Methods, the scale  $\Lambda$  is already built into the BioPCA model, in the matrix  $\underline{\Lambda}$  intervening in the  $L$  equation.  $\Lambda_{\text{PCA}}$  is set to 1 by default [38], but this leads to

a smaller  $M$  weights magnitude than by default in the IBCM model. For this reason, Fig. 22 shows that the BioPCA model requires a  $\Lambda_{\text{PCA}} > 1$  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  $\Lambda_{\text{PCA}}$  value should yield a comparable performance from both models. As derived in Eq. (D34), the PN response is reduced, in an IBCM network, by a factor

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

where we have explicitized how  $\Lambda_{\text{IBCM}}$  controls this factor by multiplying the default-scale LN activities,  $h_{\text{sp}}$ ,  $h_{\text{ns}}$ ,  $\bar{h}_{\text{d}}$ —recall that these are the specific, nonspecific, and average alignments of the IBCM fixed points, derived in Appendix D.

In comparison, Eq. (E4) shows that the PN activity in a BioPCA network is reduced along each principal direction by a factor of approximately

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

We set  $\Lambda_{\text{PCA}}$  to make these two factors equal, which occurs at

$$f_{\text{IBCM}} = f_{\text{PCA}} \Rightarrow \Lambda_{\text{PCA}} = \frac{\beta}{\alpha \sigma^2} \frac{1 - f_{\text{IBCM}}}{f_{\text{IBCM}}}. \quad (\text{J4})$$

So, to set up parameters for a numerical simulation, we compute the statistics of the chosen background ( $\langle c \rangle$ ,  $\sigma^2$ ,  $m_3$ ), the analytical predictions for the IBCM fixed points ( $h_{\text{sp}}$ ,  $h_{\text{ns}}$ ,  $\bar{h}_{\text{d}}$ ), and we set  $\Lambda_{\text{PCA}}$  to these parameter values inserted in Eq. (J4).

- 
- [1] R. F. Thompson and W. A. Spencer, Habituation: A model phenomenon for the study of neuronal substrates of behavior, *Psychol. Rev.* **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, *Neurobiol. Learn. Mem.* **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, *Curr. Opin. Neurobiol.* **58**, 135 (2019).
  - [5] M. A. Webster, Visual adaptation, *Annu. Rev. Vision Sci.* **1**, 547 (2015).
  - [6] D. Chaudhury, L. Manella, A. Arellanos, O. Escanilla, T. A. Cleland, and C. Linster, Olfactory bulb habituation to odor stimuli, *Behav. Neurosci.* **124**, 490 (2010).
  - [7] D. Pérez-González and M. S. Malmierca, Adaptation in the auditory system: An overview, *Front. Integr. Neurosci.* **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, A. Gandhi, K. Ito, S. Sanyal, J. W. Wang, V. Rodrigues, and M. Ramaswami, Plasticity of local gabaergic interneurons drives olfactory habituation, *Proc. Natl. Acad. Sci. USA* **108**, E646 (2011).
  - [9] A. Celani, E. Villerman, 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. Biloft, and J. F. Bowers, Statistical characteristics of concentration fluctuations in dispersing plumes in the atmospheric surface layer, *Boundary Layer Meteorol.* **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, *Proc. R. Soc. 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, *Front. Cell. Neurosci.* **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, *J. Neurosci.* **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, *Curr. Biol.* **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, *Front. Comput. Neurosci.* **16**, 917786 (2022).
- [26] K. A. Fulton, D. Zimmerman, A. Samuel, K. Vogt, and S. R. Datta, Common principles for odour coding across vertebrates and invertebrates, *Nat. Rev. Neurosci.* **25**, 453 (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, *Proc. Natl. Acad. Sci.* **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, *Nat. Neurosci.* **13**, 439 (2010).
- [29] A. Vinograd, Y. Livneh, and A. Mizrahi, History-Dependent odor processing in the mouse olfactory bulb, *J. Neurosci.* **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, *J. Neurosci.* **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, *Prog. Brain Res.* **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, *Curr. Biol.* **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*, *J. Neurogenet.* **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] 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).
- [38] V. Minden, C. Pehlevan, and D. B. Chklovskii, Biologically plausible online principal component analysis without recurrent neural dynamics, in *Proceedings of the 52nd Asilomar Conference on Signals, Systems, and Computers* (IEEE, Pacific Grove, CA, 2018), pp. 104–111.
- [39] J. D. Zak, G. Reddy, M. Vergassola, and V. N. Murthy, Antagonistic odor interactions in olfactory sensory neurons are widespread in freely breathing mice, *Nat. Commun.* **11**, 3350 (2020).
- [40] 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, *Curr. Biol.* **30**, 2574 (2020).
- [41] V. Singh, N. R. Murphy, V. Balasubramanian, and J. D. Mainland, Competitive binding predicts nonlinear responses of olfactory receptors to complex mixtures, *Proc. Natl. Acad. Sci.* **116**, 9598 (2019).
- [42] 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 (London)* **593**, 558 (2021).
- [43] Y. Wu, K. Chen, C. Xing, M. Huang, K. Zhao, and W. Zhou, Human olfactory perception embeds fine temporal resolution within a single sniff, *Nat. Hum. Behav.* **8**, 2168 (2024).
- [44] S. Dasgupta, C. F. Stevens, and S. Navlakha, A neural algorithm for a fundamental computing problem, *Science* **358**, 793 (2017).
- [45] 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, *J. Neurosci.* **2**, 32 (1982).
- [46] C. C. Law and L. N. Cooper, Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory, *Proc. Natl. Acad. Sci.* **91**, 7797 (1994).
- [47] L. N. Cooper and M. F. Bear, The BCM theory of synapse modification at 30: Interaction of theory with experiment, *Nat. Rev. Neurosci.* **13**, 798 (2012).
- [48] E. Oja, Simplified neuron model as a principal component analyzer, *J. Math. Biol.* **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 Comput.* **5**, 61 (1993).
- [51] 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, *Netw., Comput. Neural Syst.* **10**, 111 (1999).
- [52] M. Froc and M. C. W. van Rossum, Slowdown of BCM plasticity with many synapses, *J. Comput. Neurosci.* **46**, 141 (2019).
- [53] L. Udeigwe, P. Munro, and G. Ermentrout, Emergent dynamical properties of the BCM learning rule, *J. Math. Neurosci.* **7**, 2 (2017).
- [54] G. Si, J. K. Kanwal, Y. Hu, C. J. Tabone, J. Baron, M. Berck, G. Vignoud, and A. D. Samuel, Structured odorant response patterns across a complete olfactory receptor neuron population, *Neuron* **101**, 950 (2019).
- [55] S. Gorur-Shandilya, M. Demir, J. Long, D. A. Clark, and T. Emonet, Olfactory receptor neurons use gain control and complementary kinetics to encode intermittent odorant stimuli, *eLife* **6**, e27670 (2017).
- [56] N. M. Chapochnikov, C. Pehlevan, and D. B. Chklovskii, Normative and mechanistic model of an adaptive circuit for efficient encoding and feature extraction, *Proc. Natl. Acad. Sci.* **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, Red Hook, NY, 2018).
- [58] J. Cafaro, Multiple sites of adaptation lead to contrast encoding in the *Drosophila* olfactory system, *Physiol. Rep.* **4**, e12762 (2016).
- [59] 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, *J. Neurosci.* **35**, 179 (2015).
- [60] M. C. Ogg, J. M. Ross, M. Bendahmane, and M. L. Fletcher, Olfactory bulb acetylcholine release dishabituates odor responses and reinstates odor investigation, *Nat. Commun.* **9**, 1868 (2018).
- [61] C. Linster and T. A. Cleland, Neuromodulation of olfactory transformations, *Curr. Opin. Neurobiol.* **40**, 170 (2016).
- [62] S. D. Shea, L. C. Katz, and R. Mooney, Noradrenergic induction of odor-specific neural habituation and olfactory memories, *J. Neurosci.* **28**, 10711 (2008).
- [63] T. A. Cleland and A. Borthakur, A systematic framework for olfactory bulb signal transformations, *Front. Comput. Neurosci.* **14**, 579143 (2020).
- [64] S. R. Odell, D. Clark, N. Zito, R. Jain, H. Gong, K. Warnock, R. Carrion-Lopez, C. Maixner, L. Prieto-Godino, and D. Mathew, Internal state affects local neuron function in an early sensory processing center to shape olfactory behavior in *Drosophila* larvae, *Sci. Rep.* **12**, 15767 (2022).
- [65] R. M. Blazing and K. M. Franks, Odor coding in piriform cortex: Mechanistic insights into distributed coding, *Curr. Opin. Neurobiol.* **64**, 96 (2020).
- [66] G. Reddy, B. I. Shraiman, and M. Vergassola, Sector search strategies for odor trail tracking, *Proc. Natl. Acad. Sci. USA* **119**, e2107431118 (2022).
- [67] N. Rigolli, G. Reddy, A. Seminara, and M. Vergassola, Alternation emerges as a multi-modal strategy for turbulent odor navigation, *eLife* **11**, e76989 (2022).
- [68] Y. Li, M. Swerdloff, T. She, A. Rahman, N. Sharma, R. Shah, M. Castellano, D. Mogel, J. Wu, A. Ahmed, J. San Miguel, J. Cohn, N. Shah, R. L. Ramos, and G. H. Otazu, Robust odor identification in novel olfactory environments in mice, *Nat. Commun.* **14**, 673 (2023).
- [69] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askill, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, *et al.*, In-context learning and induction heads, [arXiv:2209.11895](https://arxiv.org/abs/2209.11895).
- [70] 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, *Curr. Biol.* **34**, 5646 (2024).
- [71] J.-B. Lalanne and P. François, Chemodetection in fluctuating environments: Receptor coupling, buffering, and antagonism, *Proc. Natl. Acad. Sci. USA* **112**, 1898 (2015).
- [72] 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 Trans. Neural Networks Learn. Syst.* **33**, 3195 (2022).
- [73] 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, *Nat. Neurosci.* **12**, 210 (2009).
- [74] A. Apicella, Q. Yuan, M. Scanziani, and J. S. Isaacson, Pyramidal cells in piriform cortex receive convergent input from distinct olfactory bulb glomeruli, *J. Neurosci.* **30**, 14255 (2010).
- [75] I. G. Davison and M. D. Ehlers, Neural circuit mechanisms for pattern detection and feature combination in olfactory cortex, *Neuron* **70**, 82 (2011).
- [76] P. H. Schönemann, A generalized solution of the orthogonal procrustes problem, *Psychometrika* **31**, 1 (1966).
- [77] [https://github.com/frbourassa/olfactory\\_habituaton](https://github.com/frbourassa/olfactory_habituaton).
- [78] F. X. P. Bourassa, P. François, G. Reddy, and M. Vergassola, Manifold learning—Simulation code, Zenodo, (2025), doi: [10.5281/zenodo.17797288](https://doi.org/10.5281/zenodo.17797288).
- [79] M. Dow, Explicit inverses of Toeplitz and associated matrices, *ANZIAM J.* **44**, 185 (2003).
- [80] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in python, *Nat. Methods* **17**, 261 (2020).
- [81] DLMF, NIST digital library of mathematical functions, <https://dlmf.nist.gov/>, edited by 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 (2020).
- [82] D. T. Gillespie, The mathematics of Brownian motion and Johnson noise, *Am. J. Phys.* **64**, 225 (1996).
- [83] C. W. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences*, 4th ed., Springer complexity (Springer, Berlin, 2009).
- [84] N. Balakrishnan and W. W. S. Chen, *Handbook of Tables for Order Statistics from Lognormal Distributions with Applications* (Springer US, Boston, MA, 1999).